Bayesian Feature Selection via Variational Inference in Omics Data

Darren Andrew Vincent Scott

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy of the University of London.

August, 2022

Department of Medical Statistics

Faculty of Epidemiology and Population Health

London School of Hygiene & Tropical Medicine

Funded by UK Medical Research Council grant MR/N013638/1

Abstract

The advent of genome sequencing has led to a dramatic change in the scale and breadth of information within biology. Omics technologies have enabled a single experiment to generate a very large amount of raw data, of increasingly complex phenomena. This data is often high-dimensional, the size raises questions about the efficiency of the computational approach used to estimate the model and the number of attributes often exceed the number of observations. The focus of the thesis is on Bayesian feature selection in high-dimensional omics data via variational inference. Our objective is to develop and implement reliable inferential tools that scale efficiently with dimensionality.

Our first algorithm identifies compositional covariates and effect sizes associated with a response of interest via auxiliary indicator variables. This is particularly useful for data sets generated from genome sequencing technology such as human microbiome, as these only contain information on the relative magnitudes of the compositional components. Novel priors account for model constraints and a Monte Carlo step, guided by the data, is introduced to estimate intractable marginal expectations.

We extend the methodology to a multidimensional response, where different compositional covariates are free to be associated with different responses. This allows the relationship between the microbiome and complex phenotypes such as lipids or metabolites to be explored in one model, facilitating a system genetics approach to understanding the flow of biological information. By a reparameterisation of the likelihood, we are able to perform fast covariance and covariate selection despite the vast model space.

A hierarchical Bayesian model is developed for clusters of individuals who exhibit different causal pathways to the same multi-dimensional endpoint. Again, we are able to reparametrise the likelihood to incorporate fast predictor and covariance selection within a large model space. We capture the different latent structures across the clusters to aid model fitting and understanding. Sparse

feature selection is performed both within each expert and in the unsupervised learning of cluster detection.

Our hope is that the software which follows the methods we have outlined will be used by practitioners to develop biological understanding and insight.

Acknowledgements

I would like to thank my supervisor Alex Lewin for all her advice and guidance. Alex has always made time for me and has been incredibly patient.

Throughout the PhD I have received a great deal of encouragement and support from my girlfriend Chloë, which has been invaluable. Thank you.

Contents

List of Abbreviations

ALL Acute Lymphoblastic Leukemia
alr Additive-log-ratio
AML Acute Myeloid Leukemia

BMI Body Mass Index

CAVI Coordinate Ascent Variational Inference
CAVI-MC Coordinate Ascent Variational Inference Monte Carlo
clr Centred-log-ratio

DAG Directed Acyclic Graph

ELBO Evidence Lower Bound Optimisation
EM Expectation Maximisation

FPR False Positive rate

HME Hierarchical Mixture of Experts

ilr Isometric-log-ratio

KL Kullback-Liebler

MCMC Markov Chain Monte Carlo
ML Maximum Likelihood
MRF Markov Random Field

OLS Ordinary Least Squares
OTU Operational Taxonomic Unit

PRMx Random Partition Model with covariates

QTL Quantitative Trait Loci

RJMCMC Reversible Jump Monte Carlo Markov Chain

SMVN Singular Multivariate Normal Distribution
SNR Signal to Noise Ratio
SUR Seemingly Unrelated Regression
SVI Stochastic Variational Inference

TPR True Positive Rate

VI Variational Inference

CHAPTER 1

Introduction

The advent of genome sequencing has led to a dramatic change in the scale and breadth of information within biology. Omics technologies such as microarrays, proteomics or high-throughput cell assays, have enabled a single experiment to generate a very large amount of raw data of increasingly complex phenomena. This data is often high-dimensional and exhibiting characteristics which are classified under the term *big data*: (a) the number of attributes greatly exceed the number of observations, (b) the size of the data set is sufficiently large to raise questions about the efficiency of the computational approach used to estimate the model. In the microbiomics setting there is an additional complexity, as the data produced from the nucleotide sequencing is compositional (Gloor et al., 2017). The magnitude of a single operational taxonomic unit (OTU) depends on the sum of all the OTUs counts, and only provides information about the relative magnitudes of the compositional components.

The two main tasks in high-dimensional statistical analysis, where variable selection is essential to knowledge discovery, are: construction of a method to predict future observations and build

understanding and insight of the model that generates the data. Since prediction accuracy is often comprised by intepretabilty and conciseness, achieving both those goals simultaneously is rarely possible. Optimal prediction and model inference are rarely achieved by a single parsimonious model, instead both benefit from model averaging where inference on issues that are not model-specific (such as prediction or covariate effects) is averaged over the set of models under consideration. Our focus lies with statistical inference, selecting a model of the process that generates the data and deducing propositions from the model. Our problem is one of variable selection, identifying the relevant covariates in a multiple regression model, where the expected total number is small or "sparse". Despite considerable work, this problem remains an active area of research as a cornerstone of many fields.

There is considerable interest in determining a subset of omics variables (or characteristics) which provide a good description of the observed phenomenon. The omics revolution has also led to a change of emphasis. Rather than using a direct phenotype of interest, variable selection for a set of "intermediate" complex phenotypes (which are usually highly correlated) offers the chance to increase our understanding of the genes, pathways and networks that underlie common human disease. In the causal framework this is analogous to identifying the mechanism that underpins the relationship between the molecular biology and the disease, where the multiple phenotypes are downstream of the covariates in the causal pathway. In terms of understanding the global molecular architecture of complex traits, this is referred to as a "system genetics approach" (Civelek and Lusis, 2014).

Compositional data contain only relative information, and are typically recorded as closed data (each data row sums to a constant). Values are not free to range from $-\infty$ to $+\infty$ and are always positive. Such data is widespread in microbiomics, given the limitations of nucleotide sequencing. Compositional data exhibits particular and important properties that cause well known problems in standard statistical analysis, these have been elucidated and discussed by a number of authors (Butler (1979), Davis (2002), Aitchison (1986), Egozcue and Pawlowsky-Glahn (2005)). In order to model compositional data with standard statistical techniques, a transformation must be performed to transfer the compositional vectors into the Euclidean space.

Aitchison (1982) introduced the additive-log-ratio (alr) and centred-log-ratio (clr) transformations, and Egozcue et al. (2003) the isometric-log-ratio (ilr) transformation. The three representations have different properties which are explored in chapter 5 of the thesis.

In the Bayesian framework, prior uncertainty regarding the values of the parameters within the regression model is expressed in terms of prior probability distributions. The uncertainty over the model space, can also be expressed with priors, and model selection performed after integrating over the uncertainty of the parameter values via Bayes factor. In high-dimensional omics data the space of models is large, posing a challenge to this method of model selection. A variety of explicit and shrinkage priors have been developed in the literature to perform sparse learning. Shrinkage priors in the Bayesian framework have been well studied since the observation that the variety of penalties imposed within the likelihood in penalized methods, are equivalent to priors on the parameters, thus leveraging the extensive methodology developed within the field. Determining the shrinkage properties involves the study of properties of prior distributions on the regression coefficients after an estimator has been applied. Explicit variable selection priors involve augmenting the model with binary inclusion variables, indicating whether each variable should be included in the model. A natural choice is the "spike-and-slab" two component mixture prior, where the first component allows nonzero entries and the second component drives the coefficients towards zero. Although the analytical intractability of the posterior distributions from these priors prevents exact inference, samples can be obtained via Markov chain Monte Carlo (MCMC) methods (Robert and Casella, 1999).

Unlike Bayesian shrinkage models which tend to admit efficient implementations of the Gibbs samplers (Park and Casella, 2008), posterior calculations in explicit selection are often more involved, since they entail simultaneous exploration of parameter and model space, and face difficulties in traversing dimensions. An exhaustive search over the space of models with an ever increasing number of predictors is impractical. George and McCulloch (1993) introduced the Gibbs sampler in the context of spike-and-slab variable selection, laying the foundations for stochastic model search. In order to avoid expensive updating of the regression coefficient vector in high-dimensions, George and McCulloch (1997) suggested integrating over the regression parameters

to sweep only through the model space. Various MCMC stochastic search techniques have been deployed to discover high probability models (Hans et al. (2007), Bottolo and Richardson (2010), Stingo and Vannucci (2011), and Lewin et al. (2016)). Alternatively Metropolis-within-Gibbs routines have been successively applied to rapidly evaluate posterior model selection uncertainty in problems of a manageable size and provide posterior model parameter estimation (Dellaportas et al. (2002), Banterle and Lewin (2018) and Zhang et al. (2020)).

Feature selection in omics data is complicated further by multiple molecular responses related through a latent structure. Capturing this within the model, offers the opportunity to increase statistical power Inouye et al. (2012) and improve model estimation and data understanding. The matrix of responses can be incorporated via a matrix normal likelihood which captures the correlation of the residuals. Identifying an intepretable model now involves sparse selection of the predictor variables and the off-diagonal covariance elements, often in the form of the precision matrix. Popular methods for Bayesian structure learning in regression involve Gaussian graphical modelling for both decomposable and non-decomposable cases, and explicit selection. Gaussian graphical determination can be viewed as a covariance selection problem (Dempster et al., 1977), where the non-zero entries in the off-diagonal of the precision matrix correspond to edges in the graph. Difficulties arise in allowing the choice of covariates associated with each response to vary in the model whilst applying some form of selection on the covariances and the sheer size of the model space. For a $T-$dimensional variable there are $2^{T(T-1)/2}$ possible conditional independence graphs. Even with a moderate number of variables, the model space is astronomical in size. To make the problem computationally feasible simplifying assumptions are made to exploit conjugacy with respect to regression coefficients and residual covariance. One simplification is for model selection to be restricted to the same subset of variables for each response. The other alternative, is to assume independence across the responses.

With the addition of multiple responses, the model space involves combinations of regression coefficients and off-diagonal covariance elements. Explicit covariance selection relies on decomposing the covariance matrix and augmenting the reparameterised likelihood with latent covariance indicator variables, enabling the range of MCMC stochastic search methods to be exploited. Structure

learning via graphical models requires an additional search algorithm which explores the graph space to distinguish important edges from irrelevant ones and detect the underlying graph with high accuracy. Various adaptations of the reversible-jump Monte Carlo Markov chain (RJMCMC) have been developed (Brooks et al. (2003), Mohammadi and Wit (2015)) to explore the transdimensional space. Alternatively, the decomposable graphical structure can be explored efficiently via a sampler introduced by Green and Thomas (2013), which makes use of the junction tree representation (Cowell et al., 2007) to allow for bolder, multi-edge proposal in the graph space.

Despite the various MCMC adaptations, for sufficiently large scale univariate (and multivariate) response model selection problems the approach can be deemed to be too slow in practise. Variational inference (VI) is an alternative technique which sacrifices some posterior accuracy in return for computational speed, yielding an estimate of the full posterior by optimising an approximate posterior over a class of distributions. The quality of the approximation is generally measured by the Kullback-Lielber (KL) divergence. Approximate solutions arise by restricting the family of densities which can be used as a proxy for the exact conditional density. By choosing conditionally conjugate prior distributions, and specifying independence across the factors through a mean field variational family, closed form iterative updates which minimise the KL divergence between the approximating densities and the exact posterior densities are obtained (Carbonetto and Stephens, 2012). Its success in solving a variety of machine learning problems with very large data sets, in topics such as neuroscience (Woolrich et al., 2004), grammar induction (Kurihara and Sato, 2006) and image denoising (Likas and Galatsanos, 2004) has led to concerted efforts in the literature to encourage its use by statisticians (Blei et al. (2017), Ormerod and Wand (2010)). The speed of VI gives it an advantage, particular for exploratory regression, where a very large model is fitted to gain an understanding of the data and identify a subset of covariates which can be explored in more detail. Carbonetto and Stephens (2012) use VI as a deterministic alternative to stochastic search algorithms for linear regression with a univariate response for large omics datasets. This is extended to multiple responses by Ruffieux et al. (2017), with the use of a hierarchy framework similar to Bottolo et al. (2011).

1.0.1 Overview

The focus of the thesis is on Bayesian feature selection in high-dimensional omics data via VI. Our objective is to develop and implement reliable inferential tools that scale efficiently with dimensionality. The thesis is structured as follows:

Chapter 2 reviews the univariate Bayesian variable selection techniques in the context of high-dimensional data. As Bayes factor is not appropriate the two main approaches, explicit variable selection and shrinkage priors are considered. The computational challenges which accompanies the non-conjugate prior specifications are detailed.

Chapter 3 is an overview of feature selection for multivariate response linear regression. This involves the selection of both, a significant matrix subset of regression coefficients via explicit variable selection and hierarchical priors, and the off-diagonal elements of the covariance matrix across the responses. Two approaches for Bayesian structure learning are discussed, Gaussian graphical modelling with decomposable graphs and explicit covariance selection.

Chapter 4 explores the basic idea behind variational inference, starting with mean-field inference and coordinate-ascent optimization. A comparison is made with the Expectation Maximisation (EM) algorithm, commonly used in maximum likelihood estimation, highlighting the similarities between the two approaches. The approach is expanded to stochastic variational inference (Hoffman et al., 2013), an stochastic optimisation alternative which scales variational inference to massive data (large number of rows). Throughout the chapter, a Gaussian mixture example is used to put the theory into context.

Chapter 5 is a brief overview of three compositional transformations; alr, clr and ilr, which take the vector from the simplex space to the Euclidean space. This is particular important when incorporating compositional data as covariates in linear regression. As the ilr transformation is a series of projections on to a non-unique orthonormal basis, it can be defined in term of balances between two groups. This interpretation, along side its link to the clr transformation is explained.

Thus, Chapters 2 to 5 cover the core statistical subjects which are utilised in the Bayesian model

building and estimation, within the three articles in the remaining chapters.

Chapter 6 is a slightly extended version of our first journal article, proposing a Bayesian hierarchical linear log-contrast model estimated by mean field Monte-Carlo co-ordinate ascent variational inference (CAVI-MC). This enables compositional microbiome features, associated with a response, to be identified alongside other covariates of interest within a Bayesian model framework. Novel priors are posited in a hierarchical framework which account for the large differences in scale of the parts within the compositional vectors and the constrained parameter space associated with the compositional covariates. A reversible-jump Monte Carlo Markov chain (RJMCMC) is added to the VI framework to estimate intractable approximate marginal expectations. This is guided by the data through univariate approximations of the variational posterior probability of inclusion. We exploit the nested nature of variational inference by proposing parameters from approximated variational densities via auxiliary parameters. Our approach is applied analysis of real data exploring the relationship of the gut microbiome to body mass index (BMI).

Chapter 7 is a second journal article, extending the univariate Bayesian hierarchical linear log-contrast model estimated by mean field Monte-Carlo co-ordinate ascent VI to multiple responses related by a latent structure. Compositional microbiome feature selection can now be performed against biological systems rather than univariate responses. Correlation between the responses is captured by a reparameterisation of the seemingly unrelated regression framework, overcoming the difficulties in multiple response covariate selection, to allow different regressors to be associated with different responses. Explicit covariance selection through spike-and-slab priors conveniently bypasses the problems which can be encountered when selecting parameters within a positive definite matrix, whilst a shrunken estimate of the precision matrix is available after a back transformation. We use priors which account for the large difference in scale and constrained parameter space associated with the compositional covariates. Intractable marginal expectations are again estimated by a RJMCMC which is guided by the data through univariate approximations of the variational posterior probability of inclusion, with proposal parameters informed by approximating variational densities via auxiliary parameters. We apply our CAVI-MC model to the "Know Your Heart" study, exploring the relationship between gut microbiome, health covariates and a

set of biomarkers.

Chapter 8 is a draft article motivated by clusters of people who exhibit different causal pathways to the same multi-dimensional endpoint. A hierarchical multivariate response Bayesian mixture of experts model is developed, which captures the cluster specific correlation structure between the multiple responses, aiding model fitting and understanding. A reparameterisation of the seemingly unrelated regression (SUR) model ensures the approach is feasible for high-dimensional omics data. Cluster specific feature selection within the experts exploits sparsity to facilitate both covariate and covariance selection, where the combination of covariates is free to vary across the experts. The unsupervised learning of detecting new information in the clustering of individuals is determined by a subset of their predictors. The model is estimated by block-mean-field coordinate ascent VI so that it scales efficiently with high-dimensional data.

Chapter 9 is a general discussion on possible future extensions of the research. Each of our research articles is accompanied by software which we plan to make into python modules. There is also scope to alter the methods so that the software can be deployed with massive data sets and in more general settings. The proposed methods for feature selection of compositional covariates can be hampered by the presence of a high degree of multicollinearity. Various approaches in the literature, which address this problem within the prior specification and could be incorporated into our model, are detailed. Currently, the performance of our multivariate response hierarchical mixture of experts (HME) model is unknown. A simulation study is proposed to evaluate the feature selection within the experts, in comparison with current mixture of regression methods. By applying the model to an omics data set which contains two types of leukemia patients, the clustering accuracy can be demonstrated.

The thesis finishes with a conclusion in **chapter 10**, highlighting the benefits of our inferential tools.

CHAPTER 2

Univariate Variable Selection

The question of detecting the location of the variable which is associated with a response can be framed generally as a model selection problem. Suppose there are a set of K models $\mathcal{M} = \{M_1, ..., M_k\}$ under consideration for data $\mathbf{y}$ which has the density $p(\mathbf{y}|\boldsymbol{\vartheta}_k, M_k)$, where $\boldsymbol{\vartheta}_k$ is a vector of unknown parameters that indexes the members of M_k. A hierarchical structure is introduced where, a prior probability $p(M_k)$ is assigned to each model, conditional on the model a prior is then assigned to the parameters of each model $p(\boldsymbol{\vartheta}_k|M_k)$ and the data is assigned a density $p(\mathbf{y}|\boldsymbol{\vartheta}_k, M_k)$. The problem of model selection is then one of identifying the model that generated the data, which can be expressed as the posterior model probability of

$$p(M_k|\mathbf{y}) = \frac{p(\mathbf{y}|M_k)p(M_k)}{\sum_k p(\mathbf{y}|M_k)p(M_k)} \tag{2.0.1}$$

where

$$p(\mathbf{y}|M_k) = \int p(\mathbf{y}|\boldsymbol{\vartheta}_k, M_k)p(\boldsymbol{\vartheta}_k|M_k)d\boldsymbol{\vartheta}_k \tag{2.0.2}$$

is the marginal likelihood of M_k. Based on these probabilities a pairwise comparison between two models M_1 and M_2 can then be performed using

$$\frac{p(M_1|y)}{p(M_2|\mathbf{y})} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} \times \frac{p(M_1)}{p(M_2)}$$

Drawing inference from the marginal posterior model probability, as in Carlin and Chib (1995), is difficult when the model space contains 2^p (or $2^{p\times T}$ with T responses) possibilities as the MCMC will rarely visit any of the models. Explicit variable selection transforms the model indicator $M \in \{1, ...k\}$ into a binary covariate indicator which characterises the model space. Therefore variable selection can be considered a problem of determining a subset of the explanatory variables $X_1, ..., X_p$, where each subset is an element of $\mathcal{M}$, which best explains the variability in the response(s) $\mathbf{y}$ within a multivariate linear regression (assuming the response is continuous and the distributional assumption of the residuals is reasonable). This can be performed by shrinkage priors with an appropriate threshold without searching through the model space. Often the underlying relationship is considered to be "sparse" with only a small number of variables effecting the response, while most have little or no effect and the prior exchangeable over the design matrix (as we are ignorant of where any influential variable may be). The interpretability of the model after variable selection is important, so that biological understanding and insight can be obtained.

In this chapter

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{2.0.3}$$

where $\mathbf{y} = (y_1, ..., y_n)^T$ is a sequence of n observed responses (univariate regression) and $\boldsymbol{X} = X_1, ..., X_p$ form the $n \times p$ design matrix, $\boldsymbol{\theta}$ is a $p \times 1$ vector of unknown coefficients and $\boldsymbol{\epsilon}$ is a vector of residuals assumed to follow a normal distribution $N(\mathbf{0}, \sigma^2 I_n)$. Hence σ^2 is an unknown positive scalar. The intercept is routinely included in all models, or the data can be centred removing the intercept from the model (assumed here) which is equivalent to integrating out the intercept with respect to an improper, uniform prior (Chipman et al., 2001). In this chapter we assume the data has been centred and rescaled so the covariates are comparable quantities. This also improves the efficiency of the MCMC sampling by reducing the autocorrelation in the chains.

The variable selection procedure can be considered to be one of determining which of the regression parameters θ_j are equal to zero. How the θ_j is parameterised in (2.0.3) defines the nature of the variable selection and its characteristics. Two main approaches currently dominate the literature and they are explored in this chapter.

2.1 Explicit Variable Selection

Explicit variable selection uses an auxiliary indicator variable γ_j with respect to the covariates, to determine which regression parameter should be included in the model (where $\gamma_j = 0$ indicates absence of the covariate and $\gamma_j = 1$ indicates the presence of the covariate). γ_j is a Bernoulli random variable governed by the rate of success or sparsity parameter $p(\gamma_j = 1) = \omega$. Each regression model is thus uniquely characterised by a vector of binary inclusion variables $\boldsymbol{\gamma}$, which characterize a specific linear combination of covariates. Prior uncertainty in ω can be used to induce sparsity into the model.

The actual variable selection can proceed in several ways. Two popular strategies applied in practice are: (1) to select a model with the highest estimated posterior probability (the highest posterior density model), (2) to select variables with estimated posterior marginal inclusion probabilities higher than 0.5 (the median probability model (Barbieri and Berger, 2004). The appropriateness of these two approaches was studied by (Barbieri and Berger, 2004) using expected mean squared error of a future observation as a loss function. Under the assumption of an orthogonal design matrices, the authors found the optimal predictive model was the median probability model rather than the highest posterior density model.

2.1.1 Parameterisation of the latent variable

The discrete mixture distribution (unlike their adaptive shrinkage counterparts which are continuous) $\gamma_j \sim Bern(\omega)$ is part of the "two group" shrinkage priors, which add information to help solve, regularise, the variable selection problem (Polson and Scott, 2011). The different approaches

within explicit variable selection are characterised by where γ_j is located, how θ_j is parameterised and the relationship between γ_j and the covariate. One approach is to define $\theta_j = \gamma_j\beta_j$ so that $\theta_j = \beta_j|\gamma_j = 1$ and $\theta_j = 0|\gamma_j = 0$ (Kuo and Mallick, 1998), which implies that the vector of indicators $\boldsymbol{\gamma}$ only enters the model via the likelihood and not through the prior for $\boldsymbol{\beta}$. The indicator and coefficient are assumed independent apriori $p(\gamma_j, \beta_j) = p(\gamma_j)p(\beta_j)$ (the posterior will be conditional on the parameter values) with independent priors placed on the γ_j and β_j. The covariate is removed from the model when the indicator variable is 0, compressing the design matrix in the posterior calculations. Motivated by the mixing of the MCMC Dellaportas et al. (Dellaportas et al., 2002) extended this approach by conditioning the prior distribution of β_j on to the indicator variable, whilst retaining the mixture of normal priors and $\theta_j = \gamma_j\beta_j$, resulting in a mixture distribution

$$p(\beta_j, \gamma_j) = p(\beta_j|\gamma_j)p(\gamma_j). \tag{2.1.1}$$

As the indicator variable ($\gamma_j = 0$) removes the covariate from the likelihood the prior does not impact the posterior, but proposes value for the covariates at the next step of sampler. The parameters of this "pseudoprior" merely serve as tuning parameters for the algorithm with $p(\beta_j|\gamma_j = 0)$ concentrated around θ_j, which is philosophically contentious. There is also an issue of identifiability of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in the likelihood as $\gamma_j = 0 \times \beta_j = 0$ and $\gamma_j = 1 \times (\beta_j \approx 0) \approx 0$ which can impact interpreting the marginal posterior.

Alternatively $\theta_j = \beta_j$ and $p(\beta_j|\gamma_j)$, giving identifiability for variables γ_j and β_j as the indicator variable only effects the prior distribution due to the hierarchical relationship expressed in (2.1.1). The prior can be characterised by a mixture distribution such as the Gaussian "spike-and-slab" where a natural choice is

$$\beta_j|\gamma_j \sim \gamma_j N(0, \tau^2 c) + (1 - \gamma_j)N(0, \tau^2). \tag{2.1.2}$$

where τ is set small ($\tau > 0$) creating a "spiked" Gaussian distribution, so that if $\gamma_j = 0$ then β_j would probably be so small that it could be safely estimated by 0. The diffuse Gaussian distribution or "slab", is from setting c to a large value ($c > 1$) so that the support incorporates

realistic parameter values when β_j is non zero whilst avoiding excessively large values of c, putting an ever increasing weight on the null model (George and McCulloch, 1993).

The hypothesis being tested is $H_0 : \beta_j \approx 0$ vs $H_1 : \beta_j \neq 0$, thus variable selection requires thresholding. Chipman et al. (2001) define the prior variance as $\tau^2 = \tau_0$ and $c\tau^2 = \tau_1^2$, and define the threshold value as

$$\delta = \sqrt{\frac{\log\left(\frac{\tau_1^2}{\tau_0^2}\right)}{\frac{1}{\tau_1^2} - \frac{1}{\tau_0^2}}} \tag{2.1.3}$$

to classify whether a regression coefficient is classified as belonging to the slab (spike) component and is not shrunk (shrunk) to zero. Clearly, elicitation of of the variance parameters is important for variable selection. Fixing these two values may result in in inconsistent variable selection. Narisetty and He (2014) propose values that are functions of n and p to ensure good performance of the model when the data dimensions increase.

Alternatively one can place a prior on τ, the choice of an exponential $\lambda^2/2$ will convert the slab into a Laplace prior. Care needs to be taken when choosing λ, too large a value will make the spike-and-slab indistinguishable and posterior inclusion probabilities will be meaningless.

Finally, the choice of $\theta_j = \beta_j$ is not possible if the spike in (2.1.2) is changed to a Dirac distribution (this is discussed in detail next) as β_j is fixed at zero if $\gamma_j = 0$, effectively forcing the indicator into the likelihood.

2.1.2 Conditional conjugacy and the Dirac distribution

Defining the “spike" as a point mass at 0 (Dirac distribution δ_0) and the prior in the form of

$$\beta_j|\gamma_j \sim N(0, \sigma_\beta^2)^{\gamma_j}\delta_0(\beta_j)^{1-\gamma_j} \qquad p(\gamma_j) = \omega^{\gamma_j}(1-\omega)^{1-\gamma_j}, \tag{2.1.4}$$

is a convenient and computational efficient conditionally conjugate parameterisation of the prior on β_j (George and McCulloch, 1997). The number of non zero covariates is p_γ and ω represents the prior probability that a coefficient is non-zero. Here σ_β^2 has a large impact on the resulting

coefficients in terms of shrinkage and variable selection properties.

The use of alternative distributions to the Gaussian slab in (2.1.4) are rare, particularly because of the conjugate properties when paired with the likelihood. Recently Ray and Szabó (2021) out performed Gaussian priors with a centred Laplace slab, when approximating the posterior distribution, using mean field co-ordinate ascent variational inference (CAVI) to estimate the model. The approximate posterior remained a Gaussian spike-and-slab, but the heavier tails of the Laplace prior prevented excess shrinkage in the marginal probability of inclusion.

Traditionally MCMC is used to compute the posterior and a choice of conjugate priors for the linear model will lead to a Gibbs sampler with iterations over the regime:

- Sample from $p(\beta_j, \gamma_j|\cdot, \mathbf{y})$ for all $j = 1, ..., p$.
- Sample the data variance parameter σ^2. An inverse gamma prior leads to an inverse gamma marginal posterior.
- Sample the sparsity parameter $p(\omega|\cdot, \mathbf{y})$. A beta prior leads to a beta marginal posterior.

The first step avoids sampling from the full conditional of γ_j

$$p(\gamma_j = 1|\beta_j, \boldsymbol{\beta}_{-j}, \omega, \sigma^2_\beta, \mathbf{y}), \tag{2.1.5}$$

which would prevent the sampler from exploring the model space, as the probability is one if $\beta_j \neq 0$ and 0 otherwise. This is not an issue in the case of the Gaussian spike (2.1.2) which allows samples of β_j to be slightly different from zero. Instead a joint update of (β_j, γ_j) is performed, iterating over $p(\gamma_j|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}_{-j}, \tau, \sigma^2_\beta, \mathbf{y})$ and then $p(\beta_j|\gamma_j, \boldsymbol{\beta}_{-j}, \boldsymbol{\gamma}_{-j}, \tau, \sigma^2_\beta, \mathbf{y})$ (Appendix 2.3.1). The probability of including a parameter in the model is

$$p(\gamma_j = 1|\mathbf{y}, \cdot) = \frac{\frac{\phi(0|0,\sigma^2_\beta)}{\phi(0|m,v)}\omega}{(1-\omega) + \frac{\phi(0|0,\sigma^2_\beta)}{\phi(0|m,v)}\omega}, \tag{2.1.6}$$

where m and v are the mean and variance of the full conditional posterior distribution β_j and

$\phi(0|m, v)$ is the density at zero of the normal distribution

$$m = vX_j^T\Big(\mathbf{y} - \sum_{s\neq j} X_s\gamma_s\beta_s\Big) \qquad v = \Big(\frac{||X_j||^2}{\sigma^2} + \frac{1}{\sigma_\beta^2}\Big). \tag{2.1.7}$$

The expression (2.1.6) is a function of the sparsity parameter ω and the prior variance σ_β^2 in the form

$$\frac{\phi(0|0, \sigma_\beta^2)}{\phi(0|m, v)}\omega, \tag{2.1.8}$$

where σ_β^2 is in both the numerator and denominator. The results can be sensitive to the choice of these values. The importance of ω is clear, too small and the posterior probability of inclusion is overly shrunk. Too large and the effect is to make it difficult for γ_j to identify the true variables in the model. The reverse is the case for σ_β^2.

The choice of $\omega = 0.5$ in the Bernoulli prior is not uniform, as it implies a prior expectation that half of the p predictors will be included in the final model. In high-dimensional settings where sparsity is expected, this value can be set close to 0. A prior can be placed on this parameter, a conjugate choice is a beta distribution. A typical choice is $Beta(1, \alpha)$ where α is set to the number of predictors. Carvalho et al. (2011) use a sparsity inducing prior, with a mixture prior of

$$\omega|\rho \sim (1 - \rho)\delta_0(\omega) + \rho\ Beta(\omega|1, \alpha). \tag{2.1.9}$$

2.1.3 Marginal model posterior

Model selection can be be performed via the marginal posterior $\mathrm{p}(\boldsymbol{\gamma}|\mathbf{y})$, which requires integrating over the other parameters in the likelihood. For convenience this is referred to as the marginal likelihood despite containing $\boldsymbol{\gamma}$ and the hyperparameters. The approach is made easier by changing the parameterisation of the spike-and-slab prior to

$$\beta_j \sim N(0, \sigma^2\tau)^{\gamma_j}\delta_0(\beta_j)^{1-\gamma_j}, \tag{2.1.10}$$

where τ is a scaling parameter which can be fixed or calculated using cross validation. However, even with a simple model (Appendix 2.3.2) this is only available up to a constant of proportionality

$$p(\boldsymbol{\gamma}|\mathbf{y}) \propto \frac{1}{\tau^{p_\gamma/2}|(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma + 1/\tau|^{1/2}} \left(\frac{1}{2}\left[\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\boldsymbol{X}_\gamma(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma + 1/\tau)^{-1}\boldsymbol{X}_\gamma^T\mathbf{y} + ab\right] \right)^{-\frac{n+a}{2}} p(\boldsymbol{\gamma}), \tag{2.1.11}$$

where a and b are the hyperparameters for the inverse gamma prior on the variance parameter σ^2. The expression $\boldsymbol{X}_\gamma$ defines the selected covariates from the non-zero entries in $\boldsymbol{\gamma}$. The posterior model probabilities $p(\boldsymbol{\gamma}|\mathbf{y})$ (which remain a function of the hyperparameters) quantify the posterior evidence for selecting each particular model, thus suggesting models with the highest values as suitable candidates. This can be evaluated with an approximation of the normalising constant d

$$p(\boldsymbol{\gamma}|\mathbf{y}) = dg(\boldsymbol{\gamma}), \tag{2.1.12}$$

by selecting a subset of $\boldsymbol{\gamma}$ values (a set of values visited from a previous simulation) and letting $g(A) = \sum_{\gamma \in A} g(\gamma)$ so that $p(A|\mathbf{y}) = dg(A)$. A consistent estimate of d is obtained by

$$\hat{d} = \frac{1}{g(A)K}\sum_{k=1}^{K} I_A(\boldsymbol{\gamma}^{(k)}), \tag{2.1.13}$$

where $I_A(.)$ is the indicator of the set A and $\boldsymbol{\gamma}^{(k)}$ is the value from the kth iteration.

MCMC techniques offer an alternative to approximating d, by simulating a chain of models with (2.1.11), to find interesting regions of the model space with an accumulation of posterior mass. The marginal posterior distribution of $\boldsymbol{\gamma}$ can be decomposed by Bayes Formula (Appendix 2.3.3) into

$$p(\gamma_j|\boldsymbol{\gamma}_{-j}, \mathbf{y}) = \frac{p(\gamma_j = 1|\boldsymbol{\gamma}_{-j})}{p(\gamma_j = 1|\boldsymbol{\gamma}_{-j}) + F(\boldsymbol{\gamma}, \boldsymbol{\gamma}')^{-1} \times p(\gamma_j = 0|\boldsymbol{\gamma}_{-j})}, \tag{2.1.14}$$

which (is exact rather than proportional) and involves the conditional prior probability $p(\gamma_j|\boldsymbol{\gamma}_{-j})$ and the marginal likelihood (2.3.12) in the Bayes factor

$$F(\boldsymbol{\gamma}, \boldsymbol{\gamma}') = \frac{p(\mathbf{y}|\gamma_j = 1, \boldsymbol{\gamma}_{-j})}{p(\mathbf{y}|\gamma_j = 0, \boldsymbol{\gamma}_{-j})}. \tag{2.1.15}$$

Expression (2.1.14) is iterated over by selecting an index i at random, and then sampling a Bernoulli random variable with probability $p(\gamma_i = 1|\boldsymbol{\gamma}_{-i}, \mathbf{y})$. As this involves the inverse of $\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma + 1/\tau = \mathbf{A}_\gamma$, the computational importance of the latent variable $\boldsymbol{\gamma}$ in reducing the dimensions of the design matrix to $n \times p_\gamma$ when the model space is large is clear. As p grows large, the ability to search the space of models $\{0,1\}^p$ diminishes. However this approach can still be effective for large p if the true model space is sparse, by restricting the models sampled.

Another popular MCMC strategy is MCMC Model Composition (MC^3), originally proposed in the context of graphical models (Madigan et al., 1995). The procedure results in a sequence of visited models $\boldsymbol{\gamma}^{(1)}, ..., \boldsymbol{\gamma}^{(M)}$, generated according to a Metropolis-Hastings routine. The proposal distribution is concentrated at close proximity to the current state of $\boldsymbol{\gamma}$, thereby restricting models differing by an inclusion of exclusion of just one variable. The candidate model $\boldsymbol{\gamma}^*$ sampled form the proposal distribution is then accepted with probability $\min[1, p(\boldsymbol{\gamma}^*|\mathbf{y})/p(\boldsymbol{\gamma}|\mathbf{y})]$, as the posterior ratio is available up to a constant of proportionality.

The stochastic search for variable selection is limited by its inability to escape from local posterior peaks, or to discover relevant but isolate regions of the model space. To resolve this issue, a population of chains can be run in parallel, each chain associated with a particular "tempered version" of the target distribution. In the model selection context the target distribution is the posterior distribution over the model space $p_t(\boldsymbol{\gamma}|\mathbf{y})$, which is now a function of the temperature t. The tempering acts to flatten the peak of the true target distribution. The higher the temperature, the easier is for the chain to escape the peaks. Furthermore, the parallel chains interact and learn from each other, making the exploration of the model space more efficient. The interaction is achieved by altering/swapping model configurations between/within the chains with different temperature at each MCMC iteration. Liang and Wong (2000) introduced the hybrid procedure Evolutionary MCMC (EMC) by combining the idea of parallel tempering together with genetic algorithms. This was applied by Bottolo and Richardson (2010) in Bayesian model selection.

Once a model has been selected by sampling over the marginal posterior distribution (2.1.11), the posterior distribution of the non-zero coefficients $p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \tau, \mathbf{y})$ is available to qualify the effect size

and associated uncertainty. It follows a multivariate t-distribution with $n+a$ degrees of freedom (Appendix 2.3.4),

$$\boldsymbol{\beta}|\boldsymbol{\gamma},\tau,\mathbf{y} \sim t_{n+a}\left(\mathbf{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\mathbf{y}, \frac{C_\gamma + ab}{n+a}\mathbf{A}_\gamma^{-1}\right), \tag{2.1.16}$$

and

$$C_\gamma = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\boldsymbol{X}_\gamma\mathbf{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\mathbf{y} \qquad \mathbf{A}_\gamma = \boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma + \frac{1}{\tau}. \tag{2.1.17}$$

2.1.4 Posterior predictive

In the presence of large uncertainty about variable selection, making predictions based on a single model can be inadequate. Predictions can be sensitive to the particular model-selection strategy and any interval from a single model can substantially undermine the the uncertainty about a prediction. The prior parameterisation means the predictive distribution for m new values $\tilde{\mathbf{y}}$, from the design settings $\tilde{\boldsymbol{X}}$, is conveniently a mixture distribution of the form

$$p(\tilde{\mathbf{y}}|\mathbf{y}) \propto \sum_\gamma p(\tilde{\mathbf{y}}|\boldsymbol{\gamma},\mathbf{y},\boldsymbol{X},\tilde{\boldsymbol{X}})p(\boldsymbol{\gamma}|\mathbf{y}) \tag{2.1.18}$$

where $p(\boldsymbol{\gamma}|\mathbf{y})$ is defined by (2.1.11) and $p(\tilde{\mathbf{y}}|\boldsymbol{\gamma},\mathbf{y},\boldsymbol{X},\tilde{\boldsymbol{X}})$ (Appendix 2.3.5) is

$$p(\tilde{\mathbf{y}}|\mathbf{y}) \propto \sum_\gamma p(\boldsymbol{\gamma}|\mathbf{y})\frac{(b^*/2)^{a^*/2}}{\Gamma(a^*/2)}\frac{\Gamma(a^*/2+m/2)}{(\sigma^2)^{m/2}|\mathbf{A}_\gamma|^{1/2}|\mathbf{A}_\gamma^*|^{1/2}}\left(\frac{1}{2}\left[\tilde{y}^T\tilde{y} + \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}_\gamma^{*T}\mathbf{A}_\gamma^*\hat{\boldsymbol{\beta}}_\gamma^* + ab\right]\right)^{-\frac{m+a+n}{2}} \tag{2.1.19}$$

where $A_\gamma^* = (\tilde{\boldsymbol{X}}_\gamma^T\tilde{\boldsymbol{X}}_\gamma + \mathbf{A}_\gamma^{-1})$ and

$$\hat{\boldsymbol{\beta}}_\gamma^{*T}\mathbf{A}_\gamma^*\hat{\boldsymbol{\beta}}_\gamma^* = (\tilde{\boldsymbol{X}}_\gamma^T\tilde{\mathbf{y}} + \boldsymbol{X}_\gamma^T\mathbf{y})^T(\tilde{\boldsymbol{X}}_\gamma^T\tilde{\boldsymbol{X}}_\gamma^T + \boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma + 1/\tau)^{-1}(\tilde{\boldsymbol{X}}_\gamma^T\tilde{\mathbf{y}} + \boldsymbol{X}_\gamma^T\mathbf{y}) \tag{2.1.20}$$

A simple approach is to generate a small Monte Carlo sample from (2.1.19) (Clyde and Parmigiani, 1998).

2.2 Shrinkage Priors

2.2.1 Penalized likelihood

In frequentist statistics, penalized likelihood methods are used to avoid explicit variable selection whilst inferring the set of active variables. They rely on the full model specification $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, where unnecessary variables are eliminated by determining which regression coefficient estimates are zero. The regularized solutions are obtained by constraining the set of admissible coefficient vectors, where the boundary optima possess the variable selection property. If $p > n$, some restrictions on the model solutions are required in order to guarantee problem determinacy. A range of attributes can be induced to reflect preferences on the solutions, these include sparsity, limited model size and smooth regression coefficients. The method of Lagrange multipliers is used to solve the constrained optimization, where the Lagrangian corresponds to the penalized log-likelihood function. The penalized log-likelihood problem in linear regression requires solving the optimisation

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^p}\left(-\frac{1}{2}\|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 - \sum_{j=1}^{p} \text{pen}_\lambda(|\beta_j|)\right), \tag{2.2.1}$$

where $\|\cdot\|$ denotes the l^2 norm and $\text{pen}_\lambda(\cdot)$ is the penalty function indexed by the regularization parameter $\lambda > 0$. As $\lambda \to 0$ the penalty term vanishes and the solution to (2.2.1) is ordinary least squares. There is large volume of statistical research of penalized likelihood approaches, producing intricate penalties motivated by arguments from asymptotic theory.

By optimising the penalized likelihood (2.2.1), the aim is to simultaneously perform variable selection (from the nonzero parameter estimates) and parameter estimation with as little bias as possible. This requires penalties which posses the variable selection property such as the ridge regression, $l^q_\lambda(|\beta_j|) = \lambda|\beta_j|^q$ for $q = 2$ (after imposing a threshold), and the least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996) $q = 1$, which has become one of the benchmark feature extraction methods. The value of the tuning parameter λ is often solved by optimising with respect to the predictive power of the model.

2.2.2 Bayesian regularization and variable selection

The expression in (2.2.1) can be interpreted as the log posterior density for $\boldsymbol{\beta}$, where the convex penalty term is determined by the choice of prior. Hence the penalized-likelihood solution can be interpreted as a posterior mode, where Bayesian regularization results from a choice of prior for $\boldsymbol{\beta}$ which is conditioned on the residual variance σ^2 and the penalty parameter λ. The conditioning on σ^2 is necessary in certain cases to obtain a unimodal posterior (Park and Casella, 2008). The λ parameter performs a similar role to the penalty parameter in classical penalized regression (2.2.1), penalizing the regression coefficient. It is this term which differentiates the model from Bayesian linear regression. Unlike the explicit prior approach, there is no prior over models or individual hypotheses $H_{0j} : \beta_j = 0$. Variable selection is performed via a posterior summarisation (mean or mode), which reduces some of the coefficients to zero. The coefficient λ needs to be large enough to penalize the coefficients β_j to zero, but not too large such that nonzero coefficients can be modeled. There are multiple options to specify the parameter, these are:

(1) A *fully Bayes* approach which treats λ as an unknown model parameter with a specified prior. This results in a solution which incorporates uncertainty about λ and results in a model which can be estimated in one step. A popular choice is the half-Cauchy $\lambda \sim$ half-Cauchy$(0, 1)$ (Gelman, 2006).

(2) An empirical Bayes method that estimates λ from the data, then fixes its value to estimate the model. The empirical Bayes estimate for λ is the maximum of marginal likelihood

$$\begin{aligned}\hat{\lambda} &= \arg\max_{\lambda} \left\{ \int \int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2, \lambda) d\boldsymbol{\beta}, d\sigma^2 \right\} \\ &= \arg\max_{\lambda} \left\{ p(\mathbf{y}|\boldsymbol{X}, \lambda) \right\}. \end{aligned} \tag{2.2.2}$$

Instead of directly optimizing (2.2.2) we can take advantage of the identity

$$p(\mathbf{y}|\boldsymbol{X}, \lambda) = \frac{p(\mathbf{y}, \boldsymbol{\beta}, \sigma^2|\boldsymbol{X}, \lambda)}{p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{X}, \lambda, \mathbf{y})}, \tag{2.2.3}$$

so that determining λ which maximises the marginal log likelihood is equivalent to maximising the augmented log likelihood $p(\mathbf{y}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}, \lambda)$ (proof in Appendix 4.9.1) in the same vein as the Expectation Maximisation (EM) algorithm (Section 4.9.1)

$$\begin{aligned} \lambda^{(k+1)} &= \arg\max_\lambda \left\{ \mathbb{E}\left(\log p(\mathbf{y}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}, \lambda) | \boldsymbol{X}, \lambda^{(k)}, \mathbf{y} \right) \right\} \\ &\approx \arg\max_\lambda \left\{ \frac{1}{M} \sum_{m=1}^{M} \log p(\mathbf{y}, \boldsymbol{\beta}^{(m)}, \sigma^{2(m)} | \boldsymbol{X}, \lambda) \right\}. \end{aligned} \tag{2.2.4}$$

This is much easier to compute, rather than integrating over β and σ^2 in the likelihood, the expectation (E-step) is taken with respect to the posterior distribution via a Monte Carlo version of the EM algorithm (which is the output from the MCMC sampler). The M-step maximises this expression over λ.

(3) *Cross-validation* (CV) is used to choose λ to minimise the estimated expected squared prediction error of a future observation (via an approximation)

$$\mathbb{E}[\mathbb{E}\{(y^f - \mathbf{x}^{fT} \hat{\boldsymbol{\beta}}_\lambda(\boldsymbol{X}, \mathbf{y}))^2 | (\boldsymbol{X}, \mathbf{y})\}], \tag{2.2.5}$$

where $(\mathbf{x}^f, y^f) \in \mathbb{R}^p \times \mathbb{R}$ is independent of $(\boldsymbol{X}, \mathbf{y})$ and has the same distribution as $(\mathbf{x}_1, y_1)$. The design matrix is treated as random and the dependence of $\hat{\boldsymbol{\beta}}_\lambda$ on the training data $(\boldsymbol{X}, \mathbf{y})$ is explicit.

2.2.3 Shrinkage priors

Bayesian lasso

The Bayesian lasso is the combination of an conditional Laplace prior (double exponential) on β_j and a Gaussian likelihood, and is analogous to $q = 1$ for (2.2.1). The marginal posterior mode of $\boldsymbol{\beta}$ performs the thresholding which selects the appropriate model (Park and Casella, 2008). The Laplace prior on β_j is equivalent to a scale mixture of Gaussian's with an exponential mixing

density (Appendix 2.3.6),

$$\beta_j|\sigma^2, \tau_j \sim N(0, \sigma^2\tau_j) \tag{2.2.6}$$

$$\tau_j \sim Ex(\lambda^2/2) \tag{2.2.7}$$

$$\begin{aligned} p(\beta_j|\sigma^2) &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\tau_j}} \exp\left(-\frac{1}{2\sigma^2\tau_j}\beta_j^2\right) \quad \frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2\tau_j}{2}\right)d\tau_j \\ &= \frac{\lambda}{2\sqrt{\sigma^2}}\exp\left(-\lambda|\beta_j|/\sqrt{\sigma^2}\right) \end{aligned} \tag{2.2.8}$$

The conditional posteriors from a fully Bayesian hierarchical approach are in Appendix 2.3.7.

The Bayesian lasso avoids computing marginal likelihoods and searching a model space. The global shrinkage parameter σ^2 controls the overall degree of sparsity in $\boldsymbol{\beta}$, where as the local shrinkage parameter τ_j acts to detect the signals. The penalty parameter λ takes the role of the complexity parameters in the frequentist lasso (Tibshirani 1994). However, the full posterior distribution under the Laplace prior does not contract at the same rate as its mode (Castillo et al., 2015), making uncertainty quantification under the Bayesian lasso unreliable.

An alternative specification of the lasso is proposed by Hans (2009) in terms of the normal-orthant distribution. Let $\mathcal{Z} = \{-1, 1\}^p$ represent the set of all 2^p possible p-vectors whose elements are ± 1. For any realisation $z \in \mathcal{Z}$, define the $\mathcal{O}_z \subset \mathbb{R}^r$. If $\boldsymbol{\beta} \in \mathcal{O}_z$, then $\beta_j \geq 0$ if $z_j = 1$ and $\beta_j < 0$ if $z_j = -1$. Then $\boldsymbol{\beta}$ follows the normal-orthant distribution with mean $\mathbf{m}$ and covariance $\mathbf{S}$, which is of the form

$$N^{[z]}(\boldsymbol{\beta}|\mathbf{m}, \mathbf{S}) = \frac{N_p(\boldsymbol{\beta}|\mathbf{m}, \mathbf{S})}{\Phi(\mathbf{m}, \mathbf{S})}\mathbf{I}_{(\boldsymbol{\beta} \in \mathcal{O}_z)}, \quad \Phi(\mathbf{m}, \mathbf{S}) = \int_{\mathcal{O}_z} N_p(\mathbf{t}|\mathbf{m}, \mathbf{S})d\mathbf{t}. \tag{2.2.9}$$

The prior parameterisation of Hans (2009) is

$$\boldsymbol{\beta}|\lambda, \sigma^2 \sim \left(\frac{\lambda}{2\sqrt{\sigma^2}}\right)^p \exp\left(-\lambda\sum_{j=1}^p \frac{|\beta_j|}{\sqrt{\sigma^2}}\right) \tag{2.2.10}$$

$$\lambda \sim Gamma(r, \delta). \tag{2.2.11}$$

Using the definition of the normal-orthant distribution, the conditional posterior of $\boldsymbol{\beta}$ is a mixture of normal-orthant distributions of the

$$\beta_j|\mathbf{y}, \boldsymbol{\beta}_{-j}, \sigma^2, \lambda \sim \phi_j N^{[+]}\left(\mu_j^+, \omega_{jj}^{-1}\right) + (1-\phi_j) N^{[-]}\left(\mu_j^-, \omega_{jj}^{-1}\right) \tag{2.2.12}$$

where

- $N^{[-]}$ and $N^{[+]}$ are the $N^{[z]}$ distribution for $z=-1$ and $z=1$ respectively,
- $\mu_j^{[+]} = \hat{\beta}_j^{ols} + \{\sum_{i\neq j}(\hat{\beta}_i^{ols} - \beta_i)(\omega_{ij}/\omega_{jj})\} + \left(-\frac{\lambda}{\sqrt{\sigma^2}\omega_{jj}}\right)$,
- ω_{ij} is the ij off-digaonal element of the matrix $\Omega = \Sigma^{-1} = \left(\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)^{-1}$,
- $\phi_j = \frac{\Phi\left(\frac{\mu_j^+}{\sqrt{\omega_{jj}}}\right)/N(0|\mu_j^+, \omega_{jj}^{-1})}{\Phi\left(\frac{\mu_j^+}{\sqrt{\omega_{jj}}}\right)/N(0|\mu_j^+, \omega_{jj}^{-1}) + \Phi\left(\frac{\mu_j^-}{\sqrt{\omega_{jj}}}\right)/N(0|\mu_j^-, \omega_{jj}^{-1})}$.

Both prediction and point estimation are performed via the posterior mean. The conditional posterior of σ^2 is not of standard form and can not be sampled directly. Hans (2009) suggests an accept/reject step to generate approximate samples from the posterior.

The elastic net combines the benefits of the lasso (ℓ_1 penalization) with ridge regression (ℓ_2 penalization) by solving the problem

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2 \tag{2.2.13}$$

with two tuning parameters. The Bayesian prior that provides the solution to the elastic net estimation problem is of the form

$$\boldsymbol{\beta}|\sigma^2 \propto \exp\left(-\frac{1}{2\sigma^2}\left(\lambda_1\sum_{j=1}^p |\beta_j| + \lambda_2\sum_{j=1}^p \beta_j^2\right)\right). \tag{2.2.14}$$

The generalized double Pareto prior (Armagan et al., 2013), introduces adaptive penalties for

each β_j coefficient by having p exponential mixing densities

$$\tau_j|\lambda_j \sim Ex(\lambda_j^2/2). \tag{2.2.15}$$

This results in the generalized double Pareto distribution prior on $\boldsymbol{\beta}$

$$\boldsymbol{\beta}|\sigma \sim \prod_{j=1}^{p} \frac{1}{2\sigma\delta/r}\left(1+\frac{1}{r}\frac{|\beta_j|}{\sigma\delta/r}\right)^{-(r+1)}, \tag{2.2.16}$$

which has a spike at zero with Student's t-like heavy tails. The conditional posteriors are in Appendix 2.3.8.

Other popular extensions to the lasso include the group lasso (Xu and Ghosh, 2015) that allows for group shrinkage, the fused lasso (Betancourt et al., 2017) that allows for spatial or temporal relationships between neighbouring parameters, and the adaptive lasso (Leng et al., 2014) that addresses variable selection consistency issues with the regular lasso.

Beyond the Bayesian lasso

The Bayesian lasso can be generalised with different densities for the mixture distribution such as a gamma, inducing a regularisation penalty which is a function of $|\beta_j|$ to maintain the property of zeroing the regression coefficients (Griffin and Brown, 2005) through the modal estimate of a multimodal posterior.

Furthermore, the lasso can be considered within an even more general framework of penalisation functions, referred to as the "one group answer" in the sparse regression context, extending the possibility of mixing densities for the scale mixture of Gaussian's. The global-local scale mixture

framework Polson and Scott (2011) is defined as

$$\beta_j|\lambda^2, \tau_j \sim N(0, \lambda^2\tau_j^2), \qquad j = 1, ..., p \tag{2.2.17}$$

$$\tau_j^2 \sim F_\tau(a, b) \tag{2.2.18}$$

$$\lambda^2 \sim F_\lambda(c, d) \tag{2.2.19}$$

where λ is a global shrinkage parameter (analogous to the regularisation penalty in (2.2.1), applying the same shrinkage to the whole vector $\boldsymbol{\beta}$) and τ_j is a local shrinkage parameter (only applying shrinkage to β_j).

Polson and Scott (2011) establish a criteria to evaluate the appropriate choice of mixture prior in the presence of sparseness "global local shrinkage rules", by trying to replicate the behaviour of explicit variable selection. With the aim of balancing the trade-off between shrinking the noise towards zero whilst leaving the large signals unshrunk, the framework identifies the horseshoe prior as a superior alternative to the lasso. Unlike the traditional Bayesian lasso, the horseshoe prior also benefits from thresholding with the expectation, the minimum mean squared error estimator.

The horseshoe prior defines the local shrinkage parameter τ_j as a standard half-Cauchy distribution $C^+(0, 1)$ on the positive reals, which has an infinitely tall spike at 0 and heavy Cauchy-like tails, Figure 2.2.1. The same half-Cauchy distribution is posited on the global shrinkage parameter λ. Its name reflects the shape of the probability density for the implied shrinkage parameter κ_j (Carvalho et al., 2010). Expressing the expectation of the marginal posterior (where λ and σ^2 are fixed at 1) as

$$\begin{aligned}\hat{\beta}_j = \mathbb{E}(\beta_j|\mathbf{y}) &= \int_0^\infty \left(1 - \frac{1}{1+\tau_j^2}\right) y_i \; p(\tau_j|\mathbf{y}) \; d\tau_j \\ &= 1 - \mathbb{E}\left(\frac{1}{1+\tau_j^2}\middle|\mathbf{y}\right) \cdot y_i,\end{aligned}$$

$\kappa_j = 1/(1+\tau_j^2)$ can be interpreted as a random shrinkage parameter, analogous to the inclusion probability ω_j in the discrete mixture of the explicit variable selection. By reparameterising the prior in terms of an implied shrinkage parameter κ_j (Appendix 2.3.9), the density of the horseshoe

prior (π_H) is $p(\kappa_j) \propto \kappa_j^{-1/2}(1-\kappa_j)^{-1/2}$, Figure 2.2.2. π_H is unbounded at both $\kappa_j = 0$ and $\kappa_j = 1$, implying large outlying β_j's will not be shrunk ($\kappa_i \approx 0$), but the remaining β_j's will have $\kappa_j \approx 1$ a posteriori and can be shrunk almost all the way to zero.

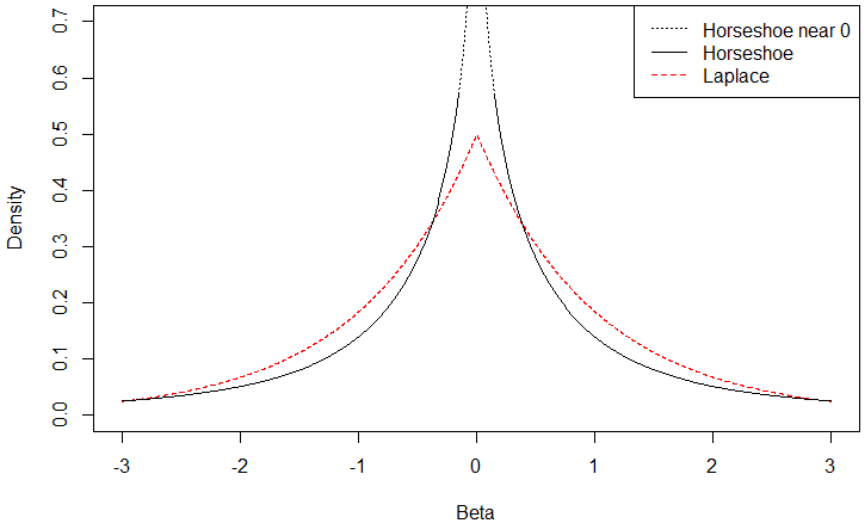

Figure 2.2.1: Plot of the marginal prior probability density of β_j from the Horseshoe prior $\tau_j \sim C^+(0,1)$ and the exponential prior (Lasso) $\tau_j \sim Exp(2)$ mixing densities.

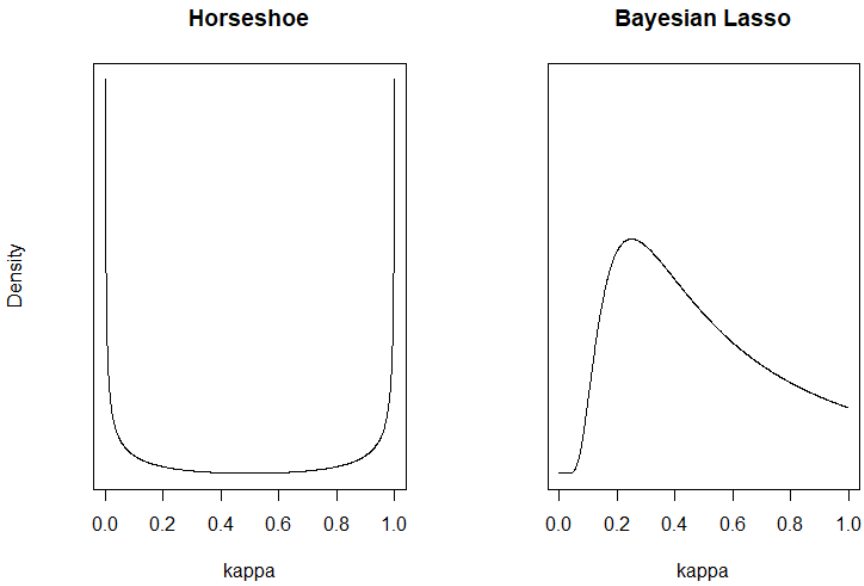

Figure 2.2.2: Comparison of the implied density for the shrinkage weights $\kappa_j \in [0,1]$ for the Bayesian Lasso π_{BL} and the Horseshoe prior π_H, where $\kappa_j = 0$ means no shrinkage and $\kappa_j = 1$ means total shrinkage to zero.

Since its inception, there have been numerous applications of this prior in statistics and machine learning, such as deep neural networks (Ghosh et al., 2018), deep generalized linear models (Tran et al., 2020) and nonparametric function estimation (Shin et al., 2020).

2.2.4 Computational challenges

Continuous shrinkage priors avoid the combinatorial search required when searching the model space with explicit variable selection. The difficulty with sampling from the model space in high dimensions (as discussed in Section 2.1.3) is in fully exploring the large binary space denoting whether a parameter is zero versus non-zero, which incurs extreme computational cost.

The posterior conditionals from the shrinkage priors are easy to derive because of the conditional structure of the model, and the combination of Gaussian likelihood and Gaussian $\boldsymbol{\beta}$ prior. A Gibbs sampler will cycle through the distributions, until a large sample from the posterior of each parameter is available. However, as these methods are often needed in high dimensions, the Gibbs sampler can become computationally costly. The conditional structure of the hierarchical priors implies dependence between the parameters which may lead to slow mixing and convergence to the desired posterior. This is particularly evident with the Horseshoe prior, leading to the proposition of more efficient slice sampling schemes (Makalic and Schmidt (2016) and Johndrow et al. (2020)).

The most cumbersome step is the sampling of the p-variate normal conditional posterior distribution of $\boldsymbol{\beta}$. This requires an inversion of the full design matrix in the form $\boldsymbol{V}^{-1} = (\boldsymbol{X}^T\boldsymbol{X}c + \boldsymbol{D})$ where c is a constant and $\boldsymbol{D}$ is a diagonal matrix (which is a function of the hyper-parameters). The presence of the hyper-parameters in $\boldsymbol{V}$ means that this matrix changes at each iteration of the sampler, preventing the data matrix $\boldsymbol{X}^T\boldsymbol{X}$ just being inverted before the start of the sampler. The Cholesky decomposition of $\boldsymbol{V}$ is performed to sample from the desired normal distribution. Whilst the step can be sped up, the decomposition of a $p \times p$ matrix is of $\mathcal{O}(p^3)$ complexity. This inversion is avoided with a spike-and-slab prior of the form (2.1.4), which gains from subsetting the design matrix to $\boldsymbol{X}_{n\times p_\gamma}$ providing large computational savings, particularly in the presence of sparsity.

A precision based sampler (Rue, 2002) can be used to obtain samples efficiently from the conditional normal posterior distribution $\boldsymbol{\beta}|\mathbf{y}, \cdot \sim N(\boldsymbol{V}\boldsymbol{X}^T\mathbf{y}, \boldsymbol{V})$. The approach combines samples from a standard multivariate normal and the Cholesky factorisation of $\boldsymbol{V}^{-1}$;

- Compute the Cholesky factorisation $\boldsymbol{V}^{-1} = \boldsymbol{L}^T\boldsymbol{L}$.
- Generate $\boldsymbol{Z} \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$.
- Set $\boldsymbol{\beta} = (\boldsymbol{L}^T)^{-1}(\boldsymbol{L}^{-1}\boldsymbol{X}^T\mathbf{y} + \boldsymbol{Z})$.

This approach can achieve high efficiency gains when the Gram matrix $\boldsymbol{X}^T\boldsymbol{X}$ is block-diagonal, assuming the prior variance $\boldsymbol{D}$ has a similar structure. The main feature of this algorithm is the requirement to invert the Cholesky factor of $\boldsymbol{V}^{-1}$, rather than the matrix $\boldsymbol{V}^{-1}$.

Bhattacharya et al. (2016) exploit the Woodbury matrix inversion lemma to generate normal variates for $\boldsymbol{\beta}|\mathbf{y}, \cdot \sim N(\boldsymbol{V}\boldsymbol{X}^T\mathbf{y}, \boldsymbol{V})$ when $p >> n$. Their algorithm requires inversion of an $n \times n$ matrix $(\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T + \boldsymbol{I}_n)$, rather than inverting the $p \times p$ matrix $\boldsymbol{V}$. Uncorrelated normal draws are generated from two diagonal covariance matrices, rather than the full covariance matrix $\boldsymbol{V}$. The worst-case complexity $\mathcal{O}(n^2p)$, is linear in p achieving savings when $p >> n$.

For ultra high-dimensional data with very large p, computation of $(\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T + \boldsymbol{I}_n)^{-1}$ remains expensive. Johndrow et al. (2020) approximate the approach of Bhattacharya et al. (2016) by reducing the dimensions of this matrix inversion via thresholding, effectively changing the sampler to combinatorial search, in a very similar fashion to the spike-and-slab prior.

2.3 Appendix

This section contains the derivations to the expressions referred to in Chapter 2.

2.3.1 Explicit variable selection - Joint update

There are two approaches to deriving the the joint update of β_j, γ_j. In the Gibbs sampler we wish to sample from $p(\gamma_j|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}_{-j}, \omega, \sigma^2_\beta, \mathbf{y})$ and then $p(\beta_j|\gamma_j, \boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}_{-j}, \omega, \sigma^2_\beta, \mathbf{y})$. Rather than integrating out β_j in the full conditional of γ_j we can define

$$\begin{aligned} p(\gamma_j = 1|\gamma_{-j}, \boldsymbol{\beta}_{-j}, \omega, \sigma^2_\beta, \mathbf{y}) &= p(\gamma_j = 1|\boldsymbol{\vartheta}, \mathbf{y}) \\ &= \frac{p(\gamma_j = 1, \beta_j = 0|\boldsymbol{\vartheta}, y)}{p(\beta_j = 0|\gamma_j = 1, \boldsymbol{\vartheta}, \mathbf{y})}. \end{aligned} \tag{2.3.1}$$

We multiply both sides of the equation by $p(\beta_j = 0|\gamma_j = 1, \boldsymbol{\vartheta}, \mathbf{y})$ and use $\boldsymbol{\vartheta}$ to denote all the other parameters except β_j. Expanding the expression, form the definition of joint probability and using proportionality

$$\begin{aligned} p(\gamma_j = 1|\boldsymbol{\vartheta}, \mathbf{y}) &= \frac{p(\boldsymbol{\vartheta}, \mathbf{y}|\gamma_j = 1, \beta_j = 0)p(\gamma_j = 1, \beta_j = 0)}{p(\boldsymbol{\vartheta}, \mathbf{y})p(\beta_j = 0|\gamma_j = 1, \boldsymbol{\vartheta}, \mathbf{y})} \\ &\propto \frac{p(\boldsymbol{\vartheta}, \mathbf{y}|\gamma_j = 1, \beta_j = 0)p(\gamma_j = 1, \beta_j = 0)}{p(\beta_j = 0|\gamma_j = 1, \boldsymbol{\vartheta}, \mathbf{y})}. \end{aligned}$$

As $p(\boldsymbol{\vartheta}, \mathbf{y}|\gamma_j = 1, \beta_j = 0) = p(\boldsymbol{\vartheta}, \mathbf{y}|\beta_j = 0)$ since γ_j is irrelevant once we condition on β_j we have

$$\begin{aligned} p(\gamma_j = 1|\boldsymbol{\vartheta}, \mathbf{y}) &\propto \frac{p(\boldsymbol{\vartheta}, y|\beta_j = 0)p(\gamma_j = 1, \beta_j = 0)}{p(\beta_j = 0|\gamma_j = 1, \boldsymbol{\vartheta}, \mathbf{y})} \\ &\propto \frac{p(\gamma_j = 1, \beta_j = 0)}{p(\beta_j = 0|\gamma_j = 1, \boldsymbol{\vartheta}, \mathbf{y})} \\ &\propto \frac{p(\beta_j = 0|\gamma_j = 1)p(\gamma_j = 1)}{p(\beta_j = 0|\gamma_j = 1, \boldsymbol{\vartheta}, \mathbf{y})}. \end{aligned}$$

As $p(\gamma_j = 1) = 1 - \omega$ and $p(\beta_j = 0|\gamma_j = 1, \boldsymbol{\vartheta}, \mathbf{y}) = N(0|m, v)$ the normalised probability is

$$p(\gamma_j = 1|\mathbf{y}, \boldsymbol{\vartheta}) = \frac{\frac{\phi(0|0,\sigma^2_\beta)}{\phi(0|m,v)}\omega}{(1-\omega) + \frac{\phi(0|0,\sigma^2_\beta)}{\phi(0|m,v)}\omega}, \tag{2.3.2}$$

where ϕ denotes the normal pdf.

The alternative is to find the update using proportionality from the joint probability

$$\log p(\beta_j, \gamma_j | \mathbf{y}, \boldsymbol{\vartheta}) \propto -\frac{1}{2\sigma^2}\left(\left\|\mathbf{y} - \sum_s X_s\gamma_s\beta_s\right\|^2\right) - \frac{\gamma_j}{2}\log\left(2\pi\sigma_\beta^2\right) - \frac{\gamma_j\beta_j^2}{2\sigma_\beta^2} + (1-\gamma_j)\log\delta_0(\beta_j) + \gamma_j\log(\omega) + (1-\gamma_j)\log(1-\omega). \tag{2.3.3}$$

Completing the square, exponentiating and rearranging gives

$$N(\beta_j|m,v)^{\gamma_j}\delta_0(\beta_j)^{1-\gamma_j}\left\{\exp\left(\frac{m^2}{2v} + \frac{\log(v)}{2} + \log(\omega) - \frac{\log\left(\sigma_\beta^2\right)}{2}\right)\right\}^{\gamma_j}(1-\omega)^{1-\gamma_j}, \tag{2.3.4}$$

where

$$m = vX_j^T\left(\mathbf{y} - \sum_{s\neq j} X_s\beta_s\right) \qquad v = \left(\frac{||X_j||^2}{\sigma^2} + \frac{1}{\sigma_\beta^2}\right). \tag{2.3.5}$$

Normalising the probabilities for γ_j gives the same answer as (2.3.2).

2.3.2 Explicit variable selection - Marginal likelihood

The expressions used in the discussion of the marginal likelihood of $\boldsymbol{\gamma}$, and all derivations in the following sections in the Appendix are from the initial parameterisation

$$\begin{aligned}
\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2 &\sim N(\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma, \sigma^2) \\
\beta_j|\sigma^2, \boldsymbol{\gamma} &\sim N(0, \sigma^2\tau)^{\gamma_j}\delta_0(\beta_j)^{1-\gamma_j} \\
\boldsymbol{\gamma} &\sim \prod_{j=1}^{p}\omega^{\gamma_j}(1-\omega)^{1-\gamma_j} \\
\sigma^2 &\sim IG(a/2, ab/2),
\end{aligned} \tag{2.3.6}$$

where τ is treated as a tuning parameter and p_γ is the number of covariates in the model. The joint distribution of $\mathbf{y}$ and $\boldsymbol{\beta}$ is

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2, \tau) =& p(\mathbf{y}|\boldsymbol{\beta}_\gamma, \sigma^2) p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2, \tau) \\ =& \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\bigg(-\frac{1}{2\sigma^2} (\mathbf{y} - \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma)^T (\mathbf{y} - \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma) \bigg) \frac{1}{(2\pi\sigma^2\tau)^{p_\gamma/2}} \exp\bigg(-\frac{1}{2\sigma^2\tau} \boldsymbol{\beta}_\gamma^T \boldsymbol{\beta}_\gamma \bigg) \\ =& \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{(2\pi\sigma^2\tau)^{p_\gamma/2}} \exp\Big(-\frac{1}{2} \Big(\sigma^{-2} \boldsymbol{\beta}_\gamma \boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma + \sigma^{-2}\tau^{-1} \boldsymbol{\beta}_\gamma^T \boldsymbol{\beta}_\gamma + \\ & - 2\sigma^{-2} \mathbf{y}^T \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma + \sigma^{-2} \mathbf{y}^T \mathbf{y} \Big) \Big). \end{aligned} \tag{2.3.7}$$

After completing the square

$$p(\mathbf{y}, \boldsymbol{\beta}|\cdot) = \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{(2\pi\sigma^2\tau)^{p_\gamma/2}} \exp\bigg(-\frac{1}{2\sigma^2} \Big((\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^T A_\gamma (\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma) - \hat{\boldsymbol{\beta}}_\gamma^T \mathbf{A}_\gamma \hat{\boldsymbol{\beta}}_\gamma + \mathbf{y}^T \mathbf{y} \Big) \bigg) \tag{2.3.8}$$

where

$$\mathbf{A}_\gamma = \boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma + \frac{1}{\tau} \qquad \hat{\boldsymbol{\beta}}_\gamma = \mathbf{A}_\gamma^{-1} (\mathbf{y}^T \boldsymbol{X}_\gamma)^T \tag{2.3.9}$$

to obtain

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2, \tau) =& \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{(2\pi\sigma^2\tau)^{p_\gamma/2}} \exp\bigg(-\frac{1}{2\sigma^2} \Big((\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^T A_\gamma (\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma) \Big) \bigg) \\ & \exp\bigg(-\frac{1}{2\sigma^2} \Big(\mathbf{y}^T \mathbf{y} - \hat{\boldsymbol{\beta}}_\gamma^T \mathbf{A}_\gamma \hat{\boldsymbol{\beta}}_\gamma \Big) \bigg) \end{aligned} \tag{2.3.10}$$

Integrating over $\boldsymbol{\beta}_\gamma$

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\gamma}, \sigma^2, \tau) =& \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{(\tau)^{p_\gamma/2}} \exp\bigg(-\frac{1}{2\sigma^2} \Big(\mathbf{y}^T \mathbf{y} - \hat{\boldsymbol{\beta}}_\gamma^T \mathbf{A}_\gamma \hat{\boldsymbol{\beta}}_\gamma \Big) \bigg) |A|^{-1/2} \\ & \int \frac{|A|^{1/2}}{(2\pi\sigma^2)^{p_\gamma/2}} \exp\bigg(-\frac{1}{2\sigma^2} \Big((\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^T \mathbf{A}_\gamma (\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma) \Big) \bigg) d\boldsymbol{\beta}_\gamma \end{aligned}$$

gives

$$p(\mathbf{y}|\boldsymbol{\gamma}, \sigma^2, \tau) = \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{(\tau)^{p_\gamma/2}} \exp\bigg(-\frac{1}{2\sigma^2} \Big(\mathbf{y}^T \mathbf{y} - \hat{\boldsymbol{\beta}}_\gamma^T \mathbf{A}_\gamma \hat{\boldsymbol{\beta}}_\gamma \Big) \bigg) |\mathbf{A}|^{-1/2}. \tag{2.3.11}$$

The joint distribution $p(y, \sigma^2|\boldsymbol{\gamma}, \tau)$ is then

$$\begin{aligned} p(\mathbf{y}, \sigma^2|\boldsymbol{\gamma}, \tau) =& p(\mathbf{y}|\boldsymbol{\gamma}, \tau, \sigma^2)p(\sigma^2) \\ =& \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{(\tau)^{p_\gamma/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}_\gamma^T \mathbf{A}_\gamma \hat{\boldsymbol{\beta}}_\gamma\right)\right) |\mathbf{A}_\gamma|^{-1/2} \frac{(ab/2)^{a/2}}{\Gamma(a/2)} (\sigma^2)^{-\frac{a}{2}-1} \\ & \exp\left(-\frac{ab}{2\sigma^2}\right) \end{aligned}$$

where we see the importance of parameterising the variance of β_j relative to σ^2 in Equation(2.1.10). Rearranging and marginalising over σ^2 gives the marginal likelihood of

$$p(\mathbf{y}|\boldsymbol{\gamma}, \tau) = \frac{(ab/2)^{a/2}\Gamma((n+a)/2)}{\Gamma(a/2)} \frac{1}{(2\pi)^{n/2}\tau^{p_\gamma/2}|\mathbf{A}_\gamma|^{1/2}} \left(\frac{1}{2}\left[\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\boldsymbol{X}_\gamma\mathbf{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\mathbf{y} + ab\right]\right)^{-\frac{n+a}{2}}. \tag{2.3.12}$$

2.3.3 Explicit variable selection - Marginal model posterior

The marginal posterior $p(\boldsymbol{\gamma}|\mathbf{y})$ is thus proportional to

$$p(\boldsymbol{\gamma}|\mathbf{y}) \propto \frac{1}{\tau^{p_\gamma/2}|\mathbf{A}_\gamma|^{1/2}} \left(\frac{1}{2}\left[\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\boldsymbol{X}_\gamma\mathbf{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\mathbf{y} + ab\right]\right)^{-\frac{n+a}{2}} p(\boldsymbol{\gamma}). \tag{2.3.13}$$

We can also obtain the conditional posterior $p(\gamma_j|\boldsymbol{\gamma}_{-j}, \mathbf{y})$

$$p(\gamma_j = 1|\boldsymbol{\gamma}_{-j}, \mathbf{y}) = \frac{p(\mathbf{y}|\gamma_j = 1, \boldsymbol{\gamma}_{-j})p(\gamma_j = 1|\boldsymbol{\gamma}_{-j})}{p(\mathbf{y}|\gamma_j = 0, \boldsymbol{\gamma}_{-j})p(\gamma_j = 0|\boldsymbol{\gamma}_{-j}) + p(\mathbf{y}|\gamma_j = 1, \boldsymbol{\gamma}_{-j})p(\gamma_j = 1|\boldsymbol{\gamma}_{-j})} \tag{2.3.14}$$

by multiplying both sides by $p(\mathbf{y}|\gamma_j = 0, \boldsymbol{\gamma}_{-j})$ to get

$$p(\gamma_j = 1|\boldsymbol{\gamma}_{-j}, \mathbf{y}) = \frac{F(\boldsymbol{\gamma}, \boldsymbol{\gamma}')p(\gamma_j = 1|\boldsymbol{\gamma}_{-j})}{p(\gamma_j = 0|\boldsymbol{\gamma}_{-j}) + F(\boldsymbol{\gamma}, \boldsymbol{\gamma}')p(\gamma_j = 1|\boldsymbol{\gamma}_{-j})} \tag{2.3.15}$$

where

$$F(\boldsymbol{\gamma}, \boldsymbol{\gamma}') = \frac{p(\mathbf{y}|\gamma_i = 1, \boldsymbol{\gamma}_{-j})}{p(\mathbf{y}|\gamma_j = 0, \boldsymbol{\gamma}_{-j})} \tag{2.3.16}$$

using the marginal likelihood obtained in Equation (2.3.12).

2.3.4 Explicit variable selection - Marginal coefficient posterior

We can obtain the marginal posterior distribution for a given model. This is particular useful when we use MCMC to search over $p(\gamma_j|\boldsymbol{\gamma}_{-j}, \mathbf{y})$. The joint distribution $p(\mathbf{y}, \boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2, \tau)$ in (2.3.10) is

$$p(\mathbf{y}, \boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2, \tau) = \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{(2\pi\sigma^2\tau)^{p_\gamma/2}} \exp\left(-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^T A_\gamma (\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma) + \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}_\gamma^T \mathbf{A}_\gamma \hat{\boldsymbol{\beta}}_\gamma\right]\right)$$

Marginalising over σ^2 after multiplying by the prior

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\beta}|\sigma^2, \tau) &= \int_{\sigma^2} p(\mathbf{y}, \boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma}, \tau) p(\sigma^2) d\sigma^2 \\ &= \int_{\sigma^2} \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{(2\pi\sigma^2\tau)^{p_\gamma/2}} \frac{(\frac{ab}{2})^{a/2}}{\Gamma(a/2)} (\sigma^2)^{-\frac{a}{2}-1} \exp\left(-\frac{ab}{2\sigma^2}\right) \\ &\quad \exp\left(-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^T A_\gamma (\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma) + \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}_\gamma^T \mathbf{A}_\gamma \hat{\boldsymbol{\beta}}_\gamma\right]\right) d\sigma^2 \end{aligned}$$

gives

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\beta}|\boldsymbol{\gamma}, \tau) =& (2\pi)^{-\frac{n}{2}} (2\pi)^{-\frac{p_\gamma}{2}} (\tau)^{-\frac{p_\gamma}{2}} \frac{(\frac{ab}{2})^{\frac{a}{2}}}{\Gamma(\frac{a}{2})} \Gamma\left(\frac{n + p_\gamma + a}{2}\right) \\ & \left[\frac{1}{2}\left((\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^T A_\gamma (\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma) + \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}_\gamma^T \mathbf{A}_\gamma \hat{\boldsymbol{\beta}}_\gamma + ab\right)\right]^{-\frac{n+p_\gamma+a}{2}}. \end{aligned}$$

Expanding the terms in the square parenthesis and noting that the term
$C_\gamma = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T \boldsymbol{X}_\gamma \mathbf{A}_\gamma^{-1} \boldsymbol{X}_\gamma^T \mathbf{y}$, is a scalar

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\beta}|\boldsymbol{\gamma}, \tau) =& (\pi)^{-\frac{n}{2}} (\pi)^{-\frac{p_\gamma}{2}} (\tau)^{-\frac{p_\gamma}{2}} \frac{(ab)^{\frac{a}{2}}}{\Gamma(\frac{a}{2})} \Gamma\left(\frac{n + p_\gamma + a}{2}\right) (C_\gamma + ab)^{-\frac{n+p_\gamma+a}{2}} \\ & \left[1 + \frac{1}{(n+a)} (\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^T (n+a)(C_\gamma + ab)^{-1} \mathbf{A}_\gamma (\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)\right]^{-\frac{n+p_\gamma+a}{2}}. \end{aligned}$$

Dividing this expression by $p(\mathbf{y}|\boldsymbol{\gamma},\tau)$ in (2.3.12) gives the marginal posterior $p(\boldsymbol{\beta}|\boldsymbol{\gamma},\tau)$ for a particular model. To be able to identify the distributional form, $p(\mathbf{y},\boldsymbol{\beta}|\boldsymbol{\gamma},\tau)$ is augmented with additional terms

$$\begin{aligned} p(\mathbf{y},\boldsymbol{\beta}|\boldsymbol{\gamma},\tau) =& (\pi)^{-\frac{n}{2}}(\pi)^{-\frac{p_\gamma}{2}}(\tau)^{-\frac{p_\gamma}{2}}\frac{(ab)^{\frac{a}{2}}|\mathbf{A}_\gamma|^{-\frac{1}{2}}\Gamma(\frac{n+a}{2})}{\Gamma(\frac{a}{2})}\Gamma\left(\frac{n+p_\gamma+a}{2}\right)(C_\gamma+ab)^{-\frac{n+a}{2}} \\ & \frac{1}{\Gamma(\frac{n+a}{2})(n+a)^{\frac{p_\gamma}{2}}}\left|\frac{(C_\gamma+ab)}{(n+a)}\mathbf{A}_\gamma^{-1}\right|^{-\frac{1}{2}} \\ & \left(1+\frac{1}{(n+a)}(\boldsymbol{\beta}_\gamma-\hat{\boldsymbol{\beta}}_\gamma)^T(n+a)(C_\gamma+ab)^{-1}\mathbf{A}_\gamma(\boldsymbol{\beta}_\gamma-\hat{\boldsymbol{\beta}}_\gamma)\right)^{-\frac{n+p_\gamma+a}{2}}. \end{aligned}$$

We have

$$p(\mathbf{y}|\boldsymbol{\gamma},\tau) = 2^{-\frac{a+n}{2}}\frac{(ab)^{a/2}\Gamma((n+a)/2)}{\Gamma(a/2)}\frac{1}{(\pi)^{n/2}\tau^{p_\gamma/2}|\mathbf{A}_\gamma|^{1/2}}\left(\frac{1}{2}\left[C_\gamma+ab\right]\right)^{-\frac{n+a}{2}}, \tag{2.3.17}$$

which gives

$$\frac{p(\mathbf{y},\boldsymbol{\beta}|\boldsymbol{\gamma},\tau)}{p(\mathbf{y}|\boldsymbol{\gamma},\tau)} = \frac{\Gamma\left(\frac{n+p_\gamma+a}{2}\right)}{(\pi)^{\frac{p_\gamma}{2}}\Gamma(\frac{n+a}{2})(n+a)^{\frac{p_\gamma}{2}}}\left|\frac{(C_\gamma+ab)}{(n+a)}\mathbf{A}_\gamma^{-1}\right|^{-\frac{1}{2}} \tag{2.3.18}$$

$$\left[1+\frac{1}{(n+a)}(\boldsymbol{\beta}_\gamma-\hat{\boldsymbol{\beta}}_\gamma)^T(n+a)(C_\gamma+ab)^{-1}\mathbf{A}_\gamma(\boldsymbol{\beta}_\gamma-\hat{\boldsymbol{\beta}}_\gamma)\right]^{-\frac{n+p_\gamma+a}{2}}, \tag{2.3.19}$$

the multivariate t-distribution with mean $\mathbf{A}_\gamma^{-1}\boldsymbol{X}_\gamma^T\mathbf{y}$ and covariance $\frac{1}{n+a}(C_\gamma+ab)\mathbf{A}_\gamma^{-1}$.

Where $\mathbf{A}_\gamma$ and $\hat{\boldsymbol{\beta}}_\gamma$ are defined in (2.3.9) as

$$\mathbf{A}_\gamma = \boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma+\frac{1}{\tau} \qquad \hat{\boldsymbol{\beta}}_\gamma = \mathbf{A}_\gamma^{-1}(\mathbf{y}^T\boldsymbol{X}_\gamma)^T \tag{2.3.20}$$

2.3.5 Explicit variable selection - Posterior predictive

The derivation of the posterior predictive comes from the parameterisation in (2.3.6). The conditional posterior distribution for $\boldsymbol{\beta}$

$$\boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2, \mathbf{y} \sim N_{p_\gamma}(\hat{\boldsymbol{\beta}}_\gamma, \sigma^2 \mathbf{A}_\gamma^{-1}) \tag{2.3.21}$$

where

$$\hat{\boldsymbol{\beta}}_\gamma = \left(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma + \frac{1}{\tau}\right)^{-1} \boldsymbol{X}_\gamma^T \mathbf{y} \qquad \mathbf{A}_\gamma = \left(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma + \frac{1}{\tau}\right) \tag{2.3.22}$$

which is taken from identifying the normal kernal from the joint posterior $p(y, \boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2)$ in (2.3.8).

The σ^2 marginal posterior distribution is

$$\sigma^2|\boldsymbol{\gamma}, \mathbf{y} \sim IG\left(a^*/2, b^*/2\right) \tag{2.3.23}$$

with

$$a^* = a + n \qquad b^* = \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}_\gamma^T \mathbf{A}_\gamma \hat{\boldsymbol{\beta}}_\gamma + ab. \tag{2.3.24}$$

The posterior predictive for a vector of new observations $\tilde{\mathbf{y}}$ (dimension m) from the design matrix $\tilde{\boldsymbol{X}}$ can be found by repeating the steps outlined in the derivation of the marginal likelihood with respect to the posterior distributions.

Integrating the marginal likelihood with respect to $p(\boldsymbol{\beta}|\sigma^2, \gamma, \mathbf{y})$

$$\int_\beta p(\tilde{\mathbf{y}}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2) p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2, \mathbf{y}) d\boldsymbol{\beta} = \int_\beta \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{1}{2\sigma^2}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{X}}_\gamma \boldsymbol{\beta}_\gamma)^T(\tilde{\mathbf{y}} - \tilde{\boldsymbol{X}}_\gamma \boldsymbol{\beta}_\gamma)\right)$$
$$\frac{1}{(2\pi\sigma^2)^{p_\gamma/2}|\mathbf{A}_\gamma|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)^T \mathbf{A}_\gamma (\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma)\right) d\boldsymbol{\beta}$$

setting

$$\hat{\boldsymbol{\beta}}_\gamma^* = (\tilde{\boldsymbol{X}}_\gamma^T \tilde{\boldsymbol{X}}_\gamma + \mathbf{A}_\gamma)^{-1}(\tilde{\boldsymbol{X}}_\gamma^T \tilde{\mathbf{y}} + \hat{\boldsymbol{\beta}}_\gamma^T \mathbf{A}_\gamma) \qquad \mathbf{A}_\gamma^* = (\tilde{\boldsymbol{X}}_\gamma^T \tilde{\boldsymbol{X}}_\gamma + \mathbf{A}_\gamma) \tag{2.3.25}$$

by rearranging, completing the square and integrating over $\boldsymbol{\beta}_\gamma$,

$$p(\tilde{\mathbf{y}}|\sigma^2,\boldsymbol{\gamma}) = \frac{1}{(\sigma^2)^{m/2}|\mathbf{A}_\gamma|^{1/2}|A^*_\gamma|^{1/2}} \exp\bigg(-\frac{1}{2\sigma^2}\left[\tilde{\mathbf{y}}^T\tilde{\mathbf{y}} + \hat{\boldsymbol{\beta}}_\gamma^T\mathbf{A}_\gamma\hat{\boldsymbol{\beta}}_\gamma - \hat{\boldsymbol{\beta}}_\gamma^{*T}\mathbf{A}^*_\gamma\hat{\boldsymbol{\beta}}^*_\gamma\right]\bigg). \tag{2.3.26}$$

Marginalising over σ^2

$$\int_{\sigma^2} p(\tilde{\mathbf{y}}|\boldsymbol{\gamma},\sigma^2)p(\sigma^2\boldsymbol{\gamma},\mathbf{y})d\sigma^2 = \int_{\sigma^2} p(\tilde{\mathbf{y}}|\boldsymbol{\gamma},\sigma^2)\frac{(b^*/2)^{a^*/2}}{\Gamma(a^*/2)}(\sigma^2)^{-\frac{m+a+n}{2}}\exp\bigg(-\frac{b^*}{2\sigma^2}\bigg)d\sigma^2$$

gives the marginal likelihood for the future observations as

$$p(\tilde{\mathbf{y}}|\boldsymbol{\gamma}) = \frac{(b^*/2)^{a^*/2}}{\Gamma(a^*/2)}\frac{\Gamma(a^*/2+m/2)}{(\sigma^2)^{m/2}|\mathbf{A}_\gamma|^{1/2}|A^*_\gamma|^{1/2}}\Big(\frac{1}{2}\left[\tilde{\mathbf{y}}^T\tilde{\mathbf{y}} + \hat{\boldsymbol{\beta}}_\gamma^T\mathbf{A}_\gamma\hat{\boldsymbol{\beta}}_\gamma - \hat{\boldsymbol{\beta}}_\gamma^{*T}\mathbf{A}^*_\gamma\hat{\boldsymbol{\beta}}^*_\gamma + b^*\right]\Big)^{-\frac{m+a+n}{2}}, \tag{2.3.27}$$

which simplifies to

$$p(\tilde{\mathbf{y}}|\boldsymbol{\gamma}) = \frac{(b^*/2)^{a^*/2}}{\Gamma(a^*/2)}\frac{\Gamma(a^*/2+m/2)}{(\sigma^2)^{m/2}|\mathbf{A}_\gamma|^{1/2}|A^*_\gamma|^{1/2}}\Big(\frac{1}{2}\left[\tilde{\mathbf{y}}^T\tilde{\mathbf{y}} + \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}_\gamma^{*T}\mathbf{A}^*_\gamma\hat{\boldsymbol{\beta}}^*_\gamma + ab\right]\Big)^{-\frac{m+a+n}{2}}, \tag{2.3.28}$$

where

$$\hat{\boldsymbol{\beta}}_\gamma^{*T}\mathbf{A}^*_\gamma\hat{\boldsymbol{\beta}}^*_\gamma = (\tilde{\boldsymbol{X}}^T\tilde{y} + \boldsymbol{X}_\gamma^T\mathbf{y})^T(\tilde{\boldsymbol{X}}_\gamma^T\tilde{\boldsymbol{X}}_\gamma^T + \boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma + 1/\tau)^{-1}(\tilde{\boldsymbol{X}}^T\tilde{\mathbf{y}} + \boldsymbol{X}_\gamma^T\mathbf{y}). \tag{2.3.29}$$

2.3.6 Shrinkage priors - Bayesian lasso prior

The Laplace prior on β_j is equivalent to a scale mixture of Gaussian's with an exponential mixing density,

$$\beta_j|\sigma^2,\tau_j \sim N(0,\sigma^2\tau_j) \quad \tau_j \sim Ex(\lambda^2/2). \tag{2.3.30}$$

Integrating over the scale by

$$P(\beta_j|\sigma^2) = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\tau_j}} \exp\left(-\frac{1}{2\sigma^2\tau_j}\beta_j^2\right) \quad \frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2\tau_j}{2}\right)d\tau_j \tag{2.3.31}$$
$$= \frac{\lambda^2}{2\sqrt{2\pi\sigma^2}} \int_0^\infty \frac{1}{\sqrt{\tau_j}} \exp\left(-\frac{1}{2}\left(\frac{|\beta_j|^2}{\sigma^2\tau_j} + \lambda^2\tau_j\right)\right)d\tau_j$$

Express $|\beta_j|^2/(\sigma^2\tau_j) + \lambda^2\tau_j = (|\beta_j|/(\sigma\sqrt{\tau_j}) - \lambda\sqrt{\tau_j})^2 + 2|\beta_j|\lambda/\sigma$ gives

$$P(\beta_j|\sigma^2) = \frac{\lambda^2}{2\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|\beta_j|\lambda}{\sigma}\right) \int_0^\infty \frac{1}{\sqrt{\tau_j}} \exp\left(-\frac{1}{2}\left(\frac{|\beta_j|}{\sigma\sqrt{\tau_j}} - \lambda\sqrt{\tau_j}\right)^2\right)d\tau_j \tag{2.3.32}$$

Use change of variable technique, set $\sqrt{\tau}_j = \nu$, $d\tau_j = 2\nu d\nu$

$$P(\beta_j|\sigma^2) = \frac{\lambda^2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|\beta_j|\lambda}{\sigma}\right) \int_0^\infty \exp\left(-\frac{1}{2}\left(\frac{|\beta_j|}{\sigma\nu} - \lambda\nu\right)^2\right)d\nu \tag{2.3.33}$$

Change of variable technique for $\epsilon = |\beta_j|/(\sigma\nu) - \lambda\nu$ which means one solution is $\nu = (-\epsilon + \sqrt{\epsilon^2 + 4\lambda|\beta_j|/\sigma})/2\lambda$. If we think of the integrand in (2.3.33) as a function of ν where the pdf can be expressed as $f_\nu(\nu) = f_\nu(\epsilon)|d\nu/d\epsilon|$, this reminds us that we require a positive Jacobian. In our case we have a Jacobian of

$$\frac{d\nu}{d\epsilon} = \left(-1 + \frac{\epsilon}{\sqrt{\epsilon^2 + 4\lambda|\beta_j|/\sigma}}\right)/(2\lambda) \tag{2.3.34}$$

As we need the absolute value before, and the expression in 2.3.34 is always negative we have

$$p(\beta_j|\sigma^2) = \frac{\lambda^2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|\beta_j|\lambda}{\sigma}\right) \int_{-\infty}^\infty \exp\left(-\frac{1}{2}\epsilon^2\right) \frac{1 - \epsilon(\epsilon^2 + 4\lambda|\beta_j|/\sigma)^{-\frac{1}{2}}}{2\lambda} d\epsilon$$
$$= \frac{\lambda^2}{2\lambda\sqrt{\sigma^2}} \exp\left(-\frac{|\beta_j|\lambda}{\sigma}\right) \int_{-\infty}^\infty \psi(\epsilon)\left(1 - \epsilon(\epsilon^2 + 4\lambda|\beta_j|/\sigma)^{-\frac{1}{2}}\right) d\epsilon$$
$$= \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{|\beta_j|\lambda}{\sigma}\right) \left(1 - \int_{-\infty}^\infty \psi(\epsilon)\epsilon(\epsilon^2 + 4\lambda|\beta_j|/\sigma)^{-\frac{1}{2}}\right) d\epsilon$$

As $\psi(\epsilon)\epsilon(\epsilon^2 + 4\lambda|\beta_j|/\sigma)^{-\frac{1}{2}}$ is an odd function this integrates to 0 hence leaving a Laplace prior.

2.3.7 Shrinkage priors - Bayesian lasso posterior

The fully Bayes prior parameterisation is

$$\begin{aligned}
\boldsymbol{\beta}|\sigma^2, \prod_{j=1}^{p} \tau_j^2 &\sim N_p(\mathbf{0}, \sigma^2 \boldsymbol{D}_\tau) \\
\tau_j^2|\lambda^2 &\sim Exponential\left(\frac{\lambda^2}{2}\right) \\
\lambda^2 &\sim Gamma(r, \delta) \\
\sigma^2 &\propto \frac{1}{\sigma^2}
\end{aligned}$$

where $\boldsymbol{D}_\tau = diag(\tau_1^2, ..., \tau_p^2)$. The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta}|\mathbf{y}, \tau_j^2, \sigma^2 &\sim N_p(\boldsymbol{V}\boldsymbol{X}^T\mathbf{y}, \sigma^2\boldsymbol{V}), \\
\frac{1}{\tau_j^2}|\mathbf{y}, \lambda, \sigma^2, \beta_j &\sim IG\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda^2\right) \\
\lambda^2|\mathbf{y}, \tau_j^2 &\sim Gamma\left(r + p, \frac{\sum_{j=1}^{p} \tau_j^2}{2} + \delta\right) \\
\sigma^2|\mathbf{y}, \boldsymbol{\beta}, \tau_j^2 &\sim IG\left(\frac{n+p}{2}, (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T\mathbf{D}_\tau^{-1}\boldsymbol{\beta})\right)
\end{aligned}$$

where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X} + \mathbf{D}_\tau^{-1})^{-1}$.

2.3.8 Shrinkage priors - Generalized double Pareto

The generalized double Pareto fully Bayes prior is specified as

$$\begin{aligned}
\boldsymbol{\beta}|\prod_{j=1}^{p} \tau_j, \sigma^2 &\sim N_p(\mathbf{0}, \sigma^2 \boldsymbol{D}_\tau) \\
\tau_j^2|\lambda_j^2 &\sim Exponential\left(\frac{\lambda_j^2}{2}\right) \\
\lambda_j^2 &\sim Gamma(r, \delta) \\
\sigma^2 &\propto \frac{1}{\sigma^2}
\end{aligned}$$

where $\boldsymbol{D}_\tau = diag(\tau_1^2, ..., \tau_p^2)$. The conditional posteriors are of the form

$$\begin{aligned}
\boldsymbol{\beta}|\mathbf{y}, \sigma^2, \tau_j^2 &\sim N_p(\boldsymbol{V}\boldsymbol{X}^T\mathbf{y}, \boldsymbol{V}\sigma^2), \\
\frac{1}{\tau_j^2}|\mathbf{y}, \lambda_j, \beta_j, \sigma^2 &\sim IG\left(\sqrt{\frac{\lambda_j^2\sigma^2}{\beta_j^2}}, \lambda_j^2\right) \\
\lambda_j^2|\mathbf{y}, \beta_j, \sigma^2 &\sim Gamma\left(r+1, \sqrt{\frac{\beta_j^2}{\sigma^2}} + \delta\right) \\
\frac{1}{\sigma^2}|\mathbf{y}, \boldsymbol{\beta}, \tau_j &\sim IG\left(\frac{n-1+p}{2}, \frac{(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})}{2} + \boldsymbol{\beta}^T\mathbf{D}_\tau^{-1}\boldsymbol{\beta})\right)
\end{aligned}$$

where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X} + \mathbf{D}_\tau^{-1})^{-1}$.

2.3.9 Shrinkage priors - Horseshoe prior

Derivation of the implied shrinkage parameter for the Horseshoe prior. This interpretation comes from the prior parameterisation in Carvalho et al. (2009),

$$\begin{aligned}
y_j &\sim N(\beta_j, \sigma^2), \\
\beta_j|\tau_j, \lambda &\sim N(0, \tau_j^2\lambda^2), \\
\tau_j &\sim C^+(0, 1).
\end{aligned}$$

This simpler form, which avoids a design matrix allows for an intuitive interpretation of τ_j in terms of κ_j.

$$\kappa_j = \frac{1}{1+\tau_j^2} \qquad \tau_j \sim C^+(0,1)$$
$$f(\tau_j) \propto \frac{1}{(1+\tau_j^2)}$$

Express τ_j in terms of κ_j

$$\tau_j = \left(\frac{1-\kappa_j}{\kappa_j}\right)^{1/2}$$

and compute the Jacobian

$$\left|\frac{\partial \tau_j}{\partial \kappa_j}\right| = \left|\frac{-1}{2\kappa_j^2}\left(\frac{1-\kappa_j}{\kappa_j}\right)^{-1/2}\right|.$$

The implied density for κ_j is thus

$$\begin{aligned} f(\kappa_j) &\propto \frac{1}{1+\left(\frac{1-\kappa_j}{\kappa_j}\right)}\left|\frac{-1}{2\kappa_j^2}\left(\frac{1-\kappa_j}{\kappa_j}\right)^{-1/2}\right| \\ &= \kappa_j^{-1/2}(1-\kappa_j)^{-1/2} \quad \square \end{aligned}$$

A more common full prior parameterisation is

$$\begin{aligned} \boldsymbol{\beta}|\prod_{j=1}^{p}\tau_j, \lambda, \sigma^2 &\sim N_p(\mathbf{0}, \sigma^2\lambda^2\boldsymbol{D}_\tau) \\ \tau_j|\lambda &\sim C^+(0,\lambda) \quad \text{for } j = 1, ..., p. \\ \lambda &\sim C^+(0,1) \\ \sigma^2 &\propto \frac{1}{\sigma^2} \end{aligned}$$

where $\boldsymbol{D}_\tau = diag(\tau_1^2, ..., \tau_p^2)$. The conditional posteriors are of the form

$$\begin{aligned} \boldsymbol{\beta}|\mathbf{y}, \sigma^2, \tau_j^2 &\sim N_p(\boldsymbol{V}\boldsymbol{X}^T\mathbf{y}, \boldsymbol{V}\sigma^2), \\ \sigma^2|\mathbf{y}, \boldsymbol{\beta}, \tau_j, \lambda &\sim IG\left(\frac{n+p}{2}, \frac{1}{2}((\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta})\right) \\ p(\tau_j|\mathbf{y}, \beta_j, \sigma^2, \lambda) &\propto \left(\frac{1}{\tau_j^2}\right)^{\frac{1}{2}}\exp\left(-\frac{\beta_j}{2\sigma^2\lambda^2\tau_j^2}\right)\frac{1}{1+\tau_j^2} \\ p(\lambda|\mathbf{y}, \boldsymbol{\beta}, \tau_j, \sigma^2) &\propto \left(\frac{1}{\lambda^2}\right)^{\frac{p}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\tau_j^2\lambda^2}\right)\frac{1}{1+\lambda^2} \end{aligned}$$

where $\boldsymbol{\Lambda} = diag(\lambda^2\tau_1^2, ..., \lambda^2\tau_p^2)$ and $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Lambda}^{-1})^{-1}$.

CHAPTER 3

Multivariate Variable Selection

3.1 Matrix Normal Spike-and-Slab

Brown, Vannucci and Fearn (Brown et al. (1998) and Brown et al. (2002)) extend the general framework of explicit variable selection for univariate regression to T multivariate outcomes

$$\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_T), \quad \mathbf{y}_t = (y_{1t}, ..., y_{nt})^T \quad \text{for} \;\; t = 1, .., T,$$

where the vector of latent indicator variables $\boldsymbol{\gamma}$ determines the set of covariates associated with all T outcomes. Conditionally on the matrix of parameters $\mathbf{B}_{p\times T}$, covariance within the columns $\boldsymbol{I}_n$ and within the rows $\boldsymbol{C}$, the standard multivariate normal regression model is assumed

$$\mathbf{Y} \sim \text{Matrix } \mathcal{N}_{n,T}(\boldsymbol{X}\mathbf{B}, \boldsymbol{I}_n, \boldsymbol{C}). \tag{3.1.1}$$

The covariance matrix $\boldsymbol{C}$ contains the variances and covariances of $y_{i1}, ..., y_{iT}$ in any $\mathbf{y}_i$,

$$\text{cov}(\mathbf{y}_i) = \boldsymbol{C} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1T} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{T1} & \sigma_{T2} & \dots & \sigma_T^2 \end{pmatrix} \tag{3.1.2}$$

for all $i = 1, 2, ..., n$ where $\mathbf{y}_i^T$ is a row in $\mathbf{Y}$ and is of dimension T. The matrix normal parameterisation in (3.1.1) sets the off-diagonals in $\text{cov}(\mathbf{y}_t)$ equal to 0. The assumption of $\text{cov}(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{O}$, for all $i \neq j$ is also made.

The conjugate prior for the matrix of regression parameters $\mathbf{B}_{p\times T}$ is

$$\mathbf{B}|\boldsymbol{\gamma}, \boldsymbol{C} \sim \text{Matrix } \mathcal{N}_{p,T}(\mathbf{B}_0, \boldsymbol{H}_\gamma, \boldsymbol{C}), \tag{3.1.3}$$

conditional on the parameters $\mathbf{B}_0$, $\boldsymbol{H}_\gamma$ and $\boldsymbol{C}$. By making the covariance across the columns dependent on $\boldsymbol{C}$, the univariate conjugate prior distribution is extended to T responses. By using the same vector of latent indicator variables $\boldsymbol{\gamma}$ for all T responses, only the covariance across the columns is embedded with $\boldsymbol{\gamma}$.

An inverse Wishart prior is placed on $\boldsymbol{C}$

$$\boldsymbol{C} \sim \mathcal{IW}(\delta; \boldsymbol{Q}), \tag{3.1.4}$$

where δ are the degrees of freedom and $\boldsymbol{Q}$ is a positive definite matrix. The scale matrix hyperparameter $\boldsymbol{Q}$ can be given the form $k\boldsymbol{I}_T$. Weak prior information requires a small value of δ, a value of 3 for δ gives $\mathbb{E}(\boldsymbol{C}) = \boldsymbol{Q}/(\delta - 2) = \boldsymbol{Q}$.

The parameterisation is completed with multivariate Bernoulli prior on $\boldsymbol{\gamma}$. A simple prior is $p(\gamma_j = 1) = \omega$, with a beta hyperprior. The presence of multiple responses does allow for the "sparsity" parameter in the prior to vary across the rows of $\mathbf{B}$, $p(\gamma_j = 1) = \omega_j \ j = 1, ..., p$.

The parameterisation of $\boldsymbol{H}_\gamma$ is analogous to the univariate spike-and-slab (2.1.4) discussed in

Chapter 2. One option is to extend the multivariate prior used by George and McCulloch (1993), by taking the row covariance matrix of $\mathbf{B}$ as

$$\boldsymbol{H}_\gamma = \boldsymbol{D}_\gamma \boldsymbol{R}_\gamma \boldsymbol{D}_\gamma \tag{3.1.5}$$

where $\boldsymbol{D}_\gamma$ is a diagonal matrix and $\boldsymbol{R}_\gamma$ a correlation matrix. The ith diagonal element of $\boldsymbol{D}_\gamma^2$ is denoted by v_{0j} when $\gamma_j = 0$ and v_{1j} when $\gamma_j = 1$. When the row components of $\mathbf{B}$ are assumed to be apriori independent $\boldsymbol{R}_\gamma \equiv \boldsymbol{I}$ and the prior matrix of coefficients $\mathbf{B}_0$ is the zero matrix, a selection prior can be motivated. Setting $v_{0j} \equiv 0$ means that the jth row of $\mathbf{B}$ has variance 0, where as $\gamma_j = 1$ indicates that the jth row has a non zero variance determined by v_{1j}. The prior distribution of $\mathbf{B}$ reduces to a singular p_γ-dimensional distribution

$$\mathbf{B}_{(\gamma)} \sim \text{Matrix } \mathcal{N}_{p_\gamma, T}(\mathbf{B}_{0\gamma}, \boldsymbol{H}_\gamma, \boldsymbol{C}) \tag{3.1.6}$$

where $\mathbf{B}_{(\gamma)}$ selects rows of $\mathbf{B}$ that have $\gamma_j = 1$. Alternatively the correlation structure of the least squares estimates can be used $\boldsymbol{R}_\gamma \propto (\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}$. This is akin to the g-prior (Zellner, 1986) and can achieve considerable computational savings for the MCMC sampler.

By choosing conjugate priors the marginal posterior probability of inclusion is explicitly available up to a constant of proportionality (Appendix 3.4.1), and can be sampled directly as it is only a function of the hyperparameters, design matrix and data

$$p(\boldsymbol{\gamma}|\mathbf{Y}) = |\boldsymbol{H}_\gamma|^{-\frac{T}{2}} |\mathbf{K}_\gamma|^{-\frac{T}{2}} |\boldsymbol{Q}*|^{-(\frac{\delta+n+T-1}{2})} p(\boldsymbol{\gamma}) \tag{3.1.7}$$

where

$$\mathbf{K}_\gamma = \boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma + \boldsymbol{H}_\gamma^{-1}. \tag{3.1.8}$$

and

$$\boldsymbol{Q}^* = \boldsymbol{Q} + \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T \boldsymbol{X}_\gamma \mathbf{K}_\gamma \boldsymbol{X}_\gamma^T \mathbf{Y} \tag{3.1.9}$$

As this is available up to a constant of proportionality, MCMC can be used to generate samples from the posterior distribution. This becomes a computational problem of searching over a 2^p binary space, which is well studied and reviewed in Section 2.1.3.

The form of the conditional posterior distribution $p(\mathbf{B}|\boldsymbol{C}, \boldsymbol{\gamma}, \mathbf{Y})$ is

$$\mathbf{B}_\gamma|\boldsymbol{C}, \boldsymbol{\gamma}, \mathbf{Y} \sim \text{Matrix } \mathcal{N}_{p_\gamma,T}(\mathbf{K}_\gamma^{-1}\mathbf{M}_\gamma, \mathbf{K}_\gamma^{-1}, \boldsymbol{C}), \tag{3.1.10}$$

where

$$\mathbf{M}_\gamma = \boldsymbol{X}_\gamma^T\mathbf{Y} + \boldsymbol{H}_\gamma^{-1}\mathbf{B}_{0\gamma} \tag{3.1.11}$$

and the marginal posterior distribution of $p(\boldsymbol{C}|\mathbf{Y})$ is

$$\boldsymbol{C} \sim \mathcal{IW}(\delta + n, \boldsymbol{Q} + \mathbf{A}_\gamma - \mathbf{M}_\gamma\mathbf{K}_\gamma^{-1}\mathbf{M}_\gamma). \tag{3.1.12}$$

The posterior predictive distribution of m future vectors of observations ($\mathbf{Y}^f_{m\times T}$) for a future design matrix $\boldsymbol{X}^f$ can be determined, by using the law of iterate expectations, to integrate the likelihood for $\mathbf{Y}^f$ with respect to the posterior distribution for $\mathbf{B}$ in (3.1.10). This gives

$$\mathbf{Y}^f|\boldsymbol{C}, \boldsymbol{\gamma} \sim \text{Matrix } \mathcal{N}_{p,T}(\boldsymbol{X}_\gamma^f\mathbf{K}_\gamma^{-1}\mathbf{M}_\gamma, \boldsymbol{I}_m + \boldsymbol{X}_\gamma^f\boldsymbol{K}_\gamma^{-1}\boldsymbol{X}_\gamma^{fT}, \boldsymbol{C}). \tag{3.1.13}$$

Integrating over $\boldsymbol{C}$ gives the posterior predictive distribution conditional on $\boldsymbol{\gamma}$, as defined by Dawid (1981)

$$\mathbf{Y}^f - \boldsymbol{X}_\gamma\mathbf{K}_\gamma^{-1}\mathbf{M}_\gamma \sim \mathcal{T}(\delta + n :, \boldsymbol{Q} + \mathbf{A}_\gamma - \mathbf{M}_\gamma\mathbf{K}_\gamma^{-1}\mathbf{M}_\gamma, \boldsymbol{Q}^*), \tag{3.1.14}$$

where $\boldsymbol{Q}^*$ is defined in (3.1.9).

To predict $\mathbf{Y}^f$ under quadratic loss, the unconditional expectation of the posterior predictive is averaged over the posterior distribution $p(\boldsymbol{\gamma}|\mathbf{Y})$ in (3.1.7)

$$\hat{\mathbf{Y}}^f = \sum_\gamma \boldsymbol{X}_\gamma^f(\mathbf{K}_\gamma^{-1}\mathbf{M}_\gamma)p(\boldsymbol{\gamma}|\mathbf{Y}). \tag{3.1.15}$$

3.2 Covariance Selection

In our context of the matrix normal regression (3.1.1), the focus of the feature selection thus far has fallen exclusively on the covariates, as the matrix $\boldsymbol{C}$ is estimated fully. However, there is considerable interest in determining the underlying relationship between the variables. In omics data, these relationships are often sparse relative to the number of variables.

3.2.1 Gaussian Graphical modelling

One approach for performing explicit covariance selection is Gaussian graphical models. These use a graph structure for modelling and making statistical inferences regarding complex relationships among variables. Two types of graphs are used in structure learning, undirected graphs which represent conditional dependence relationships among variables, and bi-directed graphs, which encode marginal dependence among variables. Under the Gaussian assumption, bi-directed graphs are determined by zeros in the covariance matrix (Cox and Wermuth (1993) and Silva and Ghahramani (2009)). Undirected graphs, which are explored in more detail, are determined by zeros in the precision matrix. If we define the multivariate normal density as

$$p(\mathbf{y}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi|\Sigma|)^{n/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right), \tag{3.2.1}$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_1 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \mathbf{V}_1 & \mathbf{R} \\ \mathbf{R}^T & \mathbf{V}_2 \end{pmatrix}, \tag{3.2.2}$$

the Schur complement allows us to define the precision matrix as

$$\Omega = \Sigma^{-1} = \begin{pmatrix} \mathbf{K}_1 & \mathbf{H} \\ \mathbf{H}^T & \mathbf{K}_2 \end{pmatrix}, \tag{3.2.3}$$

with

$$\mathbf{K}_1^{-1} = \mathbf{V}_1 - \mathbf{R}\mathbf{V}_2^{-1}\mathbf{R}^T \qquad \mathbf{H} = -\mathbf{K}_1\mathbf{R}\mathbf{V}_2^{-1}. \tag{3.2.4}$$

Using the property that conditional normal densities are also normal and completing the square, the conditional probability of $\mathbf{y}_1$ given $\mathbf{y}_2$ is

$$\mathbf{y}_1|\mathbf{y}_2 \sim \mathcal{N}_s(\boldsymbol{\mu}_1 + (-\mathbf{K}_1^{-1}\mathbf{H})(\mathbf{y}_2 - \boldsymbol{\mu}_2), \mathbf{K}_1^{-1}), \tag{3.2.5}$$

where the mean is also a function of the partitioned precision matrix. If y_1 is reduced to a scalar, and $\mathbf{y}_2 = \mathbf{y}_{2:m} = \mathbf{y}_{(-1)}$, the variance is $\omega_{1,1}^{-1}$, and the mean becomes

$$\mathbb{E}[y_1|\mathbf{y}_{(-1)}] = \mu_1 - \sum_{j=2}^{m} \frac{\omega_{1,j}}{\omega_{1,1}}(y_j - \mu_{2,j}), \tag{3.2.6}$$

as $\mathbf{H} = (\omega_{1,2}, ..., \omega_{1,m})$, which generalises to any scalar partition, with y_1 being the ith element and $\mathbf{y_2} = \mathbf{y}_{(-i)}$. This reveals explicitly, how the elements of the precision matrix characterise the conditional distribution of $y_i|\mathbf{y}_{(-i)}$. Zeros in the off-diagonal elements of the precision matrix define, and are defined by, the conditional independencies. $\omega_{ij} = 0$ if the complete conditional distribution does not depend on y_j given all the remaining elements $\mathbf{y}_{-(i,j)}$.

This property induces a unique undirected graph corresponding to each multivariate Gaussian distribution. Thus, m random variables represent m nodes, and if G is the adjacency graph pairing to the precision matrix, then the presence of an edge between two nodes implies conditional dependence and the absence of an edge implies conditional independence. A precision matrix of

$$(\mathbb{V}(\mathbf{y}))^{-1} = \begin{bmatrix} * & * & 0 & 0 & 0 \\ * & * & 0 & * & * \\ 0 & 0 & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \end{bmatrix}, \tag{3.2.7}$$

where $*$ is a non-zero element, directly translates into the graph in Figure 3.2.1.

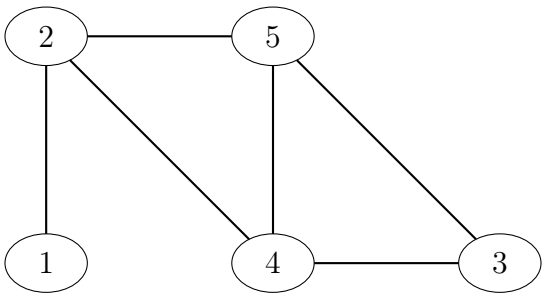

Figure 3.2.1: Undirected decomposable graph for the precision matrix defined in (3.2.7).

The non-zero entries in the off-diagonal correspond to the edges in Figure 3.2.1. For a n dimensional vector there are in total $2^{n(n-1)/2}$ possible conditional independence graphs. Even with a moderate number of variables, the discrete model space is astronomical in size.

As in the case of covariate selection from latent indicator variables, the model can be augmented with the graphical structure and an associated prior. Let $G = (V, E)$ be an undirected graph, where V is a set of vertices and $E = (i, j)$ is a set of edges for (i, j). A graph (or subgraph) is termed complete if all vertices are connected. Given this set of complete graphs, a *clique* is defined as a complete subgraph which is not completely a part of another subgraph (Carvalho et al., 2007). The following properties are thus equivalent

$$g_{ij} = 0 \Leftrightarrow (i,j) \notin E \quad \Leftrightarrow \quad y_i \perp\!\!\!\perp y_j | \mathbf{y}_{-(ij)} \Leftrightarrow \omega_{ij} = 0. \tag{3.2.8}$$

A popular approach is to restrict the space of G to decomposable graphs, as this allows for a convenient factorisation of the prior distribution. A decomposable graph is one which can be split into a set of cliques $P_1, ..., P_Q$ (Lauritzen, 1996). A clique (or prime component) is thus a complete maximal subset of a graph. Define $H_{q-1} = P_1 \cup ... \cup P_{q-1}$ and $S_q = H_{q-1} \cap P_q$. The S_q's are called separators, separating a completely-connected subgraph of G into two components, such that any path between the two components must pass through the separator. The two components and the separator form a decomposition of G. The cliques C_q can be ordered in such way that for every

$q > 1$ there exists $r < q$ such that

$$P_q \cap H_{q-1} \subset P_r \tag{3.2.9}$$

for $q = 2, ..., Q$. This is called the running intersection property Lauritzen (1996). In a (non-unique) perfect ordering $P_1; S_2, P_2; S_3, P_3; ...$ of cliques and separators, we call the clique sequence G^i and the separator sequence S^i.

The density for a mean zero random sample, $\mathbf{y}_i = (y_{1i}, ..., y_{ni})$, on the graph G is a function of multivariate Gaussian densities on the cliques and separators, with covariance matrices Σ_{PP} and Σ_{SS} on cliques and separators:

$$p(\mathbf{y}|\Sigma_G) = \frac{\prod_{P \in G^i} p(\mathbf{y}_P|\Sigma_{PP})}{\prod_{S \in S^i} p(\mathbf{y}_S|\Sigma_{SS})}. \tag{3.2.10}$$

Like the likelihood in (3.2.10), this density factors over the cliques and separators

$$p(\Sigma|G) = \frac{\prod_{P \in G^i} p(\Sigma_{PP}|G)}{\prod_{S \in S^i} p(\Sigma_{SS}|G)}. \tag{3.2.11}$$

For each clique of G (and each separator), the corresponding submatrix of the covariance Σ_{PP} has an inverse Wishart (δ, Φ_{PP}) prior.

Dawid and Lauritzen (1993) derived a conjugate prior distribution for Σ_G, termed the hyper-inverse Wishart $\text{HIW}(G, \delta, \Phi)$ with Φ a positive definite matrix and $\delta > 0$. If the k dimensional i.i.d random variables $\mathbf{y}_i \sim N(\mathbf{0}, \Sigma_G)$ for $i = 1, ..., n$ and $\Sigma_G \sim \text{HIW}_G(\delta, \Phi)$ is the prior, with Φ a positive definite $T \times T$ matrix, then the posterior is $\Sigma_G|\mathbf{Y} \sim \text{HIW}_G(\delta + n, \Phi + \mathbf{Y}^T\mathbf{Y})$, where $\mathbf{Y}$ is an $n \times T$ matrix.

Bhadra and Mallick (2013) incorporate this prior into a matrix normal Bayesian regression model (3.1.1), with a vector of latent indicator variables as described in Section 3.1. The complete

hierarchical model, for the $\mathbf{B}$ matrix is

$$\mathbf{Y} - \boldsymbol{X}_\gamma \mathbf{B}_{\gamma,G} | \boldsymbol{\gamma}, \boldsymbol{C}_G \sim \text{Matrix } \mathcal{N}_{n\times T}(\mathbf{0}, \boldsymbol{I}_n, \boldsymbol{C}_G) \tag{3.2.12}$$

$$\mathbf{B}_{\gamma,G} | \boldsymbol{\gamma}, \boldsymbol{C}_G \sim \text{Matrix } \mathcal{N}_{p_\gamma \times T}(\mathbf{0}, c\boldsymbol{I}_{p_\gamma}, \boldsymbol{C}_G) \tag{3.2.13}$$

$$\boldsymbol{C}_G | G \sim \text{HIW}_G(b, d\boldsymbol{I}_T) \tag{3.2.14}$$

$$\gamma_i \sim \text{Bernoulli}(w_\gamma) \quad \text{for} \quad i = 1, ..., p, \tag{3.2.15}$$

$$G_q \sim \text{Bernoulli}(w_G) \quad \text{for} \quad q = 1, ..., T(T-1)/2, \tag{3.2.16}$$

$$w_\gamma, w_G \sim \text{Uniform}(0, 1), \tag{3.2.17}$$

where b, c, d are fixed positive hyper-parameters and w_γ and w_G are prior weights that control the sparsity in $\boldsymbol{\gamma}$ and G respectively. The indexes i and q, denote the ith element for the vector $\boldsymbol{\gamma}$ and the qth off-diagonal edge in the lower triangular part of the adjacency matrix of the graph G. In order to preserve the positive definiteness of $\boldsymbol{C}_G$, the diagonal elements are always restricted to be 1.

The $\mathbf{B}$ parameters can be integrated out of the likelihood using iterative expectations to get

$$\mathbf{Y} | \boldsymbol{\gamma}, \boldsymbol{C}_G \sim \text{Matrix } \mathcal{N}_{n_\gamma \times T}(\mathbf{0}, \boldsymbol{I}_n + c(\boldsymbol{X}_\gamma \boldsymbol{X}_\gamma^T), \boldsymbol{C}_G). \tag{3.2.18}$$

Defining the Cholesky decomposition of the matrix $\{\boldsymbol{I}_n + c(\boldsymbol{X}_\gamma \boldsymbol{X}_\gamma^T)\}^{-1}$, when c is positive as

$$\mathbf{A}\mathbf{A}^T = \{\boldsymbol{I}_n + c(\boldsymbol{X}_\gamma \boldsymbol{X}_\gamma^T)\}^{-1}.$$

Defining $\mathbf{T} = \mathbf{A}\mathbf{Y}$,

$$\mathbf{T} | \boldsymbol{\gamma}, \boldsymbol{C}_G \sim \text{Matrix } \mathcal{N}_{n_\gamma \times T}(\mathbf{0}, \boldsymbol{I}_n, \boldsymbol{C}_G) \tag{3.2.19}$$

The choice of prior in (3.2.14) allows $\boldsymbol{C}$ to be integrated out of the likelihood, giving rise to the hyper-matrix t distribution of Dawid and Lauritzen (1993), to get

$$\mathbf{T} | \boldsymbol{\gamma}, G \sim \text{HMT}_{n\times T}(b, \boldsymbol{I}_n, d\boldsymbol{I}_T). \tag{3.2.20}$$

This is a special type of t distribution (Appendix 3.4.3) which, given the graph, splits into products and ratios over the cliques and separators as in (3.2.10). The joint search over the predictor and precision matrix elements can cycle between $\boldsymbol{\gamma}$ and G, in an MCMC sampler. The conjugate structure of the conditional posterior of $\mathbf{B}$ and Σ, allows the parameters to be sampled conditional on $\boldsymbol{\gamma}$ and G in a collapsed Gibbs sampler.

An alternative approach is to use Zellner's g-prior Zellner (1986) for multivariate regression,

$$\mathbf{B}_{\gamma,G}|\boldsymbol{\gamma}, \boldsymbol{C}_G \sim \text{Matrix } \mathcal{N}_{p_\gamma \times T}(\mathbf{0}, c(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}, \boldsymbol{C}_G) \tag{3.2.21}$$

which decreases the complexity of the marginalization of $\mathbf{y}$ over $\boldsymbol{C}_G$ (Niu et al., 2020).

These approaches rely on integrating out both $\mathbf{B}$ and $\boldsymbol{C}$, which is only possible if we restrict $\boldsymbol{\gamma}$ to be the same for each response. In a more general case, when $\boldsymbol{\gamma}$ is free to vary over the responses whilst feature selection is performed on the precision matrix, the parameters lose conjugacy and can not be integrated out. To resolve this issue, Banterle and Lewin (2018) reparameterise the matrix normal likelihood (3.1.1) by factorising the covariance matrix $\boldsymbol{C}$ iteratively, so that the likelihood is a product of independent regressions with a vector of latent indicator variables $\boldsymbol{\gamma}_t$ which varies across the responses

$$p(\mathbf{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{C}, \boldsymbol{\gamma}) = \prod_{t=1}^{T} p(\mathbf{y}_t|\boldsymbol{X}_{\gamma_t}\boldsymbol{\beta}_{\gamma_t} + \boldsymbol{U}_{(t-1)}\boldsymbol{\rho}_t, \sigma^2\boldsymbol{I}_n). \tag{3.2.22}$$

The matrix $\boldsymbol{U}_{(t-1)}$ consists of the residuals from the first $t-1$ regressions and the additional parameters are defined by

$$\left.\begin{aligned} \sigma_1^2 &\equiv c_1 \\ \sigma_t^2 &\equiv c_t - \mathbf{c}_t^T \boldsymbol{C}_{(t-1)}^{-1} \mathbf{c}_t \\ \boldsymbol{\rho}_t &\equiv \boldsymbol{C}_{(t-1)}^{-1} \mathbf{c}_t. \end{aligned}\right\} \quad t = 2, ..., T. \tag{3.2.23}$$

This corresponds to the iterative factorisation of the covariance matrix $\boldsymbol{C}$ for all $t = 2, ..., T$ with

$\boldsymbol{C}_{(T)} = \boldsymbol{C}$, $\boldsymbol{C}_{(1)} = c_1$ and $\mathbf{c}_1$ as null

$$\boldsymbol{C}_{(t)} = \begin{pmatrix} \boldsymbol{C}_{(t-1)} & \mathbf{c}_t \\ \mathbf{c}_t^T & c_t \end{pmatrix}. \tag{3.2.24}$$

A hyper-inverse Wishart prior on $\boldsymbol{C}_G$, which adds a graph structure on the precision matrix to model, means that the priors on the changed variables σ_t^2 and ρ_t are inverse gamma and normal, respectively. By using a perfect elimination ordering for the sequence of cliques and separators, an absence of an edge between the nodes (k, l) in G is equivalent to $\rho_{kl} = 0$. The addition of the graphical structure, translates directly into feature selection of the $\boldsymbol{\rho}$ parameters.

3.2.2 Explicit covariance selection

In the multivariate normal density (3.2.1), parsimony in the covariance matrix can also be identified through a Cholesky factorization of the precision matrix

$$\Omega = \Sigma^{-1} = \mathbf{A}\mathbf{D}\mathbf{A}^T, \tag{3.2.25}$$

where $\mathbf{A}$ is a lower triangular matrix with a spike-and-slab prior on the non-diagonal individual elements $a_{h,j}$ $(h > j)$ with ones along the diagonal and $\mathbf{D}$ is a diagonal matrix (Smith and Kohn, 2002). The binary indicator variable $\gamma_{h,j}$ induces the relationship

$$a_{h,j} \neq 0 \;\text{ iff }\; \gamma_{h,j} = 1, \qquad a_{h,j} = 0 \;\text{ iff }\; \gamma_{h,j} = 0 \tag{3.2.26}$$

for the elements $j = 1, .., m-1$, $h > j$, and is denoted by the γ index $\mathbf{A}_\gamma$.

Independent data can be obtained via

$$\mathbf{A}^T\mathbf{y}_i \sim N(\mathbf{0}, \mathbf{D}^{-1}), \tag{3.2.27}$$

which allows us to parameterise the likelihood of a zero mean regression as

$$p(\mathbf{Y}|\mathbf{A},\mathbf{D},\boldsymbol{\gamma}) = (2\pi)^{-\frac{nT}{2}}|\mathbf{D}|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\mathbf{y}_i^T\mathbf{A}_\gamma\mathbf{D}\mathbf{A}_\gamma^T\mathbf{y}_i\right). \tag{3.2.28}$$

Using the property $\mathbf{a}^T\mathbf{b} = \text{tr}(\mathbf{a}\mathbf{b}^T)$ and $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{CAB}) = \text{tr}(\mathbf{BCA})$, (3.2.28) can be expressed as

$$p(\mathbf{Y}|\mathbf{A},\mathbf{D},\boldsymbol{\gamma}) = (2\pi)^{-\frac{nT}{2}}\prod_{i=1}^{T}d_i^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\sum_{k=1}^{T}d_k\mathbf{a}_{\gamma,k}^T\mathbf{S}\mathbf{a}_{\gamma,k}\right), \tag{3.2.29}$$

where $\mathbf{a}_{\gamma,k}$ is the kth column of $\mathbf{A}_{\gamma,k}$ embedded with elements of $\boldsymbol{\gamma}$, $\mathbf{S} = \sum_i \mathbf{y}_i\mathbf{y}_i^T$ and d_i are the diagonal elements of $\mathbf{D}$. The matrix $\mathbf{S}$ is positive-definite almost surely if $T \leq n$. The dot product in the exponent can be expressed as

$$\mathbf{a}_{\gamma,k}^T\mathbf{S}\mathbf{a}_{\gamma,k} = \begin{cases} s_{k,k} + 2\boldsymbol{\rho}_{k,\gamma}^T\mathbf{s}_{k,\gamma} + \boldsymbol{\rho}_{k,\gamma}^T\mathbf{S}_{k,\gamma}\boldsymbol{\rho}_{k,\gamma} & \text{for } k = 1,...,T-1 \\ s_{m,m} & \text{for } k = T. \end{cases} \tag{3.2.30}$$

The dependency on $\boldsymbol{\gamma}$ is expressed though the vectors $\boldsymbol{\rho}_{k,\gamma} = (\rho_{h,k}|h > k, \gamma_{h,k} = 1)$, $\mathbf{s}_{k,\gamma} = (s_{h,k}|h > k, \gamma_{h,k} = 1)$, and the matrix $\mathbf{S}_{k,\gamma} = (s_{h,j}|h > k, j > k, \gamma_{h,j} = 1)$. The total number of unconstrained elements in $\mathbf{A}$ corresponding to model $\boldsymbol{\gamma}$ is $q_\gamma = \sum_{k=1}^{m-1} q_k$. Finally, after completing the square, the likelihood can be expressed as

$$p(\mathbf{Y}|\mathbf{A},\mathbf{D},\boldsymbol{\gamma}) = (2\pi)^{-\frac{Tn}{2}}\prod_{k=1}^{T}(d_k)^{\frac{n}{2}}\exp\left(-\frac{d_k}{2}\Big(\boldsymbol{R}_k(\boldsymbol{\gamma}) + (\boldsymbol{\rho}_{\gamma,k} - \boldsymbol{m}_{\gamma,k})^T\mathbf{S}_{\gamma,k}(\boldsymbol{\rho}_{\gamma,k} - \boldsymbol{m}_{\gamma,k})\Big)\right) \tag{3.2.31}$$

where $\boldsymbol{m}_{\gamma,k} = -\mathbf{S}_{\gamma,k}^{-1}\mathbf{s}_{\gamma,k}$ and $\boldsymbol{R}_k(\boldsymbol{\gamma}) = s_{k,k} - \mathbf{s}_{\gamma,k}^T\mathbf{S}_{\gamma,k}^{-1}\mathbf{s}_{\gamma,k}$.

This is similar to the reparameterisation used by Banterle and Lewin (2018), but now the residuals from the different responses are effectively informing the prior on the changed parameters, in an empirical Bayes approach. Smith and Kohn (2002) use a fractional conditional prior for $\boldsymbol{\rho}_\gamma$ by setting

$$p(\boldsymbol{\rho}|\boldsymbol{D},\gamma) \propto p(\mathbf{Y}|\mathbf{A},\mathbf{D},\boldsymbol{\gamma})^{\frac{1}{n}} \tag{3.2.32}$$

which mean the changed parameters are normally distributed

$$\boldsymbol{\rho}_k|\boldsymbol{D},\gamma \sim N\left(\boldsymbol{m}_{\gamma,k}, \frac{n}{d_k}\boldsymbol{S}^{-1}_{\gamma,k}\right). \tag{3.2.33}$$

The conditional posterior updates are all available in closed form. Samples can be obtained via a collapsed Gibbs sampler, where the marginal posterior is used for indicator variable to avoid the sampling issues outlined in Section 2.1. Clearly as the dimension of T increases, the computational burden required to search the whole binary space can become overwhelming, as the problem is $\mathcal{O}(T^2)$.

Wang (2015) combines graphical modelling with latent indicator variables, to try and address the computational challenges of covariate selection. The approach involves representing the graphical structure of the precision matrix by a set of latent variables $\boldsymbol{Z} = (z_{ij})_{i<j}$, where $z_{ij} = 1$ or 0 according to whether edge (i,j) belongs to E or not. The marginal prior on the precision matrix is defined as

$$p(\Omega) = C(\boldsymbol{\vartheta})^{-1}\prod_{i<j}\left\{(1-\pi)N(\omega_{ij}|0,v_0^2) + \pi N(\omega_{ij}|0,v_1^2)\right\}\prod_i\left\{\text{Exp}\left(\omega_{ii}|\frac{\lambda}{2}\right)\right\}1_{(\Omega\in M^+)} \tag{3.2.34}$$

where v^2 is the variance for a spike-and-slab normal mixture (in the same form as (2.1.2)), $\text{Exp}(\omega_{(\cdot)}|\lambda/2)$ is the exponential density and $1_{(\cdot)}$ is the indicator function. The term $C(\boldsymbol{\vartheta})$ is the normalising constant, which depends on the parameters $\boldsymbol{\vartheta} = (v_0, v_1, \omega, \lambda)$ and ensures the integration of the density over the positive-definite space M^+ is one.

The joint hierarchical prior, from which the marginal prior is derived from, can thus be defined as

$$p(\omega|\boldsymbol{Z},\boldsymbol{\vartheta}) \propto \prod_{i<j} N(\omega_{ij}|0,v^2_{z_{ij}})\prod_i \text{Exp}\Big(\omega_{ii}|\frac{\lambda}{2}\Big) \tag{3.2.35}$$

$$P(\boldsymbol{Z}|\boldsymbol{\vartheta}) \propto \prod_{i<j}\pi^{z_{ij}}(1-\pi)^{1-z_{ij}} \tag{3.2.36}$$

where the omitted constants are the respective integrals over the positive-definite space M^+.

By partitioning the matrices $\Omega, \mathbf{S} = \mathbf{Y}^T\mathbf{Y}$ and $\mathbf{V} = (v^2_{z_{ij}})$ into:

$$\Omega = \begin{pmatrix} \Omega_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{12}^T & \omega_{22} \end{pmatrix} \qquad \mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{s}_{12} \\ \mathbf{s}_{12}^T & s_{22} \end{pmatrix} \qquad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{v}_{12} \\ \mathbf{v}_{12}^T & 0 \end{pmatrix} \tag{3.2.37}$$

and imposing the change of variable

$$(\boldsymbol{\omega}_{21}, \omega_{22}) \rightarrow (\mathbf{u} = \boldsymbol{\omega}_{12}, a = \omega_{22} - \boldsymbol{\omega}_{12}^T\Omega_{11}^{-1}\boldsymbol{\omega}_{12}) \tag{3.2.38}$$

the full posterior conditionals for the new variables are:

$$p(\mathbf{u}|\cdot) \sim N(-\mathbf{C}\mathbf{s}_{12}, \mathbf{C}) \qquad p(a|\cdot) \sim \text{Ga}\Big(\frac{n}{2} + 1, \frac{s_{22} + \lambda}{2}\Big) \tag{3.2.39}$$

where $\mathbf{C} = ((s_{22} + \lambda)\Omega_{11}^{-1} + \text{diag}(\mathbf{v}_{12}^{-1}))^{-1}$.

Permuting any column to be updated to the last one and using (3.2.39), will lead to a simple block Gibbs step for generating $\Omega|\boldsymbol{Z}, \mathbf{Y}$. The conditional posterior for all z_{ij} are independent Bernoulli with probability

$$p(z_{ij} = 1|\Omega, \mathbf{Y}) = \frac{N(\omega_{ij}|0, v_1^2)\pi}{N(\omega_{ij}|0, v_1^2)\pi + N(\omega_{ij}|0, v_0^2)(1 - \pi)}. \tag{3.2.40}$$

which is very similar to the SSVS Gibbs sampler update in variable selection in (George and McCulloch, 1993), from the mixture of continuous normal priors (2.1.2) (Appendix 3.4.4). The actual sampler can be recovered by reparameterising $\mathbf{u}$ and setting $\lambda = 0$.

Here the spike-and-slab prior retains the dimension of the precision matrix, but shrinks the off-diagonal elements towards zero. Feature selection requires thresholding, once an estimator has been applied. There is no guarantee the resulting estimate will be positive definite. Since the priors place zero probability mass on any sparse matrix containing exact zeros, as opposed to the point-mass mixture priors, the posterior will be more dispersed around zero for the true non-zero off-diagonal elements. The primary advantage to this approach is its scalability over

larger T problems, as computationally faster block updates of edge-inclusion in $\boldsymbol{Z}$ are performed simultaneously, rather than one edge inclusion indicator z_{ij} at a time.

3.3 Hierarchical Priors

An alternative approach to explicitly modelling the covariance between the multiple responses $\boldsymbol{C}$, is a hierarchical model in which each response $\mathbf{y}_t$ is linked to the same design matrix through the linear model with regression coefficients $\boldsymbol{\beta}_t = (\beta_{t1}, ..., \beta_{tp})$

$$\mathbf{y}_t \sim N(\boldsymbol{X}_{\gamma_t}\boldsymbol{\beta}_{\gamma_t}, \sigma_t^2\boldsymbol{I}_n) \quad \text{for} \quad t = 1, .., T. \tag{3.3.1}$$

The latent vector variable γ_t determines the covariates associated with each response where T vectors allow a unique combination of 0 and 1's for every response. The conditional residuals for each regression equation are assumed to be independent of each other and information is borrowed across the responses whilst controlling for sparsity over the T responses by careful choice of the hierarchical prior specification. In (Bottolo et al., 2011) the sparsity parameter ω_{tj} in the prior for the latent binary indicator variable $p(\gamma_{tj}|\omega_{tj}) = \text{Bernoulli}(\omega_{tj})$ is decomposed into the marginal effects

$$\omega_{tj} = \omega_t \times \rho_j,$$

where ω_t controls the level of sparsity for each t though a suitable choice of hyperparameters (a_t, b_t), while the parameter ρ_j captures the "relative propensity" of predictor j to influence several responses at a time. The support for ω_t $(0 \leq \omega_t \leq 1)$ and ρ_j $(\rho_j \geq 0)$ is constrained so that $0 \leq \omega_{tj} \leq 1$. As in the matrix normal approach, the regression and variance parameter can be integrated out if conjugate priors are assumed so the marginal posterior for the model space is tractable up to a constant of proportionality.

Although information is shared across the responses, allowing correlation between the parameters,

this approach has been shown to be out performed by methods which explicitly incorporate the residual covariance $\boldsymbol{C}$ into the model (Banterle and Lewin, 2018).

3.4 Appendix

3.4.1 Matrix normal - Derivation of the marginal selection posterior

The matrix of outcomes $\mathbf{Y}$ is assumed to have a matrix normal probability density. Starting with the multivariate normal

$$\text{vec}(\mathbf{Y}) \sim N_{n,T}(\text{vec}(\boldsymbol{X}\mathbf{B}), \boldsymbol{C} \otimes \boldsymbol{I}_n), \tag{3.4.1}$$

this can be expressed as

$$\begin{aligned} p(\mathbf{Y}|\boldsymbol{X}\mathbf{B}, \boldsymbol{C}) =& (2\pi)^{-\frac{nT}{2}} |\boldsymbol{C} \otimes \boldsymbol{I}_n|^{-\frac{1}{2}} \exp\Bigg(-\frac{1}{2}(\text{vec}(\mathbf{Y}) - \text{vec}(\boldsymbol{X}\mathbf{B}))^T (\boldsymbol{C} \otimes \boldsymbol{I}_n)^{-1} (\text{vec}(\mathbf{Y}) - \\ & \text{vec}(\boldsymbol{X}\mathbf{B}))\Bigg) \\ =& (2\pi)^{-\frac{nT}{2}} |\boldsymbol{C}|^{-\frac{n}{2}} \exp\Bigg(-\frac{1}{2}\text{vec}(\mathbf{Y} - \boldsymbol{X}\mathbf{B})^T (\boldsymbol{C}^{-1} \otimes \boldsymbol{I}_n^{-1}) \text{vec}(\mathbf{Y} - \boldsymbol{X}\mathbf{B})\Bigg). \end{aligned}$$

Using $(\boldsymbol{B}^T \otimes \boldsymbol{A})\text{vec}(\boldsymbol{X}) = \text{vec}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B})$

$$p(\mathbf{Y}|\boldsymbol{X}\mathbf{B}, \boldsymbol{C}) = (2\pi)^{-\frac{nT}{2}} |\boldsymbol{C}|^{-\frac{n}{2}} |\boldsymbol{I}_n|^{-\frac{T}{2}} \exp\Bigg(-\frac{1}{2}\text{vec}(\mathbf{Y} - \boldsymbol{X}\mathbf{B})^T \text{vec}(\boldsymbol{I}_n^{-1}(\mathbf{Y} - \boldsymbol{X})\boldsymbol{C}^{-1})\Bigg),$$

$\text{vec}(\boldsymbol{A})^T \text{vec}(\boldsymbol{B}) = \mathbf{tr}(\boldsymbol{A}^T\boldsymbol{B})$ and $\mathbf{tr}(\boldsymbol{A}\boldsymbol{B}) = \mathbf{tr}(\boldsymbol{B}\boldsymbol{A})$ to obtain

$$p(\mathbf{Y}|\boldsymbol{X}\mathbf{B}, \boldsymbol{C}) = (2\pi)^{-\frac{nT}{2}} |\boldsymbol{C}|^{-\frac{n}{2}} \exp\Bigg(-\frac{1}{2}\text{tr}\Big((\mathbf{Y} - \boldsymbol{X}\mathbf{B})\boldsymbol{C}^{-1}(\mathbf{Y} - \boldsymbol{X}\mathbf{B})^T\Big)\Bigg),$$

which is the typical form of the probability density and is denoted

$$\mathbf{Y} \sim \text{Matrix } \mathcal{N}_{n,T}(\boldsymbol{X}\mathbf{B}, \boldsymbol{I}_n, \boldsymbol{C}). \tag{3.4.2}$$

The prior parameterisation of the model is

$$\begin{aligned}\mathbf{Y} &\sim \text{Matrix } \mathcal{N}_{n,T}(\boldsymbol{X}\mathbf{B}, \boldsymbol{I}_n, \boldsymbol{C})\\ \mathbf{B}|\boldsymbol{\gamma}, \boldsymbol{C} &\sim \text{Matrix } \mathcal{N}_{p,T}(\mathbf{B}_0, \boldsymbol{H}_\gamma, \boldsymbol{C})\\ \boldsymbol{C} &\sim \mathcal{IW}(\delta; \boldsymbol{Q}),\end{aligned}$$

where $\boldsymbol{\gamma}$ is the latent indicator variable.

Ar conjugate prior for variable selection is $\mathbf{B}|\boldsymbol{\gamma}, \boldsymbol{C} \sim \text{Matrix } \mathcal{N}_{p,T}(\mathbf{B}_0, \boldsymbol{H}_\gamma, \boldsymbol{C})$ with density

$$p(\mathbf{B}|\boldsymbol{C}, \boldsymbol{\gamma}) = (2\pi)^{-\frac{p_\gamma T}{2}}|H_\gamma|^{-T/2}|\boldsymbol{C}|^{-\frac{p_\gamma}{2}}\exp\left(-\frac{1}{2}\text{tr}\left(\boldsymbol{H}_\gamma^{-1}(\mathbf{B}_\gamma - \mathbf{B}_{0\gamma})\boldsymbol{C}^{-1}(\mathbf{B}_\gamma - \mathbf{B}_{0\gamma})^T\right)\right) \quad (3.4.3)$$

This has the effect of forcing $\boldsymbol{\gamma}$ into the likelihood via $\boldsymbol{X}_\gamma\mathbf{B}_\gamma$. The parameter $\mathbf{B}_\gamma$ can be integrated of the joint distribution given $\boldsymbol{C}$ and $\boldsymbol{\gamma}$. The exponent is

$$-\frac{1}{2}\text{tr}\left(\boldsymbol{C}^{-1}\left[(\mathbf{Y} - \boldsymbol{X}_\gamma\mathbf{B}_\gamma)^T(\mathbf{Y} - \boldsymbol{X}_\gamma\mathbf{B}_\gamma) + (\mathbf{B}_\gamma - \mathbf{B}_{0\gamma})^T\boldsymbol{H}_\gamma^{-1}(\mathbf{B}_\gamma - \mathbf{B}_{0\gamma})\right]\right), \quad (3.4.4)$$

focusing on the terms within the square parenthesis, after completing the square this can be expressed as

$$(\mathbf{B}_\gamma^T - \mathbf{K}_\gamma^{-1}\mathbf{M}_\gamma)^T\mathbf{K}_\gamma(\mathbf{B}_\gamma^T - \mathbf{K}_\gamma^{-1}\mathbf{M}_\gamma) - \mathbf{M}_\gamma\mathbf{K}_\gamma^{-1}\mathbf{M}_\gamma + \mathbf{A}_\gamma, \quad (3.4.5)$$

where

$$\mathbf{M}_\gamma = \boldsymbol{X}_\gamma^T\mathbf{Y} + \boldsymbol{H}_\gamma^{-1}\mathbf{B}_{0\gamma} \quad (3.4.6)$$

$$\mathbf{A}_\gamma = \mathbf{Y}^T\mathbf{Y} + \mathbf{B}_{0\gamma}^T\boldsymbol{H}_\gamma^{-1}\mathbf{B}_{0\gamma} \quad (3.4.7)$$

$$\mathbf{K}_\gamma = (\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma + \boldsymbol{H}_\gamma^{-1}). \quad (3.4.8)$$

The first term in (3.4.4) is the completed quadratic from of $\mathbf{B}$. Multiplying this by $-\frac{1}{2}\mathbf{tr}(\boldsymbol{C}^{-1})$, taking its exponential, and collecting the necessary powers $-p_\gamma/2$ and $T/2$ of the determinants

$\mathbf{K}_\gamma$ and $\boldsymbol{C}$ respectively, forms a normal probability density and is integrated out. This leaves

$$p(\mathbf{Y}|\boldsymbol{C},\boldsymbol{\gamma}) = |\boldsymbol{C}|^{-\frac{n}{2}}|\boldsymbol{H}_\gamma|^{-\frac{T}{2}}|\mathbf{K}_\gamma|^{-\frac{T}{2}}\exp\left(-\frac{1}{2}\text{tr}\left(\boldsymbol{C}^{-1}\left[\mathbf{A}_\gamma - \mathbf{M}_\gamma\mathbf{K}_\gamma^{-1}\mathbf{M}_\gamma\right]\right)\right). \tag{3.4.9}$$

The probability of the inverse Wishart prior for $\boldsymbol{C}$ is of the same form as (3.4.9). Marginalising over $\boldsymbol{C}$ for a given $\boldsymbol{\gamma}$ gives the likelihood conditional on a specific $\boldsymbol{\gamma}$ proportional to

$$p(\mathbf{Y}|\boldsymbol{\gamma}) = |\boldsymbol{H}_\gamma|^{-\frac{T}{2}}|\mathbf{K}_\gamma|^{-\frac{T}{2}}|\boldsymbol{Q} + \mathbf{A}_\gamma - \mathbf{M}_\gamma\mathbf{K}_\gamma^{-1}\mathbf{M}_\gamma|^{-(\frac{\delta+n+T-1}{2})}. \tag{3.4.10}$$

This can be combined with the prior to obtain the marginal posterior for the selection vector $\boldsymbol{\gamma}$.

3.4.2 Matrix normal - Intercept term

For completeness, an intercept can be included in the likelihood in (3.1.1). A conjugate prior is a multivariate normal

$$\boldsymbol{\alpha} \sim N_T(\boldsymbol{\alpha}_0, h\boldsymbol{C}) \tag{3.4.11}$$

This can integrated out and if the prior is weak the marginal posterior of inclusion is unaffected. The prior probability density is

$$p(\boldsymbol{\alpha}|\boldsymbol{C}) \propto h^{-T/2}|\boldsymbol{C}|^{-1/2}\exp\left(-\frac{1}{2h}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{C}^{-1}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)\right) \tag{3.4.12}$$

The exponent of the likelihood is

$$-\frac{1}{2}\text{tr}\left(\boldsymbol{C}^{-1}\left((\mathbf{Y}-\boldsymbol{X}\mathbf{B})^T(\mathbf{Y}-\boldsymbol{X}\mathbf{B}) - 2(\mathbf{Y}-\boldsymbol{X}\mathbf{B})^T(\mathbf{1}\boldsymbol{\alpha}^T) + (\mathbf{1}\boldsymbol{\alpha}^T)^T\mathbf{1}\boldsymbol{\alpha}^T\right)\right)$$

As $\mathbf{Y}$ and $\boldsymbol{X}$ are standardised, $\boldsymbol{X}^T\mathbf{1}$ equals zero so this simplifies to

$$-\frac{1}{2}\text{tr}\left(\boldsymbol{C}^{-1}\left((\mathbf{Y}-\boldsymbol{X}\mathbf{B})^T(\mathbf{Y}-\boldsymbol{X}\mathbf{B}) + n\boldsymbol{\alpha}\boldsymbol{\alpha}^T\right)\right). \tag{3.4.13}$$

The exponent of the prior $p(\boldsymbol{\alpha}|\boldsymbol{C})$ can be expressed

$$-\frac{1}{2}\text{tr}\left(\boldsymbol{C}^{-1}h^{-1}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)\right),$$

so

$$\begin{aligned}\log(p(\mathbf{Y}|\boldsymbol{\alpha},\mathbf{B},\boldsymbol{C})p(\boldsymbol{\alpha}|\boldsymbol{C})) = \; & c(n,T)-\left(\frac{n}{2}+\frac{1}{2}\right)\log|\boldsymbol{C}|-\frac{1}{2}\text{tr}\Bigg(\boldsymbol{C}^{-1}(\mathbf{Y}-\boldsymbol{X}\mathbf{B})^T(\mathbf{Y}-\boldsymbol{X}\mathbf{B})+\\ & \boldsymbol{C}^{-1}\Big(n\boldsymbol{\alpha}\boldsymbol{\alpha}^T+h^{-1}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)\Big)\Bigg). \qquad (3.4.14)\end{aligned}$$

Focusing on the $\boldsymbol{\alpha}$ terms in (3.4.14), completing the square gives

$$\begin{aligned}n\boldsymbol{\alpha}\boldsymbol{\alpha}^T+h^{-1}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0) =& (n+h^{-1})\Big(\boldsymbol{\alpha}\boldsymbol{\alpha}^T-2(h(n+h^{-1}))^{-1}\boldsymbol{\alpha}_0^T\boldsymbol{\alpha}+\\ & +(h(n+h^{-1}))^{-1}\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^T\\ =& (n+h^{-1})(\boldsymbol{\alpha}-\bar{\boldsymbol{\alpha}})(\boldsymbol{\alpha}-\bar{\boldsymbol{\alpha}})^T-h^{-2}(n+h^{-1})^{-1}\boldsymbol{\alpha_0}\boldsymbol{\alpha_0}^T+\\ & +h^{-1}\boldsymbol{\alpha_0}\boldsymbol{\alpha_0}^T, \qquad (3.4.15)\end{aligned}$$

where

$$\bar{\boldsymbol{\alpha}}=(h(n+h^{-1}))^{-1}\boldsymbol{\alpha}_0.$$

The exponential of the first term in (3.4.15) with $|\boldsymbol{C}|^{-1/2}$ in the prior, can be integrated out. The second and third term in (3.4.15) tend to 0 as h becomes large in the weak prior and can be ignored.

3.4.3 Matrix normal - Hyper-matrix t distribution

Given n observations and the graph G, we know the sequence of cliques $P_1,...,P_Q$ and seperators $S_2,...,S_Q$. For any $A\subset P_j$, the nodes in A are selected and $\mathbf{T}_A^n$ corresponds to the $n\times|A|$ matrix, where $|A|$ denotes the cardinality of the set A. The hyper-matrix t density on a given clique P_j,

with degrees of freedom b and scale matrices $\boldsymbol{I}_n$ and $d\boldsymbol{I}_q$, is defined as

$$f(\mathbf{t}^n_{P_j}) = \frac{\Gamma_{|P_j|}((b+n+|P_j|-1)/2)}{\pi^{\frac{n}{2}}\Gamma_{|P_j|}((b+|P_j|-1)/2)\det\left(d\boldsymbol{I}^{n/2}_{|P_j|}\right)} \times \left[\det\left(\boldsymbol{I}_n + (\mathbf{t}^n_{P_j})(d\boldsymbol{I}_{|P_j|})^{-1}(\mathbf{t}^n_{P_j})^T\right)\right]^{-\frac{(b+n+|P_j|-1)}{2}}. \tag{3.4.16}$$

3.4.4 Block Gibbs sampler updates for precision matrix

The matrices Ω, $\mathbf{S} = \mathbf{Y}^T\mathbf{Y}$ and $\mathbf{V} = (v^2_{z_{ij}})$ are partitioned into the blocks:

$$\Omega = \begin{pmatrix} \Omega_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}^T_{12} & \omega_{22} \end{pmatrix} \qquad \mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{s}_{12} \\ \mathbf{s}^T_{12} & s_{22} \end{pmatrix} \qquad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{v}_{12} \\ \mathbf{v}^T_{12} & 0 \end{pmatrix} \tag{3.4.17}$$

The joint distribution is proportional to

$$p(\mathbf{Y}, \boldsymbol{Z}, \Omega) \propto |\Omega|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{S}\Omega)\right) \prod_{i<j} N(\omega_{ij}|0, v^2_{z_{ij}})\pi^{z_{ij}}(1-\pi)^{1-z_{ij}} \prod_i \frac{\lambda}{2}\exp\left(-\frac{\lambda\omega_{ii}}{2}\right). \tag{3.4.18}$$

The Ω posterior update is performed on the last column, so it is proportional to $(\boldsymbol{\omega}_{12}, \omega_{22})$. The determinant of the block matrix Ω can be expressed as

$$|\Omega| = (\omega_{22} - \omega^T_{12}\Omega^{-1}_{11}\omega_{12})|\Omega_{11}| \tag{3.4.19}$$

(Powell, Philip, 2011). After expanding the matrix product $\mathbf{S}\Omega$, the full conditional is proportional to

$$p(\boldsymbol{\omega}_{12}, \omega_{22}|\mathbf{Y}, \cdot) \propto (\omega_{22} - \omega^T_{12}\Omega^{-1}_{11}\omega_{12})^{\frac{n}{2}} \exp\left(-\frac{1}{2}\left(s_{22}\omega_{22} + \boldsymbol{\omega}^T_{12}\mathbf{D}^{-1}\boldsymbol{\omega}_{12} + \mathbf{s}_1 2^T\omega_{12}\right)\right)\exp\left(-\frac{\lambda\omega_{22}}{2}\right),$$

where $\mathbf{D}^{-1}$ is the diagonal matrix of the inverse of the vector $\mathbf{v}_{12}$, $\mathbf{D}^{-1} = \mathrm{diag}(\mathbf{v}^{-1}_{12})$.

Using a change of variable of $\mathbf{u} = \boldsymbol{\omega}_{12}, a = \omega_{22} - \boldsymbol{\omega}_{12}^T \Omega_{11}^{-1} \boldsymbol{\omega}_{12}$, the joint posterior is

$$p(\mathbf{u}, a|\mathbf{Y}, \cdot) \propto a^{\frac{n}{2}} \exp\left(-\frac{1}{2}\left(\mathbf{u}^T \mathbf{D}^{-1} \mathbf{u} + 2\mathbf{s}_{12}^T \mathbf{u} + (\lambda + s_{22})(a + \mathbf{u}^T \Omega_{11}^{-1} \mathbf{u})\right)\right), \tag{3.4.20}$$

with a Jacobian equal to 1. Making (3.4.20) proportional to $\mathbf{u}$ and completing the square gives

$$\mathbf{u}|\mathbf{Y}, \cdot \sim N(-\mathbf{C}\mathbf{s}_{12}, \mathbf{C}), \tag{3.4.21}$$

where $\mathbf{C} = ((s_{22} + \lambda)\Omega_{11}^{-1} + \mathbf{D}^{-1})^{-1}$.

The conditional posterior for a is,

$$a|\mathbf{Y}, \cdot \sim \text{Ga}\left(\frac{n}{2} + 1, \frac{s_{22} + \lambda}{2}\right). \tag{3.4.22}$$

Unlike in the explicit variable selection, we do not require a joint update. As the prior for $\omega_{ij}|z_{ij} = 0$ is a normal distribution, rather than a Dirac spike at 0, the latent indicator variable does not enter the likelihood. The posterior for $z_{ij} = 1$ is thus proportional to

$$p(z_{ij} = 1|\Omega, \mathbf{Y}) \propto N(\omega_{ij}|0, v_1^2)\pi. \tag{3.4.23}$$

Normalising, gives the probability

$$p(z_{ij} = 1|\Omega, \mathbf{Y}) = \frac{N(\omega_{ij}|0, v_1^2)\pi}{N(\omega_{ij}|0, v_1^2)\pi + N(\omega_{ij}|0, v_0^2)(1 - \pi)}. \tag{3.4.24}$$

Finally, the SSVS Gibbs sampler of George and McCulloch (1993) can be recovered by setting $\lambda = 0$, reparameterising $\boldsymbol{\beta} \equiv (\beta_1, ..., \beta_{p-1}) = -\mathbf{u}$ and noting that $s_{22} = n$, for standardised data. If $\Omega_{11}^{-1} = \frac{1}{n}\mathbf{S}_{11}$ then

$$\boldsymbol{\beta}|\boldsymbol{z}_{12}, \mathbf{Y} \sim N\left((\mathbf{S}_{11} + \text{diag}(\mathbf{v}_{12}^{-1}))^{-1}\mathbf{s}_{12}, (\mathbf{S}_{11} + \text{diag}(\mathbf{v}_{12}^{-1}))\right). \tag{3.4.25}$$

As selection is performed on the last column of the precision matrix, defining the corresponding edge inclusion vector $\boldsymbol{\gamma} \equiv (\gamma_1, ..., \gamma_{p-1})^T = (z_{1p}, ..., z_{p-1,p})^T$ implies

$$p(\gamma_j = 1|\boldsymbol{\beta}) = \frac{N(\beta_j|0, v_1^2)\pi}{N(\beta_j|0, v_1^2)\pi + N(\beta_j|0, v_0^2)(1-\pi)}. \tag{3.4.26}$$

CHAPTER 4

Variational Inference

In high-dimensional settings such as omics data with multiple outcomes, the computational time needed to perform MCMC can often be prohibitively slow, even after quite restrictive assumptions. Variational Inference (VI) is an alternative approach to produce posterior information at a much reduced computational cost. By approximating the posterior through optimization, the speed of computing the posterior is improved at the cost of a loss of accuracy, as samples from the proxy conditional density are not from the "exact" posterior.

4.1 Evidence Lower Bound Optimisation

A family $\mathcal{D}$ of densities is specified over the latent variables. Each $q(\mathbf{z}) \in \mathcal{D}$ is a candidate approximation to the exact conditional $p(\mathbf{z}|\mathbf{y})$, where $\mathbf{z}$ are the latent variables (or parameters) and $\mathbf{y}$ is the observed data. The aim is to find the candidate probability density which is closest

in Kullback-Leibler (KL) divergence to the exact conditional distribution

$$q^*(\mathbf{z}) = \underset{q^*(\mathbf{z})\in\mathcal{D}}{\arg\min}\ KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{y})) \tag{4.1.1}$$

$$= \underset{q^*(\mathbf{z})\in\mathcal{D}}{\arg\min} \int q(\mathbf{z}) \log\left(\frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{y})}\right) d\mathbf{z}. \tag{4.1.2}$$

As Equation (4.1.1) contains the very posterior density $p(\mathbf{z}|\mathbf{y})$ we wish to avoid, we rearrange to form the Evidence Lower Bound or ELBO ($\mathcal{L}$). Maximising the ELBO is equivalent to minimizing the KL divergence

$$\mathcal{L}(q) = \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{z}, \mathbf{y})] - \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})]. \tag{4.1.3}$$

By rewriting the ELBO (4.1.3) as a sum of the expected log likelihood of the data and the KL divergence between the prior $p(\mathbf{z})$ and $q(\mathbf{z})$, we are able to see that the variational objective mirrors the usual balance between likelihood and prior

$$\mathcal{L}(q) = \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{z})] + \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y}|\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})]$$

$$= \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y}|\mathbf{z})] - \mathrm{KL}[q(\mathbf{z})||p(\mathbf{z})]. \tag{4.1.4}$$

Which values of $\mathbf{z}$ will the ELBO encourage $q(\mathbf{z})$ to place its mass over? The first term is an expected likelihood, encouraging densities that place their mass on configurations of the latent variables that explain the observed data. The second term is the negative divergence between the variational density and the prior; it encourages densities close to the prior.

By expanding the KL divergence between the variational distribution $q(\mathbf{z})$ and target conditional $p(\mathbf{z}|\mathbf{y})$, the decomposition of the log marginal probability of the observed data, which holds for any choice of $\mathbf{z}$ is

$$\log p(\mathbf{y}) = \mathcal{L}(q) + KL(q(\mathbf{z}||p(\mathbf{z}|\mathbf{y}))). \tag{4.1.5}$$

Since $KL(q||p) \geq 0$, the ELBO forms a lower bound to $\log p(\mathbf{y})$.

4.2 Mean-Field Variational family

To complete the specification of the optimisation a variational family is required. The *mean field variational family* is often used, where the latent variables are mutually independent and each governed by a distinct factor in the variational density.

$$q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j) \tag{4.2.1}$$

Each latent variable z_j follows its own variational factor, the density $q_j(z_j)$ with its own finite variational parameter(s) called free parameter(s), which are the arguments of the ELBO.

The mean-field family is expressive because it can capture any marginal density of the latent variables, but it is unable to capture any correlation between them. The marginal variances of the approximation often under-represent those of the target density. The KL divergence from the approximation to the posterior penalizes placing mass in $q(\mathbf{z})$ on areas where $p(\mathbf{z}|\mathbf{y})$ has little mass but penalizes less the reverse (Figure 4.2.1).

In the simple bivariate normal case any correlation will twist the pdf, contorting the shape from a circle to an ellipse. If a mean field family is assumed across the two parameters of interest (z_1, z_2) the approximation cannot extend to the full shape of the pdf without placing lots of density in areas where the target density has little mass. Figure 4.2.1 illustrates the limitation of the mean field variational family in the case of a bivariate positively correlated Gaussian distribution. This property comes from examining the fraction $q(\mathbf{z})/p(\mathbf{z}|\mathbf{y})$ in Equation (4.1.2). This is infinite if $p(\mathbf{z}|\mathbf{y}) = 0$ and $q(\mathbf{z}) > 0$. In order to prevent the expression from exploding at the tails $q(\mathbf{z})$ must be heavier then $p(\mathbf{z}|\mathbf{y})$, inducing a "zero forcing characteristic" for $q(\mathbf{z})$.

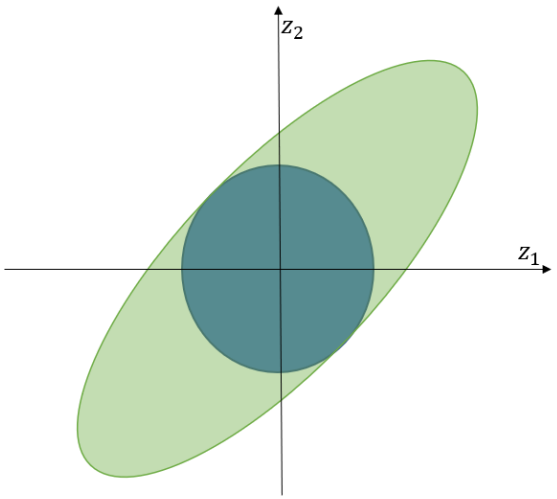

Figure 4.2.1: Image of a mean-field approximation to a two-dimensional Gaussian posterior with positive correlation where both distributions are 2σ contours of the Gaussian. The exact posterior is in green and the mean-field approximation is in blue. The ellipses shows the effect of a mean-field factorisation, where the variance of the approximate distribution has been underestimated.

The fully factorised approximation from (4.2.1) is attractive because it leads to a tractable optimisation problem to solve, but as described above, it is also very restrictive. The mean-field variational family can also handle vector variables. In each case, a multivariate conditional distribution is defined in terms of $p(\mathbf{y}|\mathbf{z}_j)$, and the corresponding factor $q(\mathbf{z}_j)$ will also be multivariate, rather than factorised with respect to the elements in the vector. This motivates *structured* or *fixed-form Variational Bayes*, where dependencies between parameters are explicitly incorporated within blocks and independence is retained across the blocks (Salimans and Knowles (2013), Bishop and Winn (2006), Hoffman and Blei (2015), Xing et al. (2002)). For example, in the case of explicit Bayesian variable selection in multivariate regression, an approximating posterior block which captures the natural dependency between the latent indicator variable γ_j and the corresponding regression coefficient β_j is

$$q(\beta_j, \gamma_j) = q(\beta_j|\gamma_j)q(\gamma_j). \tag{4.2.2}$$

This leads to a natural type of approximation for hierarchical Bayesian models, where the hierarchical structure of the prior often suggests a good hierarchical structure for the posterior

approximation.

4.3 Coordinate Ascent Mean-field Variational Inference

One approach for solving the optimisation of (4.1.3) is coordinate ascent mean-field variational inference. Each factor of the mean-field variational density is iteratively optimised while holding the others fixed, climbing the ELBO to a local optimum (Bishop, 2006). By using iterative expectations (Blei et al., 2017), the coordinate updates which maximises ELBO can be derived. First we rewrite the ELBO as

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y}, \mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] \\
&= \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{y}, \mathbf{z})] - \sum_{j=1}^{n} \mathbb{E}_{q(z_j)}[\log q(z_j)] \qquad (4.3.1) \\
&= \mathbb{E}_{q(z_j)}[\mathbb{E}_{q(\mathbf{z}_{-j})}[\log p(\mathbf{y}, \mathbf{z}_{-j}, z_j)|z_j]] - \sum_{j=1}^{m} \mathbb{E}_{q(z_j)}[\log q(z_j)]. \qquad (4.3.2)
\end{aligned}$$

Using the mutual independence of each variational density in (4.2.1), we can express the ELBO for the jth factor as

$$\mathcal{L}(q_j) = \mathbb{E}_{q(z_j)}[\mathbb{E}_{q(\mathbf{z}_{-j})}[\log p(\mathbf{y}, \mathbf{z}_{-j}, z_j)]] - \mathbb{E}_{q(z_j)}[\log q(z_j)] + \text{constant}. \qquad (4.3.3)$$

Rewriting (4.3.3) in terms of the negative KL divergence,

$$\begin{aligned}
\mathcal{L}(q_j) &\propto \mathbb{E}_{q(z_j)}[\log\big(\exp(\mathbb{E}_{q(\mathbf{z}_{-j})}[\log p(\mathbf{y}, \mathbf{z}_{-j}, z_j)])\big)] - \mathbb{E}_{q(z_j)}[\log q(z_j)] \\
&\propto D_{KL}(\exp(\mathbb{E}_{q(\mathbf{z}_{-j})}[\log p(\mathbf{y}, \mathbf{z}_{-j}, z_j)])||q(z_j)), \qquad (4.3.4)
\end{aligned}$$

Thus we maximise the ELBO with respect to $q_j(z_j)$ when we make the negative KL as small as

possible, which is when we set

$$q_j(z_j)^* \propto \exp(\mathbb{E}_{q(\mathbf{z}_{-j})}[\log p(\mathbf{y}, \mathbf{z}_{-j}, z_j)]) \tag{4.3.5}$$

$$\propto \exp(\mathbb{E}_{q(\mathbf{z}_{-j})}[\log p(z_j|\mathbf{y}, \mathbf{z}_{-j})]). \tag{4.3.6}$$

The *complete conditional* of the jth latent variable z_j is its conditional density given all the other latent and observed variables $p(z_j|\mathbf{y}, \mathbf{z}_{-j})$. The log of the optimal solution for factor $q_j(z_j)$ is obtained by taking the expectation with respect to all of the other factors $\{q_i(z_i)\}$ for $i \neq j$ which marginalises over the other densities, each weighted according to their respective probability density.

Algorithm 1: Coordinate ascent variational inference CAVI

Input : A model $p(\mathbf{y}, \mathbf{z})$, a data set $\mathbf{y}$

Output : A variational density $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$

Intialize: Variational factors $q_j(z_j)$

while *the ELBO has not converged* **do**

  for $j \in \{1, ..., m\}$ **do**

    Set $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j|\mathbf{z}_{-j}, \mathbf{y})]\}$

  end

  Compute ELBO(q)=$\mathbb{E}_q[\log p(\mathbf{z}, \mathbf{y})] - \mathbb{E}_q[\log q(\mathbf{z})]$

end

return $q(\mathbf{z})$

CAVI is performed by iterating through the variational factors from (4.3.6), maximising the ELBO with respect to each coordinate direction whilst fixing the other coordinate values. For each run we compute the ELBO, using Equation (4.3.1) with the updated free parameters, until this converges to the local optimum (Algorithm 1). The coordinate ascent updates can therefore also be obtained by taking the partial derivative with respect to the free parameter (local and global), holding the other parameters fixed, as this achieves the same ascent over the ELBO. This property helps motivate stochastic variational inference (SVI) which traverses the ELBO in the

direction of the natural gradient.

The CAVI algorithm (Algorithm 1), is very similar to the Gibbs sampler, in that both use the full conditionals, and update one parameter at a time. In each approach the variables can be vectors or arrays, thus allowing for correlation between parameters. Most recently Lee (2021) outlined common structure between the two schemes, from a set-theory perspective.

Posterior approximations from structured mean-field are often more accurate than a factorized approximation (where each latent variable is independent). However, the requirement of being able to evaluate the joint expectations analytically with respect to the grouped variables within the block $q(z_{j1}, ..., z_{jb})$ is often very restrictive. Hoffman and Blei (2015) incorporate a variety of dependencies between a vector of global variables and each set of local variables, by exploring different mean-field structures, to identify the properties of the respective updates. In allowing each vector of local variables to depend on the global variables, the lower bound contains expectations that are no longer possible to compute. To optimise the ELBO, a Monte Carlo expectation is incorporated into the algorithm.

Suppose each *complete conditional*, which is used to update CAVI, is in the exponential family form. Each optimal variational factor is then in the same parametric form as its corresponding complete conditional (Hoffman et al., 2013) (Appendix 4.9.4), making it easier to derive the corresponding CAVI algorithm and enabling VI to be scaled up to massive data.

4.4 Understanding CAVI with an EM Comparison

VI is often compared to the frequentist Expectation Maximisation (EM) algorithm commonly used to compute the maximum likelihood (ML) estimate in the presence of missing data. The approach involves augmenting the log likelihood $\ell(\theta; \mathbf{y})$ with latent variables $\mathbf{z}$ and taking the expectation with respect to $p(\mathbf{z}|\mathbf{y}, \theta^{(t)})$ from the previous iteration

$$Q(\theta|\theta^{(t)}) = \mathbb{E}[\ell(\theta; \mathbf{y}, \mathbf{z})|\mathbf{y}, \theta^{(t)}] \tag{4.4.1}$$

and then maximising $Q(\theta|\theta^{(t)})$ with respect to θ to get

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}). \tag{4.4.2}$$

A more detailed explanation is in Appendix 4.9.1. The log-likelihood can be decomposed into

$$\begin{aligned}\log p(\mathbf{y}|\theta) &= \int \log\left(\frac{p(\mathbf{y},\mathbf{z}|\theta)}{q(\mathbf{z})}\right) q(\mathbf{z})d\mathbf{z} - \int \log\left(\frac{p(\mathbf{z}|\mathbf{y},\theta)}{q(\mathbf{z})}\right) q(\mathbf{z})d\mathbf{z} \\ &= \mathcal{L}(q,\theta) + KL(q||p)\end{aligned} \tag{4.4.3}$$

where $q(\mathbf{z})$ is any probability distribution and $KL(q||p)$ is the KL divergence between $p(\mathbf{z}|\mathbf{y},\theta)$ and $q(\mathbf{z})$. As $KL(q||p) \geq 0$, $\mathcal{L}(q,\theta)$ is a lower bound of the log-likelihood. The expression of the marginal likelihood (4.4.3) is similar to (4.1.5) in VI, with the addition of a frequentist parameter argument θ alongside the q probability distribution in the lower bound function.

The E-step can thus be viewed as maximising $\mathcal{L}(q,\theta)$ with respect to the $q(\mathbf{z})$ argument. Just as in the VI case, this is maximized when $KL(q||p) = 0$, but now $q(\mathbf{z})$ is equal to a posterior distribution conditional on frequentist parameter values from the previous iteration $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{y},\theta^{(t)})$. In the subsequent M-step, $q(\mathbf{z})$ is held fixed and the lower bound $\mathcal{L}(q,\theta)$ is maximised with respect to the θ argument to give some new value $\theta^{(t+1)}$.

This decomposition makes the choice of $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{y},\theta^{(t)})$ in the EM algorithm explicit. While $p(\mathbf{z}|\mathbf{y},\theta)$ maybe easier to infer than $p(\mathbf{y}|\theta)$, in many problems this is not possible. The requirement can be avoided by using mean field theory to find approximate solutions for q instead, which gives rise to the Variational EM algorithm (Beal and Ghahramani, 2003).

If a mean field variational family (4.2.1) is assumed, $\mathcal{L}(q,\theta)$ can be rearranged in terms of $q_j(z_j)$

as

$$\begin{aligned}\mathcal{L}(q,\theta) &= \int \prod_i q_i(z_i)\Big[\log p(\mathbf{y},\mathbf{z}|\theta) - \sum_i \log(q_i(z_i))\Big]d\mathbf{z} \\ &= \int q(z_j)\log\big(\exp\big(\mathbb{E}_{q(\boldsymbol{z}_{-j})}[\log p(\mathbf{y},\boldsymbol{z}|\theta)]\big)\big)dz_j - \int q_j(z_j)\log q_j(z_j)dz_j + \\ &\quad - \sum_{i\neq j}\int q_i(z_i)\log q_i(z_i)dz_i \\ &= -KL(q_j(z_j)||\tilde{p}(\mathbf{y},\mathbf{z}|\theta)) - \sum_{i\neq j}\int q_i \log q_i \; dz_i \end{aligned} \tag{4.4.4}$$

where

$$\tilde{p}(\mathbf{y},\mathbf{z}|\theta) = \exp\left(\int \log p(\mathbf{y},\mathbf{z}|\theta)\prod_{i\neq j} q_i(z_i)dz_i\right). \tag{4.4.5}$$

The bound in (4.4.4) is maximised when the KL distance becomes zero, as is the case for $q_j(z_j) = \tilde{p}(\mathbf{y}, z_j|\theta)$, making the optimal distribution

$$q_j(z_j)^* \propto \exp\big(\mathbb{E}_{q(\mathbf{z}_{-j})}[\log p(\mathbf{y},\mathbf{z}|\theta)]\big). \tag{4.4.6}$$

which is similar to the VI update (4.3.5).

4.4.1 Mixture of Gaussians example

A Gaussian mixture model example is used to highlight the VI concepts discussed in the chapter. A detailed exposition is in Appendix 4.9.2. Consider a mixture of univariate Gaussian distributions. There are k mixture components, corresponding to k Gaussian distributions with means $\boldsymbol{\mu} = \{\mu_1, ..., \mu_k\}$ and variances $\boldsymbol{\sigma}^2 = \{\sigma_1^2, ..., \sigma_k^2\}$. To generate an observation y_i from the model, choose a cluster assignment with probability vector $\pi_1, ..., \pi_k$. ($\mathbf{z}_i$ as a k-vector indicator, all zeros except for a one in the position corresponding to y_i's cluster.)

The marginal likelihood is

$$L(\mathbf{y};\boldsymbol{\theta}) = \prod_{i=1}^{n}\sum_{j=1}^{k}\pi_j\left(\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left(-\frac{1}{2\sigma_j^2}(y_i-\mu_j)^2\right)\right), \tag{4.4.7}$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\pi})$. The likelihood can be augmented with a latent variable $\mathbf{z}$

$$\begin{aligned} p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta}) &= p(\mathbf{y}|\boldsymbol{\theta},\mathbf{z})p(\mathbf{z}) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{k}\left(\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left(-\frac{1}{2\sigma_j^2}(y_i-\mu_j)^2\right)\right)^{z_{ij}}\pi_j^{z_{ij}}. \end{aligned} \tag{4.4.8}$$

Parameter estimation can be achieved via the EM algorithm. The E-step involves the expectation with respect to $p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{(t)})$

$$\begin{aligned} \mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)},y_i] &= p(z_{ij}=1|y_i,\boldsymbol{\theta}^{(t)}) \\ &= \frac{p(y_i|z_{ij}=1,\boldsymbol{\theta}^{(t)})p(z_{ij}=1|\boldsymbol{\theta}^{(t)})}{p(y_i|\boldsymbol{\theta}^{(t)})} \\ &= \frac{\left(\frac{1}{\sqrt{2\pi\sigma_j^{2(t)}}}\exp\left(-\frac{1}{2\sigma_j^{2(t)}}(y_i-\mu_j^{(t)})^2\right)\right)\pi_j^{(t)}}{\sum_{j=1}^{k}\pi_j^{(t)}\left(\frac{1}{\sqrt{2\pi\sigma_j^{2(t)}}}\exp\left(-\frac{1}{2\sigma_j^{2(t)}}(y_i-\mu_j^{(t)})^2\right)\right.}. \end{aligned} \tag{4.4.9}$$

In the M-step, the expected complete log likelihood is maximised with respect to the parameters $\boldsymbol{\theta}$. Taking the corresponding partial derivatives equal to zero and using Lagrange multipliers for the constraint $\sum_j \pi_j = 1$, the following equations are derived for the updates of the M-step

$$\pi_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)},y_i] \tag{4.4.10}$$

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n}\mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)},y_i]y_i}{\sum_{i=1}^{n}\mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)},y_i]} \tag{4.4.11}$$

$$\sigma_j^{2(t+1)} = \frac{\sum_{i=1}^{n}\mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)},y_i](y_i-\mu_j^{(t)})^2}{\sum_{i=1}^{n}\mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)},y_i]}. \tag{4.4.12}$$

Alternatively, the full Bayesian hierarchical model can be posited where $\mathbf{z}$ and $\boldsymbol{\theta}$ in the likelihood (4.4.8) are treated as hidden random variables (denoted $\boldsymbol{\vartheta}$), with a prior specification of

$$\begin{aligned} \mu_j &\sim N(0, \tau^2) \quad \sigma_j^2 \sim IG(a, b), \\ \boldsymbol{\pi} &\sim Dir(\alpha_1, ..., \alpha_k), \\ \mathbf{z}_i &\sim \text{Multinomial } (1, \eta_1, ..., \eta_k), \end{aligned} \tag{4.4.13}$$

for the groups $j = 1, ..., k$ and observations $i = 1, ..., n$.

The assumed mean field variational family form of

$$q(\boldsymbol{\vartheta}) = \left\{ \prod_{i=1}^{n} q(\mathbf{z}_i) \right\} q(\boldsymbol{\pi}) \left\{ \prod_{j=1}^{k} q(\mu_j) q(\sigma_j^2) \right\}, \tag{4.4.14}$$

allows a dependency between the parameters within the vectors of $\boldsymbol{\pi}$ and $\mathbf{z}_i$. A choice of conjugate priors leads to the q approximating densities with local updates of

$$\begin{aligned} q(\mathbf{z}_i) &= \text{Multinomial}(1, \eta_1^*, ..., \eta_k^*) \\ \eta_j^* &= \frac{\frac{1}{\sqrt{2\pi(\sigma_j^2)^{(1)}}} \exp\left(-\frac{1}{2(\sigma_j^2)^{(1)}} (y_i - 2y_i(\mu_j)^{(1)} + (\mu_j)^{(2)}) \right) (\pi_j)^{(1)} \eta_j}{\sum_{j=1}^{K} \frac{1}{\sqrt{2\pi(\sigma_j^2)^{(1)}}} \exp\left(-\frac{1}{2(\sigma_j^2)^{(1)}} (y_i - 2y_i(\mu_j)^{(1)} + (\mu_j)^{(2)}) \right) (\pi_j)^{(1)} \eta_j}. \end{aligned} \tag{4.4.15}$$

where $(\cdot)^{(1)}$ denotes the q expectation with respect to all the other factors.The E-step in (4.4.9) is equivalent to the VI local parameter update in (4.4.15). The VI update substitutes the ML estimate with the q expectation and includes the hyperparameter η_j from the prior $p(\boldsymbol{\pi})$.

The global updates for the complete conditionals are

$$q(\mu_j) = N\left(\mu_j^* = \frac{\sum_{i=1}^n (z_{ij})^{(1)} y_i}{\sum_{i=1}^n (z_{ij})^{(1)} + (\sigma_j^2)^{(1)}/(\tau^2)^{(1)}}, \quad \tau_j^{2*} = \left(\frac{\sum_{i=1}^n (z_{ij})^{(1)}}{(\sigma_j^2)^{(1)}} + \frac{1}{(\tau^2)^{(1)}}\right)^{-1}\right) \tag{4.4.16}$$

$$q(\boldsymbol{\pi}) = Dir\left(\alpha_1 + \sum_{i=1}^n (z_{i1})^{(1)}, ..., \alpha_k + \sum_{i=1}^n (z_{ik})^{(1)}\right) \tag{4.4.17}$$

$$q(\sigma_j^2) = IG\left(a_j^* = \sum_{i=1}^n \frac{(z_{ij})^{(1)}}{2} + a, \quad b_j^* = b + \sum_{i=1}^n \frac{(z_{ij})^{(1)}(y_i - 2y_i(\mu_j)^{(1)} + (\mu_j)^{(2)})}{2}\right) \tag{4.4.18}$$

As the q densities approximate the posterior uncertainty around the parameters, we can obtain suitable estimators to compare with the equivalent EM updates, where the expectation of z_{ij} is denoted by $\mathbb{E}_q[z_{ij}|\boldsymbol{\vartheta}_{-j}, y_i]$

$$\mathbb{E}_q[\pi_j] = \frac{\alpha_j + \sum_{i=1}^n \mathbb{E}_q[z_{ij}|\boldsymbol{\vartheta}_{-j}, y_i]}{\sum_{j=1}^k \alpha_j + n} \tag{4.4.19}$$

$$\mathbb{E}_q[\mu_j] = \frac{\sum_{i=1}^n \mathbb{E}_q[z_{ij}|\boldsymbol{\vartheta}_{-j}, y_i] y_i}{\sum_{i=1}^n \mathbb{E}_q[z_{ij}|\boldsymbol{\vartheta}_{-j}, y_i] + (\sigma_j^2)^{(1)}/(\tau^2)^{(1)}} \tag{4.4.20}$$

$$\arg\max_{\sigma_j^2} q(\sigma_j^2; a_j^*, b_j^*) = \frac{2b + (y_i - 2y_i(\mu_j)^{(1)} + (\mu_j)^{(2)})}{2 + 2a + \sum_{i=1}^n \mathbb{E}_q[z_{ij}|\boldsymbol{\vartheta}_{-j}, y_i]}. \tag{4.4.21}$$

The maximisation step is equivalent to the global update in VI augmented with the hyperparameters from the respective priors.

The VI algorithm can be interpreted in terms of gradients of the local and global parameters. The E-step corresponds to setting the gradient of the local parameters equal to 0 by solving, given the value of the global parameters (equivalent to the coordinate move of the latent variable in the EM algorithm). In the M-step the gradient of the global parameters is set to 0 by the update, given the value of the local parameters.

4.4.2 Mixture of Gaussians estimation comparison

The EM algorithm is the preferred method for estimation of univariate and multivariate mixtures in the frequentist setting. The M-steps in the EM algorithm for the univariate mixture of normals

are (4.4.10) to (4.4.12), and in the multivariate case they are

$$\boldsymbol{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^n \mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)}, \mathbf{y}_i]\mathbf{y}_i}{\sum_{i=1}^n \mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)}, \mathbf{y}_i]}$$
$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^n \mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)}, \mathbf{y}_i](\mathbf{y}_i - \boldsymbol{\mu}_j^{(t)})(\mathbf{y}_i - \boldsymbol{\mu}_j^{(t)})^T}{\sum_{i=1}^n \mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)}, \mathbf{y}_i]}.$$

There are well known limitations with this ML estimation approach which do not apply in the Bayesian framework. The EM algorithm breaks down whenever $\sigma_j^{2(t+1)}$ is zero or $\boldsymbol{\Sigma}_j^{(t+1)}$ is singular or nearly singular, which happens when $\mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)}, \mathbf{y}_i]$ is close to zero for too many observations (indexed by i). Then at the next iteration the computation of $\mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t+1)}, \mathbf{y}_i]$ is no longer possible. Such difficulties arise in particular, if the EM algorithm is applied to a finite mixture of Gaussians overfitting the number of components.

A further difficulty with ML estimation for univariate mixtures of normals, first identified by Kiefer and Wolfowitz (1956), is that the mixture likelihood function (4.4.7) is unbounded and has many spurious modes. The unboundedness of the mixture likelihood function is also relevant for mixtures of multivariate normals, as each observation $\mathbf{y}_i$ gives rise to a singularity on the boundary of the parameter space. Thus the ML estimate as a global maximizer of the mixture likelihood function does not exist. Several local maximizers may exist for a given sample, and a major difficulty is to identify if the correct one has been found. However, Kiefer (1978) showed that a particular local maximizer of the mixture likelihood function is consistent, efficient and asymptomatically normal if the mixture is not overfitting.

To avoid these issues, Hathaway (1985) proposed the constrained ML estimation of univariate mixtures of normals based on the inequality constraint

$$\min_{k,j} \frac{\sigma_k}{\sigma_j} \geq c > 0. \tag{4.4.22}$$

and proves strong consistency of the resulting estimator. For a mixtures of normals, Hathaway (1985) constrains all eigenvalues of $\Sigma_k \Sigma_j^{-1}$ to be greater than a positive constant.

In the Bayesian approach, the use of a proper prior distribution on each component variance, usually in the form of an inverse gamma for univariate mixtures or inverse Wishart priors for multivariate mixtures, has two desirable effects. First, the conditional posterior distribution of the variance components is always proper. In the context of the Gibbs sampler, sampling yields a well-defined variance even if the group is empty or contains to few observations to obtain a well defined sample variance.

Second, the unbounded nature of the mixture likelihood function is caused by complete ignorance about the variance ratio (4.4.22). The priors in the Bayesian approach allows us to include some prior information on this ratio, however vague. In comparison to the likelihood, the posterior density will be more regular.

In the Bayesian paradigm, estimation of the model can be achieved either using an MCMC algorithm or VI. In MCMC methods, both the Gibbs sampler and the Metropolis-Hastings algorithm are often required in combination (Gormley and Murphy, 2010). As in any mixture model setting, the so called label switching problem (Stephens (2000a) and Frühwirth-Schnatter (2011)) must be considered when employing such algorithms. This is the non-identifiability of a finite mixture distribution caused by the invariance of a mixture distribution to relabelling the components. In our example $k = 2$, $\boldsymbol{\theta}_k = (\mu_k, \sigma_k^2)$ and $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \pi_1, \pi_2)$. If $\boldsymbol{\theta_1} \neq \boldsymbol{\theta}_2$ and $\boldsymbol{\vartheta}^* = (\boldsymbol{\theta}_2, \boldsymbol{\theta}_1, \pi_2, \pi_1)$, which is obtained by interchanging the order of the components, then the distribution induced by $\boldsymbol{\vartheta}$ and $\boldsymbol{\vartheta}^*$ is the same although the two parameters are distinct

$$\begin{aligned} p(y_i|\boldsymbol{\vartheta}^*) =& \pi_2 f_N(y_i; \mu_2, \sigma_2^2) + \pi_1 f_N(y_i; \mu_1, \sigma_1^2) = \\ & \pi_1 f_N(y_i; \mu_1, \sigma_1^2) + \pi_2 f_N(y_i; \mu_2, \sigma_2^2) = p(y_i|\boldsymbol{\vartheta}). \end{aligned}$$

Because of this invariance, a mixture of two normals is not identifiable in the strict sense (Rothenberg, 1971). For the general finite mixture distribution with k components, there exists $k!$ equivalent ways of arranging the components. The posterior distributions are thus always multimodal, with a multiple of $k!$ symmetric modes in the case of exchangeable priors. This can lead to convergence issues as the Markov chains may have trouble visiting all these modes in a symmetric

manner, despite the symmetry being guaranteed from the shape of the posterior.

In order to obtain an identifiable model for inference, formal identifiability constraints can be imposed. However, this may not lead to unique labelling (Celeux (1998), Stephens (2000b)) and paradoxically, prevents any formal claim that the MCMC has converged. Alternative approaches include random permutation of the labels (Frühwirth-Schnatter, 2001) and more sophisticated and complex MCMC methods to improve mix of the sampler (Celeux et al., 2000). The label switching issue is partially bypassed in the VI approach, which relies on scaling the slope of the ELBO rather than exploring the multi-modal posterior space, to reach a local optimum. This is analogous to the MCMC approach, when insufficient proposal variance prevents the sampler from leaving the local optimum.

4.5 ELBO and the Natural Gradient

Up to now all latent variables, either global or local, have been defined as $\mathbf{z}$. For clarity, a vector of local latent variables $\boldsymbol{\gamma}$ is introduced (such as the indicator $\mathbf{z}_i$ in the mixture modelling example 4.4.1) and a vector of global parameters $\boldsymbol{\beta}$ with hyperparameters $\boldsymbol{\alpha}$ (which are "natural" parameters of the exponential family form). The updates for the hyperparameters have been excluded as these will just be a function of the global and local parameters.

The variational posterior for the latent variables $\boldsymbol{\gamma}_i$, governed by the local parameters $\boldsymbol{\phi}_i$, is $q(\boldsymbol{\gamma}_i|\boldsymbol{\phi}_i)$ and the variational posterior for the vector of global parameters is $q(\boldsymbol{\beta}|\boldsymbol{\lambda})$ with the "global free parameters" $\boldsymbol{\lambda}$. The joint posterior is

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{y}) = p(\boldsymbol{\beta}) \prod_{i=1}^{n} p(\boldsymbol{\gamma}_i, y_i|\boldsymbol{\beta}). \tag{4.5.1}$$

Choosing conjugate priors to ensure the complete conditionals are in the exponential family

$$p(\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\gamma},\boldsymbol{\alpha}) = h(\boldsymbol{\beta})\exp\big(\eta_g(\boldsymbol{\gamma},\mathbf{y},\boldsymbol{\alpha})^T t(\boldsymbol{\beta}) - a_g(\eta_g(\boldsymbol{\gamma},\mathbf{y},\boldsymbol{\alpha}))\big),$$
$$p(\gamma_{ij}|y_i,\boldsymbol{\beta},\boldsymbol{\gamma}_{i,-j}) = h(\gamma_{ij})\exp\big(\eta_l(y_i,\boldsymbol{\beta},\boldsymbol{\gamma}_{i,-j})^T t(\gamma_{ij}) - a_l(\eta_l(y_i,\boldsymbol{\beta},\boldsymbol{\gamma}_{i,-j})),\big)$$

and specifying the mean-field variational form of

$$q(\boldsymbol{\gamma},\boldsymbol{\beta}) = q(\boldsymbol{\beta}|\boldsymbol{\lambda})\prod_{i=1}^{n}\prod_{j=1}^{k} q(\gamma_{ij}|\boldsymbol{\phi}_{ij}), \tag{4.5.2}$$

where the approximating q distributions are also in the exponential family form, allows the ELBO to be expressed as a function of the global natural free parameters $\boldsymbol{\lambda}$ (using $\mathbb{E}_{q(\boldsymbol{\beta}|\boldsymbol{\lambda})}[t(\boldsymbol{\beta})] = \nabla_{\boldsymbol{\lambda}} a_g(\boldsymbol{\lambda})$)

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\lambda}) &\propto \mathbb{E}_q[\log p(\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\gamma})] - \mathbb{E}_q[\log q(\boldsymbol{\beta})] \\
&\propto \mathbb{E}_q[\eta_g(\boldsymbol{\gamma},\mathbf{y},\boldsymbol{\alpha})^T t(\boldsymbol{\beta})] - \mathbb{E}_q[a_g(\eta_g(\boldsymbol{\gamma},\mathbf{y},\boldsymbol{\alpha}))] - \boldsymbol{\lambda}^T\mathbb{E}_q[t(\boldsymbol{\beta})] + a_g(\boldsymbol{\lambda}) \\
&\propto \mathbb{E}_{q(\boldsymbol{\gamma}|\boldsymbol{\phi})}[\eta_g(\boldsymbol{\gamma},\mathbf{y},\boldsymbol{\alpha})]\nabla_\lambda a_g(\boldsymbol{\lambda}) - \boldsymbol{\lambda}^T\nabla_\lambda a_g(\boldsymbol{\lambda}) + a_g(\boldsymbol{\lambda}).
\end{aligned}$$

The global coordinate ascent update is determined by taking the derivative with respect to $\boldsymbol{\lambda}$, setting it to zero and solving. The CAVI parameter updates, in their exponential family form, are thus

$$\boldsymbol{\lambda} = \mathbb{E}_{q(\boldsymbol{\gamma}|\boldsymbol{\phi})}[\eta_g(\boldsymbol{\gamma},\mathbf{y},\boldsymbol{\alpha})] = \big[\boldsymbol{\alpha}_1 + \sum_{i=1}^{N}\mathbb{E}_{q(\boldsymbol{\gamma}_i|\boldsymbol{\phi}_i)}[t(\boldsymbol{\gamma}_i,y_i)], \alpha_2 + n\big]^T, \tag{4.5.3}$$
$$\boldsymbol{\phi}_{ij} = \mathbb{E}_{q(\beta|\lambda)}[\eta_l(y_i,\boldsymbol{\beta},\boldsymbol{\gamma}_{i,-j})]. \tag{4.5.4}$$

The local update is found by applying the same approach to $\mathcal{L}(\gamma_{ij})$.

An alternative to the CAVI is ascent by natural gradient. Gradient ascent relies on the Euclidean distance metric which is not suitable for the ELBO as the optimisation objective is with respect to the probability measure, (4.1.3). The natural gradient accounts for the geometric structure of probability parameters (Amari, 1998) by warping the parameter space so that moving the

same distance in different directions amounts to equal change in symmetrized KL divergence. In conditionally conjugate models, the natural gradient of the global parameters is calculated by premultiplying the gradient of the ELBO (with respect to the global parameter)

$$\nabla_\lambda \mathcal{L}(\boldsymbol{\lambda}) = \nabla^2_\lambda a_g(\boldsymbol{\lambda}) \big(E_{q_{(\gamma|\phi)}}[\eta_g(\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\alpha})] - \boldsymbol{\lambda} \big) \tag{4.5.5}$$

by the inverse of the Fisher information of $q(\boldsymbol{\beta}|\boldsymbol{\lambda})$ (inverse covariance matrix of the sufficient statistic $(\nabla^2_\lambda a_g(\boldsymbol{\lambda}))^{-1}$ or Riemannian metric (Amari, 1982))

$$\begin{aligned} g(\boldsymbol{\lambda}) &= \big(\nabla^2_\lambda a_g(\boldsymbol{\lambda})\big)^{-1} \nabla^2_\lambda a_g(\boldsymbol{\lambda}) \big(E_{q_{(\gamma|\phi)}}[\eta_g(\boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\alpha})] - \boldsymbol{\lambda} \big) \\ &= E_{q_{(\gamma|\phi)}}[\eta_g(\boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\alpha})] - \boldsymbol{\lambda}. \end{aligned} \tag{4.5.6}$$

The local updates are computed in the same fashion.

In a gradient based optimisation algorithm, for each iteration optimise the local parameters first, then update global parameters by small increments ϵ_m in the direction of the natural gradient conditional on the local parameter updates

$$\begin{aligned} \boldsymbol{\lambda}_{m+1} &= \boldsymbol{\lambda}_m + \epsilon_{m+1} g(\boldsymbol{\lambda}_m) \\ &= (1 - \epsilon_{m+1})\boldsymbol{\lambda}_m + \epsilon_{m+1} \Big[\boldsymbol{\alpha}_1 + \sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{\gamma}_i|\phi_i)}[t(\boldsymbol{\gamma}_i, y_i)], \alpha_2 + n \Big]^T \\ &= (1 - \epsilon_{m+1})\boldsymbol{\lambda}_m + \epsilon_{m+1} \hat{\boldsymbol{\lambda}}_m. \end{aligned} \tag{4.5.7}$$

4.6 Stochastic Variational Inference

The natural gradient has the same computational cost as the coordinate update, it still requires summing over the entire data set to re-estimate the global variational free parameters. Stochastic variational inference (SVI) solves this problem by using the natural gradient in a stochastic optimisation algorithm. A subsample of the data is repeatedly taken to form noisy but cheap

to compute estimates of the natural gradient of the ELBO, which are followed with a decreasing step size. The subsample may comprise a single draw (or more), where the update is a weighted average of the current and new update (4.6.1). This is equivalent to a CAVI update where the date set comprises n replicates of the sampled data point $(y_i, \boldsymbol{\gamma}_i)$,

$$\hat{g}(\boldsymbol{\lambda}) = \boldsymbol{\alpha} + n\mathbb{E}_{q(\gamma_i|\phi_i)}[t(\boldsymbol{\gamma}_i, y_i), 1]^T - \boldsymbol{\lambda}, \tag{4.6.1}$$

where the local parameters are for the single randomly sampled data point.

Algorithm 2: SVI for Conditionally Conjugate Models

Input : A model $p(\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\beta})$, a data set $\mathbf{y}$, a step size schedule for ϵ_m.

Output : Global $q_\lambda(\boldsymbol{\beta}|\lambda)$ and local $\prod_i \prod_j q(\gamma_{ij}|\boldsymbol{\phi}_{ij})$ variational densities.

Intialize: Variational parameters $\boldsymbol{\lambda}_0$, $\boldsymbol{\phi}_0$. Number of iterations m.

for $m = 1, ..., \infty$ **do**

Sample a data point $y_i^{(r)}$ randomly, $i \sim \text{Unif}(1, ..., n)$, from the data set. Optimize the associated local variational parameters:

$$\phi_{ij} = \mathbb{E}_{q(\beta|\lambda^{(m)})}[\eta_l(y_i^{(r)}, \boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}_{i,-j}^{(r)})] \ \forall \ j.$$

Compute the intermediate global parameter as though $y_i^{(r)}$ had been replicated n times:

$$\hat{\boldsymbol{\lambda}}_m = \mathbb{E}_{q(\gamma_i|\phi_i)}[\eta_g(y_i^{(r)}, \boldsymbol{\gamma}_i^{(r)}, \boldsymbol{\alpha})] = [\boldsymbol{\alpha_1} + n\mathbb{E}_{q(\boldsymbol{\gamma}_i|\boldsymbol{\phi}_i)}[t(y_i^{(r)}, \boldsymbol{\gamma}_i^{(r)})], \ \alpha_2 + n].$$

Update the current estimate of the variational parameter (which computes the natural gradient):

$$\boldsymbol{\lambda}_{m+1} = (1 - \epsilon_{m+1})\boldsymbol{\lambda}_m + \epsilon_{m+1}\hat{\boldsymbol{\lambda}}_m.$$

end

return $\boldsymbol{\lambda}$, $\boldsymbol{\phi}$

The global parameters are updated by replacing $g(\boldsymbol{\lambda})$ in (4.5.7) with $\hat{g}(\boldsymbol{\lambda})$. The step size sequence are set to satisfy the conditions of Robbins and Monro (1951) to guarantee that the algorithm

converges to a local optimum, as the ELBO is convex

$$\sum_m \epsilon_m = \infty; \quad \sum_m \epsilon_m^2 < \infty. \tag{4.6.2}$$

In SVI the global parameter updates are now a function of their previous value rather than the values of all the other parameters (CAVI), the pseudocode is in Algorithm 2. There are several ways to parameterise the learning rate which satisfy (4.6.2), Hoffman et al. (2013) set

$$\epsilon_m = (m + \tau)^{-\kappa}. \tag{4.6.3}$$

The forgetting rate $\kappa \in (0.5, 1]$ controls how quickly old information is forgotten and the delay $\tau \geq 0$, down-weights early iterations.

To improve its stability, the SVI algorithm can be extended to multiple samples (mini batches) where S samples of the data are made $\mathbf{y}_{m,1:S}$ with or without replacement. This is particularly important when the dimensions of the response extend beyond 1 dimension.

The mini-batch must be drawn uniformly at random with size S satisfying $1 \leq S << n$. Larger values of S reduce the variance of the stochastic natural gradient. Computational savings are obtained when $S << n$, when $S = n$ the SVI reduces to CAVI when the learning rate is set to 1.

At each iteration compute the local variational parameters $\boldsymbol{\phi}_s(\boldsymbol{\lambda}_m)$ for each data point, compute the intermediate global parameters $\hat{\boldsymbol{\lambda}}_s$ for each data point $y_{m,s}$

$$\boldsymbol{\lambda}_{m+1} = (1 - \epsilon_{m+1})\boldsymbol{\lambda}_m + \frac{\epsilon_{m+1}}{S} \sum_s \hat{\boldsymbol{\lambda}}_s, \tag{4.6.4}$$

and finally average the $\hat{\boldsymbol{\lambda}}_s$ in the update (Hoffman et al., 2013). The stochastic natural gradients associated with each point y_s have an expected value equal to the gradient. Therefore, the average of these stochastic natural gradients has the same expectation and the algorithm remains valid.

Mixture of Gaussians example

Returning to our mixture of Gaussians example in Section (4.4.1), the SVI updates leads to sampling a data point y_i uniformly from the data set and computing the local variational update using (4.4.15). The global parameters are computed as though y_i is replicated n times (as a batch contains a single sample)

$$\begin{aligned}
q(\mu_j) =& N\left(\mu_j^* = \frac{n(z_{ij})^{(1)} y_i}{n(z_{ij})^{(1)} + (\sigma_j^2)^{(1)}/(\tau^2)^{(1)}}, \quad \tau_j^{2*} = \left(\frac{n(z_{ij})^{(1)}}{(\sigma_j^2)^{(1)}} + \frac{1}{(\tau^2)^{(1)}}\right)^{-1}\right) \\
q(\boldsymbol{\pi}) =& Dir\left(\alpha_1 + n(z_{i1})^{(1)}, ..., \alpha_k + n(z_{ik})^{(1)}\right) \\
q(\sigma_j^2) =& IG\left(a_j^* = \frac{n(z_{ij})^{(1)}}{2} + a, \quad b_j^* = b + n\frac{(z_{ij})^{(1)}(y_i - 2y_i(\mu_j)^{(1)} + (\mu_j)^{(2)})}{2}\right)
\end{aligned}$$

The variational parameters are then mapped to their exponential family natural form, in the case of $q(\mu_j)$ the natural parameters in the form of

$$\begin{bmatrix} \frac{\mu_j^*}{\tau_j^{2*}} \\ -\frac{1}{2\tau^{*2}} \end{bmatrix}, \tag{4.6.5}$$

and updated using (4.6.4).

4.7 Adaptive Learning Rates and Mini-batches

The convergence speed is influenced by the choice of the learning rate ϵ_m and the mini-batch size. Due to the law of large numbers, as the size of the mini-batch increases the noise of the stochastic gradient reduces, allowing larger learning rates. The learning procedure is improved by optimally adapting the learning rate for a fixed batch size, rather than optimally adapting the mini-batch size for a given learning rate.

Using the method developed in Ranganath et al. (2013), the learning rates ϵ_m can be adapted to the sampled data by minimising the expected distance between the stochastic update $\boldsymbol{\lambda}_{m+1}$ in

(4.6.4) to the optimal global variational parameter (CAVI update) $\boldsymbol{\lambda}_m^*$.

$$\begin{aligned}\boldsymbol{\lambda}_m^* &= \mathbb{E}_{q(\gamma|\phi)}[\eta_g(\boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\alpha})] \\ &= \Big[\boldsymbol{\alpha}_1 + \sum_{i=1}^{n} \mathbb{E}_{q(\gamma_i|\phi_i)}[t(\boldsymbol{\gamma}_i, y_i)], \alpha_2 + n\Big]^T \\ &= \boldsymbol{\alpha} + \sum_{i=1}^{n} \bar{t}_{\phi_i^{\lambda_m}}(y_i), \end{aligned} \tag{4.7.1}$$

where $\bar{t}_{\phi_i^{\lambda_m}}(y_i)$ is the vector $(\mathbb{E}_{q(\gamma_i|\phi_i)}[t(\gamma_i, y_i)], 1)$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \alpha_2)$. The learning rate is estimated by minimising the expected error between the cheaper stochastic update $\boldsymbol{\lambda}_{m+1}$ and the expensive batch update $\boldsymbol{\lambda}_m^*$.

Defining the squared norm of the error as

$$J(\epsilon_m) \triangleq (\boldsymbol{\lambda}_{m+1} - \boldsymbol{\lambda}_m^*)^T(\boldsymbol{\lambda}_{m+1} - \boldsymbol{\lambda}_m^*), \tag{4.7.2}$$

where the intermediate global parameter update is

$$\hat{\boldsymbol{\lambda}}_m = \boldsymbol{\alpha} + n\bar{t}_{\phi_i^{\lambda_m}}(y_i), \tag{4.7.3}$$

the adapting learning rate ϵ_m^* is obtained by minimizing $\mathbb{E}_n[J(\epsilon_m|\boldsymbol{\lambda}_m)]$. This leads to a stochastic update that is close in expectation to the batch update.

After conditioning on $\boldsymbol{\lambda}_m$, the randomness in $J(\epsilon_m)$ comes from the intermediate global parameter $\hat{\boldsymbol{\lambda}}_m$. Its mean and covariance (Appendix 4.9.6) are

$$\begin{aligned}\mathbb{E}_n[\hat{\boldsymbol{\lambda}}_m|\boldsymbol{\lambda}_m] &= \boldsymbol{\lambda}_m^*, \\ \mathrm{Cov}_n[\hat{\boldsymbol{\lambda}}_m|\boldsymbol{\lambda}_m] &= \mathbb{E}_n[(\hat{\boldsymbol{\lambda}}_m - \boldsymbol{\lambda}_m^*)(\hat{\boldsymbol{\lambda}}_m - \boldsymbol{\lambda}_m^*)^T] \triangleq \Sigma.\end{aligned}$$

Minimizing $\mathbb{E}_n[J(\epsilon_m|\boldsymbol{\lambda}_m)]$ with respect to ϵ_m (Appendix 4.9.6) gives

$$\epsilon_m^* = \frac{(\boldsymbol{\lambda}_m^* - \boldsymbol{\lambda}_m)^T(\boldsymbol{\lambda}_m^* - \boldsymbol{\lambda}_m)}{(\boldsymbol{\lambda}_m^* - \boldsymbol{\lambda}_m)^T(\boldsymbol{\lambda}_m^* - \boldsymbol{\lambda}_m) + tr(\Sigma)} \tag{4.7.4}$$

The learning rate ϵ_m^* shrinks through the trace term, when the intermediate parameter has a high variance around the batch update $\boldsymbol{\lambda}_m^*$. The learning rate grows when the batch update $\boldsymbol{\lambda}_m^*$ is far from the current parameter $\boldsymbol{\lambda}_m$. This learning rate however depends on the batch update $\boldsymbol{\lambda}_m^*$ and the variance of the intermediate parameters around it, both unknown quantities. The adaptive learning rate involves estimating these quantities.

Let $\hat{g}(\boldsymbol{\lambda}_m)$ be the sampled natural gradient defined in (4.6.1). The expected value of the difference between the current parameter and the intermediate global update is

$$\mathbb{E}_n[\hat{\boldsymbol{\lambda}}_m - \boldsymbol{\lambda}_m | \boldsymbol{\lambda}_m] = \mathbb{E}_n[\hat{g}(\boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m] = -\boldsymbol{\lambda}_m + \boldsymbol{\lambda}_m^*. \tag{4.7.5}$$

Its covariance is equal to the covariance of the intermediate parameters $\hat{\boldsymbol{\lambda}}_m$

$$\text{Cov}_n[\hat{g}(\boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m] = \text{Cov}_n[\hat{\boldsymbol{\lambda}}_m|\boldsymbol{\lambda}_m] = \Sigma \tag{4.7.6}$$

which allows the denominator of the adaptive learning rate to be expressed as

$$\mathbb{E}_n[\hat{g}(\boldsymbol{\lambda}_m)^T\hat{g}(\boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m] = \mathbb{E}_n[\hat{g}(\boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m]^T\mathbb{E}_n[\hat{g}(\boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m] + tr(\Sigma).$$

The adaptive learning rate in Equation (4.7.4) can be rewritten as

$$\epsilon_m^* = \frac{\mathbb{E}_n[\hat{g}(\boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m]^T\mathbb{E}_n[\hat{g}(\boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m]}{\mathbb{E}_n[\hat{g}(\boldsymbol{\lambda}_m)^T\hat{g}(\boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m]}. \tag{4.7.7}$$

The expectations can be approximated within the stochastic algorithm with moving averages Schaul et al. (2013). Let the moving averages for $\mathbb{E}_n[\hat{g}(\boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m]$ and $\mathbb{E}_n[\hat{g}(\boldsymbol{\lambda}_m)^T\hat{g}(\boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m]$ be denoted by $\bar{g}_m$ and $\bar{h}_m$ respectively. Let τ_m be the window size of the exponential moving average at time t. The updates are

$$\mathbb{E}_n[\hat{g}(\boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m] \approx \bar{g}_m = (1 - \tau_m^{-1})\bar{g}_{m-1} + \tau_m^{-1}\hat{g}(\boldsymbol{\lambda}_m) \tag{4.7.8}$$

$$\mathbb{E}_n[\hat{g}(\boldsymbol{\lambda}_m)^T\hat{g}(\boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m] \approx \bar{h}_m = (1 - \tau_m^{-1})\bar{h}_{m-1} + \tau_m^{-1}\hat{g}(\boldsymbol{\lambda}_m)^T\hat{g}(\boldsymbol{\lambda}_m). \tag{4.7.9}$$

Plugging these into (4.7.7), the adaptive learning rate can be approximated with

$$\epsilon_m^* \approx \frac{\bar{g}_m^T \bar{g}_m}{\bar{h}_m}.$$

As the moving averages are less reliable after larger steps, the memory size are updated using

$$\tau_{m+1} = \tau_m(1 - \epsilon_m^*) + 1 \tag{4.7.10}$$

The description of the adaptive learning rates assumes a single data point, but this generalises easily using

$$\hat{\boldsymbol{\lambda}}_m = \sum_{s=1}^{S} \frac{\hat{\boldsymbol{\lambda}}_s}{S}, \tag{4.7.11}$$

where $\hat{\boldsymbol{\lambda}}_s$ is the intermediate parameter for the s sampled data point and S is the size of the mini-batch.

The moving averages are initialised by Monte Carlo estimates of the expectations at the initialization of the global parameters $\boldsymbol{\lambda}_1$ and τ_1 is initialised to be the number of samples used to construct the Monte Carlo estimate. The full algorithm is in Appendix 4.9.5.

4.8 Modern Variational Inference

A major issue that often arises in mean field variational inference is that not all expectations in the sum of the log likelihood terms are available in closed form. For notation simplicity, the log joint likelihood of the latent variables $\mathbf{z}$ and the data $\mathbf{y}$, given the hyperparameters $\boldsymbol{\alpha}$, is expressed as

$$\log p(\mathbf{z}, \mathbf{y}|\boldsymbol{\alpha}) = \sum_k f_k(\mathbf{z}_{B_k}, \mathbf{y}_{A_k}) \tag{4.8.1}$$

where A_k indexes the data appearing in function k, B_k indexes the latent variables appearing in function k and $\boldsymbol{\alpha}$ is dropped for simplicity as it is fixed. The index k corresponds to groups of units within the log joint likelihood, rather than variables or distributions. The ELBO as a function

of the latent variables $\mathbf{z}$ (as opposed to separate local and global variables) and free variational parameters $\boldsymbol{\lambda}$ is thus expressed as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\lambda}) &= \mathbb{E}_{q(\mathbf{z}|\lambda)}[\log p(\mathbf{z}, \mathbf{y}|\boldsymbol{\alpha})] - \mathbb{E}_{q(\mathbf{z}|\lambda)}[\log q(\mathbf{z}|\boldsymbol{\lambda})] \\ &= \sum_k \mathbb{E}_{q(\mathbf{z}_k|\lambda_k)}[f_k(\mathbf{z}_{B_k}, \mathbf{y}_{A_k})] - \sum_j \mathbb{E}_{q(z_j|\lambda_j)}[\log q_j(z_j|\boldsymbol{\lambda}_j)]. \end{aligned} \tag{4.8.2}$$

For each function f_k those $z_j \notin \mathbf{z}_{B_k}$ will have their corresponding q_j removed from the expectation. For those $z_j \in \mathbf{z}_{B_k}$, the expectation of f_k results in a new function of variational parameters $\boldsymbol{\lambda}_j \in \boldsymbol{\lambda}_{B_k}$.

A typical solution to an intractable expectation in (4.8.2) is to replace the problematic function with a nicer functional lower bound of the same variable. For example, if $\mathbb{E}_{q(z_j)}[f_k(z_j)]$ (where $\mathbf{y}_{A_k}$ is dropped for clarity) is intractable, a function $g(z_j, \xi)$ replaces f_k and is a point-wise lower bound, $f_k(z_j) \geq g(z_j, \xi)$ for all z_j (Jaakkola and Jordan (2000) and Marlin et al. (2011)). The function g usually takes an auxiliary variable ξ, which determines how tightly g approximates f_k and is tuned along with other parameters during inference. Although inference can now proceed, a limitation of introducing bounds is that the true variational objective function is no longer being optimized, which may lead to a significantly worse posterior approximation. An alternative to a lower bound approximation when the expectation $\mathbb{E}_{q(z_j)}[f_k(z_j)]$ is intractable is an unbiased stochastic approximation of $\nabla_{\lambda_j}\mathcal{L}(\boldsymbol{\lambda})$ allowing for an optimization of (4.8.2). This leads to two main ideas to construe the gradient of the ELBO with respect to q, avoiding model-specific analysis.

4.8.1 Black box variational inference

Through incorporating the score function, an unbiased stochastic approximation of the gradient of the intractable joint log likelihood term can be performed. The estimator, known as the likelihood-ratio estimator, is popular as does not impose any restriction on $f_k(z_j)$ or the approximating density $q(z_j)$.

To simplify notation the indices are dropped; f is the intractable function of z and z has a

variational distribution q taking parameters $\boldsymbol{\lambda}$. The gradient of the expectation

$$\nabla_\lambda \mathbb{E}_q[f(z)] = \int \nabla_\lambda q(z|\lambda) f(z) dz \tag{4.8.3}$$

can not be approximated via Monte Carlo as the gradient of a density is not a density function. By using the identity $\nabla_\lambda q(z|\boldsymbol{\lambda}) = q(z|\boldsymbol{\lambda}) \nabla_\lambda \log q(z|\boldsymbol{\lambda})$, we can stochastically approximate this expectation using Monte Carlo integration

$$\nabla_\lambda \mathbb{E}_{q(z)}[f(z)] \approx \frac{1}{S} \sum_{s=1}^{S} f(z^{(s)}) \nabla_\lambda \log q(z^{(s)}|\boldsymbol{\lambda}), \tag{4.8.4}$$

where $z^{(s)} \sim q(z|\boldsymbol{\lambda})$ for $s = 1, ..., S$. As the variational update comprises the expectation over the likelihood (described in Section 4.5 or derived in Appendix 4.9.3), the gradient of the ELBO can be written as an expectation over the variational model $q(z|\boldsymbol{\lambda})$ with the addition of the score function $\nabla_\lambda \log q(z|\boldsymbol{\lambda})$ (Paisley et al., 2012)

$$\nabla_\lambda \mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(z)} \left[\nabla_\lambda \log q(z|\boldsymbol{\lambda}) (f(z) - \log q(z|\boldsymbol{\lambda})) \right], \tag{4.8.5}$$

with Monte Carlo integration used to obtain noisy estimates of the ELBO. The basic procedure is to sample from $q(z|\lambda)$, evaluate the score function ($\nabla_\lambda \log q(z|\lambda)$ and ELBO. A Monte Carlo estimate of the gradient is then

$$\nabla_\lambda \mathcal{L}(\boldsymbol{\lambda}) \approx \frac{1}{S} \sum_{s=1}^{S} \left[(f(z^{(s)}) - \log q(z^{(s)}|\boldsymbol{\lambda})) \nabla_\lambda \log q(z^{(s)}|\boldsymbol{\lambda}) \right]. \tag{4.8.6}$$

Black box variational inference (Ranganath et al., 2014) incorporates the stochastic optimisation into a general algorithm (Algorithm 3), avoiding the work required to derive the variational posteriors and ELBO. This method yields a Monte Carlo estimator of the gradient of the ELBO which facilitates stochastic updates for each parameter. The only requirements are the log variational distribution and the the log of the joint probability of the data and the latent variables must be differentiable with respect to the variational parameters.

Algorithm 3: Black box variational inference

Input : A model $p(\mathbf{y}, \mathbf{z})$, a data set $\mathbf{y}$, a mean field variational family q.

Intialize: Variational parameters $\boldsymbol{\lambda}$ randomly, step size schedule ρ_j

while *the ELBO has not converged* **do**

  for $s = 1$ *to* S **do**

    $\mathbf{z}^{(s)} \sim q(\mathbf{z}|\boldsymbol{\lambda})$

  end

  Compute the noisy stochastic gradient

  $\tilde{g}_j = \frac{1}{S} \sum_s (\log p(\mathbf{y}, \mathbf{z}^{(s)}) - \log q(\mathbf{z}^{(s)}|\boldsymbol{\lambda}_j)) \nabla_\lambda \log q(\mathbf{z}^{(s)}|\boldsymbol{\lambda}_j)$

  Update the variational parameters

  $\boldsymbol{\lambda}_{j+1} = \boldsymbol{\lambda}_j + \rho_j \tilde{g}_j$

end

return $q(\mathbf{z})$

Reducing the variance of the gradient estimator is essential to the fast convergence of the algorithm. Rao-Blackwellization (Casella and Robert, 1996) exploits the factorisation of the variational distribution. Control variates (Ross, 2006) use the log probability of the variational distribution. The idea of adaptive learning rates and mini batches described in Section 4.7 are also applicable because of the stochastic form of the update. This approach works for both discrete and continuous models.

4.8.2 Reparameterisation gradient

If the model has differentiable latent variables, then it is generally advantageous to leverage gradient information from the model in order to better traverse the optimization space. One approach to this is the reparameterisation gradient, referred to as *stochastic backpropogation* (Rezende et al., 2014) or *stochastic gradient variational Bayes* (Kingma and Welling, 2014). This involves reparameterising the latent variable in terms of a base distribution and a differentiable transformation (such as a location scale transformation) in order to simplify the expectation of the gradient (Ap-

pendix 4.9.7). For example if $p(z)$ is a multivariate Gaussian $z \sim N(\boldsymbol{\mu}, \Sigma)$, then the location-scale transformation using a standard multivariate normal is

$$z \sim N(z|\boldsymbol{\mu}, \Sigma) \Leftrightarrow z = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, I) \tag{4.8.7}$$

where $\Sigma = \boldsymbol{L}\boldsymbol{L}^T$. In general this can be written as

$$\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}) \quad z = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{\epsilon} = h(\boldsymbol{\epsilon}; \boldsymbol{\theta}). \tag{4.8.8}$$

The random variable $\boldsymbol{\epsilon}$ is independent of the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$. The deterministic function $h(\boldsymbol{\epsilon}; \boldsymbol{\theta})$ encapsulates the parameters instead, and following the process is equivalent to directly drawing z from the original distribution. The estimator can be adapted to many other continuous distributions. The equivalent expectations are

$$\mathbb{E}_{p(z)}[f(z)] \Leftrightarrow \mathbb{E}_{p(\epsilon)}[f(h(\boldsymbol{\epsilon}; \boldsymbol{\theta}))] \tag{4.8.9}$$

and after applying the chain rule, the derivative is thus

$$\nabla_\theta \mathbb{E}_{p(z)}[f(z)] = \mathbb{E}_{p(\epsilon)}[\nabla_z f(z) \nabla_\theta h(\boldsymbol{\epsilon}; \boldsymbol{\theta})]. \tag{4.8.10}$$

Returning to the variational parameters ($\boldsymbol{\lambda}$) in the notation defined in the beginning of Section 4.8, with the sampling path $g(\epsilon; \boldsymbol{\lambda})$ and a base distribution $p(\epsilon)$. If $f(z)$ and $\log q(z)$ are differentiable with respect to z then the reparameterisation gradient of the ELBO can be expressed as

$$\nabla_\lambda \mathcal{L}(\lambda) = \mathbb{E}_{p(\epsilon)} \left[\underbrace{\nabla_z \left(f(z) - \log q(z|\boldsymbol{\lambda}) \right) |_{z=h(\epsilon;\lambda)}}_{\text{gradient of instantaneous ELBO}} \times \underbrace{\nabla_\lambda h(\epsilon; \boldsymbol{\lambda})}_{\text{gradient of transformation}} \right]. \tag{4.8.11}$$

Unlike the score gradient approach, we take the derivative of the ELBO function in (4.8.11) which must be differentiable.

The reparameterisation trick cannot be applied to discrete variables, since any reparameterisa-

tion includes discontinuous operations for which the gradient cannot be estimated. An alternative approach proposed by Tokui and Sato (2016) avoids the discontinuity by marginalizing out the variable of interest. The gradient $\nabla_z f(z)$ depends on the model, but can be computed using automatic differentiation tools (Baydian et al., 2018). This has led to powerful software packages for easy-to-use variational inference using automatic differentiation, where a small amount of code replaces a large amount of mathematical derivation (Duvenaud and Adams, 2016). An important advantage of stochastic backpropogation is that for models with continuous latent variables, it has the lowest variance among competing estimators. Rezende and Mohamed (2015) combine stochastic backpropogation with normalizing flows of different lengths to obtain increasingly complex posterior approximations.

Titsias and Lázaro-Gredilla (2014) propose an alternative stochastic optimization algorithm for correlated non-conjugate inference in continuous parameter space. Through a change of variable, the integration within the KL divergence between the target and transformed approximation is performed by Monte Carlo simulation. The approach is a more general version of the variational Gaussian approximation of Challis and Barber (2013) which does not rely on an analytically tractable integral for $f(z)$. By adopting the stochastic variational updates described in Section 4.6 the approach, which now also incorporates stochasticity by sampling from the variational distribution, is referred to as *doubly stochastic variational inference*.

4.9 Appendix

4.9.1 The EM algorithm

An understanding of VI can be developed by comparing with the frequentist EM algorithm. The EM algorithm is an iterative algorithm, introduced in (Dempster et al., 1977), and is designed to compute the (ML) estimate when there is missing data. It consists of a series of iteration where the parameter values get repeatedly updated until a convergence criteria is met. The algorithm converges to a local maximum of the likelihood function, thus if the function is unimodal the EM

algorithm will converge to the ML estimate.

The EM algorithm is primarily used for maximising the likelihood when the task becomes easier given more information associated with existing data. This situation is called an incomplete data problem because we do not have this extra information. Instead we augment the likelihood $L(\theta; y)$ or equivalently log-likelihood $\ell(\theta; \mathbf{y})$, where $\mathbf{y}$ is the data and θ is the parameter(s), with a latent variable z. The expectation of the likelihood is then taken, conditional on the observed data and the current value of the parameter $\mathbb{E}[\ell(\mathbf{y}, z|\theta)|\mathbf{y}, \theta^{(t)}]$ with respect to the latent variable. We then maximise $\mathbb{E}[\ell(\mathbf{y}, z|\theta)|\mathbf{y}, \theta^{(t)}]$ with respect to θ using the value of the latent variable we have obtained from the expectation to get $\theta^{(t+1)}$.

The E- and M- steps can be formally specified as;

- **E-step**: Calculation of $Q(\theta|\theta^{(t)})$ as a function of θ; $Q(\theta|\theta^{(t)}) = \mathbb{E}[\ell(\theta; \mathbf{y}, z)|\mathbf{y}, \theta^{(t)}]$ with respect to $p(z|\mathbf{y}, \theta^{(t)})$ distribution.
- **M-step**: Maximization of $Q(\theta|\theta^{(t)})$ with respect to θ to get $\theta^{(t+1)}$; $\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)})$.

How does the algorithm work?

The EM algorithm iterates over $Q(\theta|\theta^{(t)})$ which contains the joint probability of the data and the augmented variable as increasing $Q(\theta|\theta^{(t)})$ increases the marginal log-likelihood $\ell(\theta; y)$.

The marginal likelihood can be expressed in terms of the augmented variable z as

$$p(\mathbf{y}|\theta) = \frac{p(\mathbf{y}, z|\theta)}{p(z|\theta, \mathbf{y})}$$

and by taking logs as

$$\log(p(\mathbf{y}|\theta)) = \log(p(\mathbf{y}, z|\theta)) - \log(p(z|\theta, \mathbf{y})). \tag{4.9.1}$$

Taking the expectation with respect to the posterior latent variable z, conditional on the observed

data and the current parameter estimates $\theta^{(t)}$, returns an expression which is function of θ

$$\begin{aligned}\mathbb{E}[\log(p(\mathbf{y}|\theta))|\mathbf{y},\theta^{(t)}] =& \mathbb{E}[\log(p(\mathbf{y},z|\theta))|\mathbf{y},\theta^{(t)}] - \mathbb{E}[\log(p(z|\theta,\mathbf{y}))|\mathbf{y},\theta^{(t)}] \\ \log(p(\mathbf{y}|\theta)) =& \int \log(p(\mathbf{y},z|\theta))p(z|\mathbf{y},\theta^{(t)})dz - \int \log(p(z|\theta,\mathbf{y}))p(z|\mathbf{y},\theta^{(t)})dz \\ =& Q(\theta|\theta^{(t)}) + H(\theta|\theta^{(t)}).\end{aligned}$$

This equation holds for any value of θ including $\theta = \theta^{(t)}$ so

$$\log\big(p(\mathbf{y}|\theta^{(t)})\big) = Q(\theta^{(t)}|\theta^{(t)}) + H(\theta^{(t)}|\theta^{(t)})$$

Subtracting the two equations

$$\log(p(\mathbf{y}|\theta)) - \log\big(p(\mathbf{y}|\theta^{(t)})\big) = Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) + H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})$$

The term $H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})$ can be ignored if it is greater than or equal to 0 and θ is chosen so $Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})$ is non-decreasing. Thus, first we prove $H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}) \geq 0$.

$$\begin{aligned}H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}) =& \bigg[- \int \log(p(z|\theta,\mathbf{y})) + \int \log\big(p(z|\theta^{(t)},\mathbf{y})\big)\bigg] p(z|\mathbf{y},\theta^{(t)})dz \\ =& - \bigg[\int \log(p(z|\theta,\mathbf{y})) - \int \log\big(p(z|\theta^{(t)},\mathbf{y})\big)\bigg] p(z|\mathbf{y},\theta^{(t)})dz \\ =& - \mathbb{E}\bigg[\log\bigg(\frac{p(z|\theta,\mathbf{y})}{p(z|\mathbf{y},\theta^{(t)})}\bigg)|\mathbf{y},\theta^{(t)}\bigg]\end{aligned}$$

Using Jensens inequality for a concave function $\mathbb{E}[f(x)] \leq f(\mathbb{E}(x))$ and remembering the minus sign which gives $-\mathbb{E}[f(x)] \geq -f[\mathbb{E}(x)]$

$$\begin{aligned}-\mathbb{E}\bigg[\log\bigg(\frac{p(z|\theta,\mathbf{y})}{P(z|\mathbf{y},\theta^{(t)})}|\mathbf{y},\theta^{(t)}\bigg)\bigg] &\geq -\log\bigg(\mathbb{E}\bigg[\frac{p(z|\theta,\mathbf{y})}{p(z|\mathbf{y},\theta^{(t)})}\bigg]|\mathbf{y},\theta^{(t)}\bigg) \\ &\geq \log\bigg(\int p(z|\theta,\mathbf{y})dz\bigg) \\ &\geq 0\end{aligned}$$

Therefore $H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}) \geq 0$ as desired $\square$

We are able to optimise $Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})$ if our choice of θ also increases the the incomplete log likelihood. We maximise $Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})$ as a function of its first argument to get $\theta^{(t+1)}$. The $Q(\theta^{(t)}|\theta^{(t)})$ term is ultimately lossed in the E- and M-steps as we differentiate with respect to θ.

$$
\begin{aligned}
Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) &\geq 0 \\
\int \log\big(p(\mathbf{y}, z|\theta^{(t+1)})\big) p(\mathbf{y}|z, \theta^{(t)}) dz - \int \log\big(p(\mathbf{y}, z|\theta^{(t)})\big) p(z|\mathbf{y}, \theta^{(t)}) dz &\geq 0 \\
\int \log\left(\frac{p(z, \mathbf{y}|\theta^{(t+1)})}{p(z, \mathbf{y}|\theta^{(t)})}\right) p(z|\mathbf{y}, \theta^{(t)}) dz &\geq 0
\end{aligned}
$$

By Bayes theorem

$$
\begin{aligned}
\int \log\left(\frac{p(z|\mathbf{y}, \theta^{(t+1)}) p(\mathbf{y}|\theta^{(t+1)})}{p(z|\mathbf{y}, \theta^{(t)}) p(\mathbf{y}|\theta^{(t)})}\right) p(z|\mathbf{y}, \theta^{(t)}) dz &\geq 0 \\
\int \log\left(\frac{p(\mathbf{y}|\theta^{(t+1)})}{p(\mathbf{y}|\theta^{(t)})}\right) p(z|\mathbf{y}, \theta^{(t)}) dz + \int \log\left(\frac{p(z|\mathbf{y}, \theta^{(t+1)})}{p(z|\mathbf{y}, \theta^{(t)})}\right) p(z|\mathbf{y}, \theta^{(t)}) dz &\geq 0.
\end{aligned}
$$

Concentrating on the second part of the equation, using $\log(x) \leq x - 1$ we have

$$
\begin{aligned}
\int \log\left(\frac{p(z|\mathbf{y}, \theta^{(t+1)})}{p(z|\mathbf{y}, \theta^{(t)})}\right) p(z|\mathbf{y}, \theta^{(t)}) dz &\leq \int \left(\frac{p(z|\mathbf{y}, \theta^{(t+1)})}{p(z|\mathbf{y}, \theta^{(t)})} - 1\right) p(z|\mathbf{y}, \theta^{(t)}) dz \\
&= \int p(z|\mathbf{y}, \theta^{(t+1)}) - p(z|\mathbf{y}, \theta^{(t)}) dz \\
&= 0.
\end{aligned}
$$

The first integral is not positive. Therefore the second integral must be non-negative

$$
\begin{aligned}
\int \log\left(\frac{p(\mathbf{y}|\theta^{(t+1)})}{p(\mathbf{y}|\theta^{(t)})}\right) p(z|\mathbf{y}, \theta^{(t)}) dz &\geq 0 \\
\log\left(\frac{p(\mathbf{y}|\theta^{(t+1)})}{p(\mathbf{y}|\theta^{(t)})}\right) \int p(z|\mathbf{y}, \theta^{(t)}) dz &\geq 0 \\
\log\big(p(\mathbf{y}|\theta^{(t+1)})\big) - \log\big(p(\mathbf{y}|\theta^{(t)})\big) &\geq 0 \quad \square
\end{aligned}
$$

Since $p(\mathbf{y}|\theta) = L(\theta;\mathbf{y})$ then $\ell(\theta^{(t+1)};\mathbf{y}) \geq \ell(\theta^{(t)};\mathbf{y})$, the log likelihood increases from one iteration to the next. Choosing θ to maximise $Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})$ leads to a "non-decrease" in the marginal likelihood, regardless of the second term $H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})$.

4.9.2 Derivations for the EM Gaussian mixture model comparison

The expected value of the completed log likelihood with respect to the posterior distribution $p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{(t)})$ is given by

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \mathbb{E}[\log L(\boldsymbol{\theta};\mathbf{y},\mathbf{z})|\mathbf{y},\boldsymbol{\theta}^{(t)}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{K} \mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)},\mathbf{y}] \log\left(\pi_j^{(t)}\right) + \sum_{i=1}^{n}\sum_{j=1}^{K} \mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)},\mathbf{y}] \log N(y_i;\mu_j^{(t)},\sigma_j^{2(t)}). \end{aligned} \tag{4.9.2}$$

The M-step requires taking the derivative of the expected complete log-likelihood with respect to the parameters $\boldsymbol{\theta}$. The ML estimate of the $\boldsymbol{\pi}$ parameters is determined by using the Lagrange multiplier for the constraint $\sum_j \pi_j = 1$. The Lagrangian is

$$L(\boldsymbol{\pi},\lambda) = f(\boldsymbol{\pi},\mathbf{y}) - \lambda(g(\boldsymbol{\pi},\mathbf{y}) - c) \tag{4.9.3}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{K} \mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)},\mathbf{y}] \log \pi_j - \lambda\left(\sum_j \pi_j - 1\right), \tag{4.9.4}$$

after derivatives with respect to $\boldsymbol{\pi}$ and λ,

$$\sum_{i=1}^{n} \frac{\mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)},\mathbf{y}]}{\pi_j} - \lambda = 0 \tag{4.9.5}$$

$$\sum_{j=1}^{K} \pi_j - 1 = 0 \tag{4.9.6}$$

which gives $\lambda = n$ and

$$\pi_j^{(t)} = \sum_{i=1}^{n} \frac{\mathbb{E}[z_{ij}|\boldsymbol{\theta}^{(t)},\mathbf{y}]}{n}. \tag{4.9.7}$$

The Bayesian prior conjugate specification of (4.4.13) and mean field family (4.4.14) leads to the

following complete conditional

$$\begin{aligned}\log q(\mu_j) &\propto \mathbb{E}_{q(-\mu_j)}\left[\sum_{i=1}^{n} z_{ij}\left(-\frac{1}{2\sigma_j^2}(y_i-\mu_j)^2\right)-\frac{1}{\tau^2}\mu_j^2\right] \\ &\propto \mathbb{E}_{q(-\mu_j)}\left[-\frac{1}{2}\left(\mu_j^2\left(\frac{\sum_{i=1}^n z_{ij}}{\sigma_j^2}+\frac{1}{\tau^2}\right)-2\mu_j\frac{\sum_{i=1}^n z_{ij}}{\sigma_j^2}\right)\right]\end{aligned}$$

Exponentiating and completing the square gives $q(\mu_j) = N(\mu_j^*, \tau_j^*)$ with updates

$$\mu_j^* = \frac{\sum_{i=1}^n (z_{ij})^{(1)} y_i}{\sum_{i=1}^n (z_{ij})^{(1)} + (\sigma_j^2)^{(1)}/(\tau^2)^{(1)}} \qquad \tau_j^{2*} = \left(\frac{\sum_{i=1}^n (z_{ij})^{(1)}}{(\sigma_j^2)^{(1)}} + \frac{1}{(\tau^2)^{(1)}}\right)^{-1}. \tag{4.9.8}$$

where $(.)^{(1)}$ denotes a q expectation, $(\mu_j)^{(1)} = \mu_j^*$ and $(\mu_j)^{(2)} = \mu_j^{*2} + \tau_j^{2*}$.

$$\begin{aligned}\log q(\sigma_j^2) &\propto \mathbb{E}_{q(-\sigma_j^2)}\left[\sum_{i=1}^{n} z_{ij}\left(-\frac{1}{2}\log\sigma_j^2-\frac{(y_i-\mu_j)^2}{2\sigma_j^2}\right)+(-a-1)\log\sigma_j^2-\frac{b}{\sigma_j^2}\right] \\ &\propto \log\sigma_j^2\left(-\sum_{i=1}^{n}\frac{(z_{ij})}{2}-a-1\right)-\frac{1}{\sigma_j^2}\left(\sum_{i=1}^{n}\frac{(z_{ij})^{(1)}}{2}(y_i-2y_i(\mu_j)^{(1)}+(\mu_j)^{(2)})+b\right)\end{aligned}$$

which is the log kernel of an Inverse Gamma density. Thus $q(\sigma_j^2) = IG(a_j^*, b_j^*)$ with updates

$$a_j^* = \sum_{i=1}^{n}\frac{(z_{ij})^{(1)}}{2}+a, \quad b_j^* = b+\sum_{i=1}^{n}\frac{(z_{ij})^{(1)}(y_i-2y_i(\mu_j)^{(1)}+(\mu_j)^{(2)})}{2}. \tag{4.9.9}$$

with

$$(\sigma_j^{-2})^{(1)} = \frac{a_j^*}{b_j^*}. \tag{4.9.10}$$

The probabilities of belonging to each of the mixtures $\boldsymbol{\pi}$,

$$\begin{aligned}\log q(\boldsymbol{\pi}) &\propto \mathbb{E}_{q(-\boldsymbol{\pi})}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\log(\pi_j)+\sum_{j=1}^{k}(\alpha_j-1)\log(\pi_j)\right] \\ &\propto \sum_{j=1}^{k}\log(\pi_j)\left(\alpha_j+\sum_{i=1}^{n}(z_{ij})^{(1)}-1\right)\end{aligned}$$

is proportional to the log Dirichlet distribution. The complete conditional $q(\boldsymbol{\pi}) = Dir(\alpha_1^*, ..., \alpha_k^*)$

where

$$\alpha_j^* = \alpha_j + \sum_{i=1}^{n} (z_{ij})^{(1)}. \tag{4.9.11}$$

with

$$(\pi_j)^{(1)} = \frac{\alpha_j + \sum_{i=1}^{n} (z_{ij})^{(1)}}{\sum_{m=1}^{k} \alpha_m + n} \tag{4.9.12}$$

as $\sum_i \sum_j (z_{ij})^{(1)} = n$.

For the local update $\mathbf{z}_i$

$$\begin{aligned}
\log q(\mathbf{z}_i) &\propto \mathbb{E}_{q(-\mathbf{z}_i)} \left[\sum_{j=1}^{k} \left\{ z_{ij} \Big(-\frac{1}{2} \log\big(2\pi\sigma_j^2\big) - \frac{1}{2\sigma_j^2} (y_i - \mu_j)^2 \Big) + z_{ij} \log(\pi_j) + z_{ij} \log(\eta_j) \right\} \right] \\
&\propto \sum_{j=1}^{k} z_{ij} \Big(-\frac{1}{2} \log\big(2\pi(\sigma_j^2)^{(1)}\big) - \frac{1}{2(\sigma_j^2)^{(1)}} (y_i^2 - 2y_i(\mu_j)^{(1)} + (\mu_j)^{(2)}) + \log(\pi_j) + \log(\eta_j) \Big)
\end{aligned}$$

thus $q(\mathbf{z}_i) = \text{Multinomial}(1, \eta_1^*, ..., \eta_k^*)$ with normalised probabilities

$$\eta_j^* = \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(\sigma_j^2)^{(1)}}} \exp\Big(-\frac{1}{2(\sigma_j^2)^{(1)}} (y_i - 2y_i(\mu_j)^{(1)} + (\mu_j)^{(2)}) \Big) (\pi_j)^{(1)} \eta_j}{\prod_{i=1}^{n} \sum_{j=1}^{K} \frac{1}{\sqrt{2\pi(\sigma_j^2)^{(1)}}} \exp\Big(-\frac{1}{2(\sigma_j^2)^{(1)}} (y_i - 2y_i(\mu_j)^{(1)} + (\mu_j)^{(2)}) \Big) (\pi_j)^{(1)} \eta_j} \tag{4.9.13}$$

with $(z_{ij})^{(1)} = \eta_j^*$ □

4.9.3 Bayesian updates in exponential family form

Bayesian posterior updating can be performed generically in the exponential family form and the updated natural parameters in the exponential family form can then be mapped to the posterior parameters in the standard form. If we define a prior parameterisation as

$$\begin{aligned}
\boldsymbol{\eta} &\sim F(\boldsymbol{\eta}|\boldsymbol{\lambda}) \\
\mathbf{x}_i &\sim G(\mathbf{x}_i|\boldsymbol{\eta}) \text{ for } i \in \{1, ..., n\}.
\end{aligned}$$

where $\boldsymbol{\lambda}$ are the prior hyperparmeters in their natural form. The posterior distribution of $\boldsymbol{\eta}$ given the data $\mathbf{x}_{1:n}$ is

$$p(\boldsymbol{\eta}|\mathbf{x}_{1:n}, \boldsymbol{\lambda}) \propto F(\boldsymbol{\eta}|\boldsymbol{\lambda}) \prod_{i=1}^{n} G(\mathbf{x}_i|\boldsymbol{\eta}). \tag{4.9.14}$$

If this distribution is in the same family as F then F and G are a conjugate pair. The conjugate prior, particularly in SVI where this parameterisation simplifies the algebra, can be expressed relative to the likelihood as

$$p(\mathbf{x}_i|\boldsymbol{\eta}) = h_l(\mathbf{x}_i) \exp\{\boldsymbol{\eta}^T t(\mathbf{x}_i) - a_l(\boldsymbol{\eta})\} \tag{4.9.15}$$

$$p(\boldsymbol{\eta}|\boldsymbol{\lambda}_1, \lambda_2) = h_c(\boldsymbol{\eta}) \exp\{\boldsymbol{\lambda}_1^T \boldsymbol{\eta} + \lambda_2(-a_l(\boldsymbol{\eta})) - a_c(\boldsymbol{\lambda})\} \tag{4.9.16}$$

$$= h_c(\boldsymbol{\eta}) \exp\{\boldsymbol{\lambda}^T [\boldsymbol{\eta}, (-a_l(\boldsymbol{\eta}))] - a_c(\boldsymbol{\lambda})\} \tag{4.9.17}$$

$$= h_c(\boldsymbol{\eta}) \exp\{\boldsymbol{\lambda}^T t(\boldsymbol{\eta}) - a_c(\boldsymbol{\lambda})\}$$

where $a_l(\boldsymbol{\eta})$ is the same function as appears in the respective likelihood Equation (4.9.15) . The natural parameter $\boldsymbol{\lambda} = \langle \boldsymbol{\lambda}_1, \lambda_2 \rangle$ has dimension $\dim(\boldsymbol{\eta}) + 1$ (λ_2 is scalar) and the sufficient statistic of $p(\boldsymbol{\eta}|\boldsymbol{\lambda}_1, \lambda_2)$ is $\langle \boldsymbol{\eta}, -a(\boldsymbol{\eta}) \rangle$.

The posterior is

$$\begin{aligned}
p(\boldsymbol{\eta}|\boldsymbol{x}_{1:n}, \boldsymbol{\lambda}) &\propto p(\boldsymbol{\eta}|\boldsymbol{\lambda}) \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\lambda}) \\
&= h(\boldsymbol{\eta}) \exp\{\boldsymbol{\lambda}_1^T \boldsymbol{\eta} + \lambda_2(-a_l(\boldsymbol{\eta})) - a_c(\boldsymbol{\lambda})\} \cdot \prod_{i=1}^{N} h(\mathbf{x}_i) \exp\{\boldsymbol{\eta}^T t(\mathbf{x}_i) - a_l(\boldsymbol{\eta})\} \\
&\propto h(\boldsymbol{\eta}) \exp\{(\boldsymbol{\lambda}_1 + \sum_{i=1}^{N} t(\mathbf{x}_i))^T \boldsymbol{\eta} + (\lambda_2 + n)(-a_l(\boldsymbol{\eta}))\} \\
&\propto h(\boldsymbol{\eta}) \exp\{\hat{\boldsymbol{\lambda}}^T [\boldsymbol{\eta}, -a_l(\boldsymbol{\eta})]\}
\end{aligned}$$

This is the same exponential family as the prior with parameters of $\hat{\boldsymbol{\lambda}}$

$$\hat{\boldsymbol{\lambda}}_1 = \boldsymbol{\lambda}_1 + \sum_{i=1}^{n} t(\mathbf{x}_i) \tag{4.9.18}$$

$$\hat{\lambda}_2 = \lambda_2 + n. \tag{4.9.19}$$

This is a reparameterisation of the common approach outlined in Bernado and Smith (1994) who define conjugacy priors relative to the likelihood as

$$p(\mathbf{x}_i|\boldsymbol{\eta}) = h_l(\mathbf{x}_i)g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^T t(\mathbf{x}_i)\} \tag{4.9.20}$$

$$p(\boldsymbol{\eta}|\boldsymbol{\lambda}_1, \lambda_2) = K(\boldsymbol{\lambda})^{-1} g(\boldsymbol{\eta})^{\lambda_2} \exp\{\boldsymbol{\lambda}_1^T t(\boldsymbol{\eta})\} \tag{4.9.21}$$

where

$$K(\boldsymbol{\lambda}) = \int g(\boldsymbol{\eta})^{\lambda_2} \exp\{\boldsymbol{\lambda}_1^T t(\boldsymbol{\eta})\} d\boldsymbol{\eta}. \tag{4.9.22}$$

as $g(\boldsymbol{\eta}) = \exp(-a_l(\boldsymbol{\eta}))$.

A simple Gaussian example with unit variance can be expressed as

$$p(x|\mu) = \frac{\exp\{-x^2/2\}}{\sqrt{2\pi}} \exp\{\mu x - \mu^2/2\} \tag{4.9.23}$$

The conjugate prior is $h_c(\eta)\exp\{\lambda_1^T\eta + \lambda_2(-a_l(\eta)) - a_c(\lambda)\}$. We could set $\lambda_1^* = \lambda_1$ and $\lambda_2^* = -\lambda_2/2$ so the sufficient statistics are (η, η^2). The posterior parameters are

$$\begin{aligned} \hat{\lambda}_1 &= \lambda_1 + \sum_{i=1}^{n} x_i \\ \hat{\lambda}_2 &= \lambda_2 + n \\ \hat{\lambda}_2^* &= -\frac{(\lambda_2 + n)}{2} \end{aligned}$$

If we choose a prior Gaussian with mean and variance (μ_0, σ_0^2) then rearranging into the expo-

nential family form for the prior we have

$$\begin{aligned}\lambda_1 &= \mu_0/\sigma_0^2 \\ \lambda_2 &= -1/2\sigma_0^2 \\ \lambda_2^* &= 1/\sigma_0^2.\end{aligned}$$

Here the posterior hyperparameters are a function of the natural parameters of the posterior, just as in the prior. This feature is used in Section 4.5 to show that if the prior is chosen to be in conjugate pair the update for the variational parameters is in the same exponential family form.

4.9.4 Complete conditional and the exponential family form

$$\begin{aligned}q(z_j) &\propto \exp\{\mathbb{E}_{q(\mathbf{z}_{-j})} \log p(z_j|\mathbf{z}_{-j}, \mathbf{y}))\} \\ &\propto \exp\{\log h(z_j) + \mathbb{E}_{q(\mathbf{z}_{-j})}[\eta_j(\mathbf{z}_{-j}, \mathbf{y})^T]t(z_j) - \mathbb{E}_{q(\mathbf{z}_{-j})}[a(\eta_j(\mathbf{z}_{-j}, \mathbf{x}))]\} \\ &\propto h(z_j)\exp\{\mathbb{E}_{q(\mathbf{z}_{-j})}[\eta_j(\mathbf{z}_{-j}, \mathbf{y})^T]t(z_j)\},\end{aligned}$$

where $t(z_j)$ is the sufficient statistic. If we let $\mathbf{v}_j$ denote the variational parameter for the jth variational factor, when we update each factor we set its parameter equal to the expected parameter of the complete conditional

$$\mathbf{v}_j = \mathbb{E}_{q(\mathbf{z}_{-j})}[\eta_j(\mathbf{z}_{-j}, \mathbf{y})], \tag{4.9.24}$$

where there is one sufficient statistic per variational factor.

4.9.5 Adaptive learning rate stochastic variational inference algorithm

The SVI algorithm to estimate the local ($\boldsymbol{\phi}$) and global ($\boldsymbol{\lambda}$) free parameters with an adaptive learning rate.

Algorithm 4: SVI for Conditionally Conjugate Models with Adaptive Learning rates

Input : A model $p(\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\beta})$, a data set $\mathbf{y}$.
Output : Global variational density $q_\lambda(\boldsymbol{\beta}|\lambda), \prod_i \prod_j q_\phi(\gamma_{ij}|\boldsymbol{\phi}_{ij})$
Intialize: Variational parameters $\boldsymbol{\lambda}_1$, $\boldsymbol{\phi}_1$, window size τ_1, moving averages $\bar{g}_0, \bar{h}_0$.
for $m = 1, ..., \infty$ **do**

Sample a data point $y_i^{(r)}$ randomly, $i \sim \text{Unif}(1, ..., n)$, from the data set. Optimize its associated local variational parameters:

$$\boldsymbol{\phi}_{ij}^{\lambda_m} = \mathbb{E}_{q(\beta|\lambda_m)}[\eta_l(y_i^{(r)}, \boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}_{i,-j}^{(r)})] \; \forall \; j$$

Compute the intermediate global parameter as though y_i had been replicated n times:

$$\hat{\boldsymbol{\lambda}}_m = \boldsymbol{\alpha} + n \left[\mathbb{E}_{q(\gamma_i|\phi_i)} \left(t(y_i^{(r)}, \boldsymbol{\gamma}_i^{(r)}) \right), 1 \right]^T.$$

Update the moving averages $\bar{g}_m$ and $\bar{h}_m$:

$$\bar{g}_m = (1 - \tau_m^{-1})\bar{g}_{m-1} + \tau_m^{-1}\hat{g}(\boldsymbol{\lambda}_m)$$
$$\bar{h}_m = (1 - \tau_m^{-1})\bar{h}_{m-1} + \tau_m^{-1}\hat{g}(\boldsymbol{\lambda}_m)^T\hat{g}(\boldsymbol{\lambda}_m)$$

Set the estimate step size:

$$\epsilon_m^* = \frac{\bar{g}_m^T \bar{g}_m}{\bar{h}_m}$$

Update the window size:

$$\tau_{m+1} = \tau_m(1 - \epsilon_m^*) + 1$$

Update the current estimate of the global parameters (which computes the natural gradient):

$$\boldsymbol{\lambda}_{m+1} = (1 - \epsilon_m^*)\boldsymbol{\lambda}_m + \epsilon_t^* \hat{\boldsymbol{\lambda}}_m$$

end
return $\boldsymbol{\lambda}$, $\boldsymbol{\phi}$

4.9.6 Adaptive learning rate derivations

The expectation of the intermediate global parameter $\hat{\boldsymbol{\lambda}}_m$ is

$$\begin{aligned}
\mathbb{E}_n[\hat{\boldsymbol{\lambda}}_m] &= \sum_{i=1}^{n} \left(\boldsymbol{\alpha} + n \left[\mathbb{E}_{q(\boldsymbol{\gamma}_i|\boldsymbol{\phi}_i)} \left(t(\boldsymbol{\gamma}_i, y_i) \right), 1 \right]^T \right) \; p(I = i) \\
&= \frac{1}{n} \left(n\boldsymbol{\alpha} + n \sum_{i=1}^{n} \left[\mathbb{E}_{q(\gamma_i|\phi_i)} \left(t(\boldsymbol{\gamma}_i, y_i) \right), 1 \right]^T \right) \\
&= \boldsymbol{\alpha} + \left[\sum_{i=1}^{n} \mathbb{E}_{q(\boldsymbol{\gamma}_i|\phi_i)} \left(t(\gamma_i, y_i) \right), 1 \right]^T \\
&= \boldsymbol{\lambda}_m^*. \qquad (4.9.25)
\end{aligned}$$

Minimise $\mathbb{E}_n[J(\epsilon_m)|\boldsymbol{\lambda}_m]$ with respect to ϵ_m.

$$J(\epsilon_m) = ((1-\epsilon_m)\boldsymbol{\lambda}_m + \epsilon_m\hat{\boldsymbol{\lambda}}_m - \boldsymbol{\lambda}_m^*)^T((1-\epsilon_m)\boldsymbol{\lambda}_m + \epsilon_m\hat{\boldsymbol{\lambda}}_m - \boldsymbol{\lambda}_m^*) \tag{4.9.26}$$

Expanding (4.9.26) and making proportional to ϵ_m

$$\begin{aligned}\mathbb{E}_n[J(\epsilon_m)|\boldsymbol{\lambda}_m] \propto \epsilon_m^2\mathbb{E}_n[(\hat{\boldsymbol{\lambda}}_m - \boldsymbol{\lambda}_m)^T(\hat{\boldsymbol{\lambda}}_m - \boldsymbol{\lambda}_m)|\boldsymbol{\lambda}_m] + 2\epsilon_m\mathbb{E}_n[(\hat{\boldsymbol{\lambda}}_m - \boldsymbol{\lambda}_m)^T\boldsymbol{\lambda}_m|\boldsymbol{\lambda}_m] + \\ - 2\epsilon_m\mathbb{E}_n[(\hat{\boldsymbol{\lambda}}_m - \boldsymbol{\lambda}_m)\boldsymbol{\lambda}_m^*|\boldsymbol{\lambda}_m]\end{aligned}$$

Using

$$\begin{aligned}\mathbb{E}_n[\hat{\boldsymbol{\lambda}}_m^T\hat{\boldsymbol{\lambda}}_m|\boldsymbol{\lambda}_m] &= tr(\mathbb{E}_n[\hat{\boldsymbol{\lambda}}_m\hat{\boldsymbol{\lambda}}_m^T|\boldsymbol{\lambda}_m]) \\ &= tr(\Sigma) + \boldsymbol{\lambda}_m^{*T}\boldsymbol{\lambda}_m^*\end{aligned}$$

and setting the derivative to 0

$$\begin{aligned}0 = \epsilon_m(\mathbb{E}_n[\hat{\boldsymbol{\lambda}}_m^T\hat{\boldsymbol{\lambda}}_m|\boldsymbol{\lambda}_m] - 2\mathbb{E}_n[\hat{\boldsymbol{\lambda}}_m]^T\boldsymbol{\lambda}_m + \boldsymbol{\lambda}_m^T\boldsymbol{\lambda}_m) + (\boldsymbol{\lambda}_m^*)^T\boldsymbol{\lambda}_m - \boldsymbol{\lambda}_m^T\boldsymbol{\lambda}_m + \\ - \boldsymbol{\lambda}_m^{*T}\boldsymbol{\lambda}_m^* + \boldsymbol{\lambda}_m\boldsymbol{\lambda}_m^*.\end{aligned}$$

Rearranging for ϵ_m^* gives

$$\epsilon_m^* = \frac{(\boldsymbol{\lambda}_m^* - \boldsymbol{\lambda}_m)^T(\boldsymbol{\lambda}_m^* - \boldsymbol{\lambda}_m)}{tr(\Sigma) + (\boldsymbol{\lambda}_m^* - \boldsymbol{\lambda}_m)^T(\boldsymbol{\lambda}^* - \boldsymbol{\lambda}_m)} \quad \square \tag{4.9.27}$$

4.9.7 The reparameterisation trick

In order to present an unbiased, differentiable and scalable estimator for the ELBO in variational inference Kingma and Welling (2014) use a reparameterisation trick. This approach avoids undifferentiable expectations and offers an alternative to using the score function to write the gradient of the ELBO as an expectation.

If we wish to take the gradient with respect to λ of the expectation

$$\mathbb{E}_{p(z)}[f_\lambda(z)],$$

where p is a density and $f_\lambda(z)$ is a function of the random variable z with parameter λ. Provided $f_\lambda(z)$ is differentiable, the gradient can be computed,

$$\begin{aligned}\nabla_\lambda \mathbb{E}_{p(z)}[f_\lambda(z)] &= \nabla_\lambda \left[\int_z p(z) f_\lambda(z) dz\right] \\ &= \int_z p(z) \left[\nabla_\lambda f_\lambda(z)\right] dz \\ &= \mathbb{E}_{p(z)} \left[\nabla_\lambda f_\lambda(z)\right].\end{aligned}$$

The gradient of the expectation is equal to the expectation of the gradient. If the density p is also parameterised by λ the product rule means,

$$\begin{aligned}\nabla_\lambda \mathbb{E}_{p(z)}[f_\lambda(z)] &= \nabla_\lambda \left[\int_z p_\lambda(z) f_\lambda(z)\right] dz \\ &= \int_z f_\lambda(z) \nabla_\lambda p_\lambda(z) dz + \mathbb{E}_{p_\lambda(z)} \left[\nabla_\lambda f_\lambda(z)\right]. \qquad (4.9.28)\end{aligned}$$

The first term of (4.9.28), containing the derivative of the density p, is not guaranteed to be an expectation. Monte Carlo methods require that we can sample from $p_\lambda(z)$, rather than differentiate the density. This is not a problem if we have an analytic solution to $\nabla_\lambda p_\lambda(z)$, but this is not true in general.

Kingma and Welling (2014) use a reparameterisation trick, for continuous densities, to remove this term. By introducing a random variable ϵ and making z a deterministic function given ϵ

$$\epsilon \sim p(\epsilon|\cdot) \qquad z = h_\lambda(\epsilon; \lambda),$$

the expectation with respect to z is equivalent to

$$\mathbb{E}_{p_\lambda(z)}[f(z)] = \mathbb{E}_{p(\epsilon)}[f(g_\lambda(\epsilon; \lambda))]. \qquad (4.9.29)$$

The law of the unconscious statistician states that the expectation of a function of a random variable can be computed without knowing its distribution, if we use a valid sampling path and a base distribution.

As the expectation is with respect to the distribution of ϵ because of the change of variable, the additional term from the product rule is avoided. The derivative of the expectation is thus

$$\begin{aligned}\nabla_\lambda \mathbb{E}_{p_\lambda(z)}\left[f(z)\right] &= \mathbb{E}_{p_{(\epsilon)}}\left[\nabla_\lambda f(g_\lambda(\epsilon;\lambda))\right] \\ &= \mathbb{E}_{p_{(\epsilon)}}\left[\nabla_\lambda g_\lambda(\epsilon;\lambda)\nabla_z f(z)\right] \quad \square \end{aligned} \tag{4.9.30}$$

CHAPTER 5

Transformations for Compositional Data

The properties of compositional data arise from the fact that they represent the relative magnitudes of the parts. A row vector $\mathbf{x} = [x_1, ..., x_d]$ is defined as a d-part composition, when all its components are strictly positive real numbers and thus only contain relative information. If a is a real positive number, $[x_1, x_2, ..., x_d]$ and $[ax_1, ax_2, ..., ax_d]$ convey the same information and are thus indistinguishable. The ratio of any two components of a subcomposition is the same as the ratio of the corresponding two components in the full composition. This set of vectors is called the simplex of d parts and is denoted $\mathcal{S}^d$. The geometry of this space has been established over the last three decades (Aitchison and Shen (1980), Aitchison and Bacon-Shone (1984), Egozcue and Pawlowsky-Glahn (2005)), and is often termed Aitchison geometry. Operations and metric characteristics have been developed so that the simplex space has the structure of a Euclidean space of dimension $d - 1$.

In order to exploit statistical approaches for unconstrained data a transformation is required, so that the composition is represented as a real vector. As the study of compositions is concerned

with the relative magnitudes, it seems sensible to work in terms of log ratios as it benefits from the simple relationship

$$\text{Var}\left(\log\left(\frac{x_i}{x_j}\right)\right) = \text{Var}\left(\log\left(\frac{x_j}{x_i}\right)\right). \tag{5.0.1}$$

There are three main log-ratio transformations available; Aitchison (1982) introduced the additive-log-ratio (alr), and centred-log-ratio (clr) transformations, and Egozcue et al. (2003) the isometric-log-ratio (ilr) transformation. Their form and properties are briefly reviewed in this Chapter. In the approaches developed in Chapters 6 and 7, only the ilr transformation is used. The Appendix includes an introduction to the geometry of the simplex, proposed by Aitchison (1986), which is analogous to working in the Euclidean space.

5.1 Additive-log-ratio

The alr transformation $\mathcal{S}^d \to \mathbb{R}^{d-1}$, is defined by

$$\mathbf{z} = \text{alr}(\mathbf{x}) = \left[\log\left(\frac{x_1}{x_d}\right) \ \ldots \ \log\left(\frac{x_{d-1}}{x_d}\right)\right] \tag{5.1.1}$$

where the ratios involve the division of each of the first $d-1$ components by the final component. The choice of denominator is arbitrary, and could be any specified component. The inverse transformation $\text{alr}^{-1} : \mathbb{R}^{d-1} \to \mathcal{S}^d$ is

$$\mathbf{x} = \text{alr}^{-1}(\mathbf{z}) = \mathcal{C}\left[\exp(z_1) \ \ldots \ \exp(z_{d-1}) \ 1\right], \tag{5.1.2}$$

where $\mathcal{C}$ denotes the closure operation, which divides each component of a vector by the sum of the components (scaling the vector to 1).

The *additive-log-ratio* term comes from the expression of its inverse (5.1.2). Each part of of the composition is

$$x_i = \text{alr}^{-1}(\mathbf{z})_i = \frac{\exp(z_i)}{\sum_{i=1}^{d-1}\exp(z_i) + 1}, \tag{5.1.3}$$

where the denominator is the effect of the closure. The term *additive* is from the denominator, which is the sum of exponentials, in contrast with other transformations where this feature is multiplicative or hybrid.

The alr transformation is not symmetric in the components, as the reference part x_d is in the denominator of the component logratios. Another reference part can be chosen, leading to different alr transformations. The alr transformation reduces perturbation and powering to ordinary operations in the $d-1$ dimensional real space:

$$\text{alr}((\alpha \otimes \mathbf{x}_1) \oplus (\beta \otimes \mathbf{x}_2)) = \alpha \cdot alr(\mathbf{x}_1) + \beta \cdot alr(\mathbf{x}_2) \tag{5.1.4}$$

for any compositions $\mathbf{x}_1, \mathbf{x}_2$ and any real constants α and β. However it has the inconvenience of not being invariant under permutation of components and fails to preserve distances, so dot products and norms in the Euclidean space are not the same in the simplex. It is not an isometric transformation and the alr co-ordinates are an oblique basis of the simplex.

The alr transformation has proved particularly useful for a wide variety of regression problems with compositional covariates. The linear log-contrast model Aitchison and Bacon-Shone (1984) with second-order terms, involves rearranging the alr transformed covariates so that the model is symmetric and takes the form

$$\mathbf{y} = \alpha + \sum_{j=1}^{d} \log(\boldsymbol{x}_j)\beta_j + \sum_{j} \sum_{k>j} (\log(\boldsymbol{x}_k) - \log(\boldsymbol{x}_j))\beta_{jk} + \boldsymbol{\epsilon} \qquad (\beta_1 + ... + \beta_d = 0), \tag{5.1.5}$$

subject to the sum to zero constraint of the elements of $\boldsymbol{\beta}$. Here $\mathbf{y}$ is vector of continuous responses, $\boldsymbol{\epsilon}$ is a vector of error terms and $\boldsymbol{x}_j$ is a column in the design matrix, where each row is a compositional sample.

The model (5.1.5) is well suited to understanding the effects of a subcomposition (subvector such as $\mathcal{C}(x_{c+1}, ..., x_d)$) on the response. If $\beta_j = 0$ and $\beta_{jk} = 0$ for $j = 1, .., c$ and $k > j$, the expected response depends on the composition only through the subcomposition. This motivates our choice of the model for the research articles in Chapter 6 and Chapter 7.

In the frequentist setting, the estimation of $\boldsymbol{\beta}$ in (5.1.5) is obtained via the method of Lagrange multipliers. The linear log-contrast model has been generalized to a high-dimensional setting via regularisation. Lin et al. (2014) introduced the sparse linear log-contrast model with variable selection via ℓ^1, this has been extended to multiple linear constraints for sub-compositional coherence across predefined groups of predictors (Shi et al., 2016). A general approach to convex optimisation, where the model has been extended to the high-dimensional setting via regularization has recently been proposed by Combettes and Müller (2021).

5.2 Centred-log-ratio

To address these issues, Aitchison (1986) introduced the clr transformation $\mathcal{S}^d \to \mathbb{R}^d$, defined by

$$\boldsymbol{\xi} = \text{clr}(\mathbf{x}) = \left[\log\left(\frac{x_1}{\left(\prod_{i=1}^d x_i\right)^{1/d}} \right) \; ... \; \log\left(\frac{x_d}{\left(\prod_{i=1}^d x_i\right)^{1/d}} \right) \right], \tag{5.2.1}$$

preserving operations and metrics from the simplex into the real space. This is an isometric transformation of the simplex with the Aitchison metric, onto a real sub-space with the ordinary Euclidean metric, hence

$$\text{clr}((\alpha \otimes \mathbf{x}) \oplus (\beta \otimes \mathbf{y})) = \alpha \cdot \text{clr}(\mathbf{x}) + \beta \cdot \text{clr}(\mathbf{y}) \tag{5.2.2}$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \text{clr}(\mathbf{x}), \text{clr}(\mathbf{y}) \rangle \tag{5.2.3}$$

$$||\mathbf{x}||_a = ||\text{clr}(\mathbf{x})|| \tag{5.2.4}$$

$$d(\mathbf{x}, \mathbf{y})_a = d(\text{clr}(\mathbf{x}), \text{clr}(\mathbf{y})) \tag{5.2.5}$$

The inverse clr transformation is

$$\mathbf{x} = \text{clr}^{-1}(\mathbf{v}) = \mathcal{C} \exp(\mathbf{v}). \tag{5.2.6}$$

The clr transformation is symmetrical in the components, but the price is a sum to zero constraint on the components of the transformed sample ($\sum_i \xi_i = 0$). Any transformed composition will lie on a plane which goes through the origin of $\mathbb{R}^d$ and is orthogonal to the vector of unities $[1, ..., 1]$. This property can effect the analysis of random compositions, as the covariance matrix of $\boldsymbol{\xi}$ is singular. Furthermore, clr transformations are subcompositionally incoherent. When different subsets of parts are considered the clr transformed results will differ in general, which can have sever consequences for bivariate data analysis such as pair-wise correlation coefficients (Filzmoser et al., 2010).

5.3 Isometric-log-ratio

The clr transformation assigns each composition in $\mathcal{S}^d$ to a row vector in $\mathbb{R}^d$ which sums to zero. This implies that we can find $d-1$ linearly independent vectors using the clr coordinates to obtain an orthonormal basis of the linear subspace. The isometric-log-ratio transformation (ilr) (Egozcue et al., 2003) is the projection of the compositional vector $\mathbf{x} \in \mathcal{S}^d$ onto an Aitchison orthonormal basis $\mathbf{e}_1, ..., \mathbf{e}_{d-1} \in \mathcal{S}^d$ from the Aitchison dot product (Aitchison, 1982) (Appendix 5.4.2),

$$\text{ilr}(\mathbf{x}) = [\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, ..., \langle \mathbf{x}, \mathbf{e}_{d-1} \rangle_a]. \tag{5.3.1}$$

Hence $\text{ilr}(\mathbf{e}_i) = \overrightarrow{\mathbf{e}_i}$, for $i = 1, ..., d-1$; $\overrightarrow{\mathbf{e}_i}$ being the ith vector in the canonical basis in $\mathbb{R}^{d-1}$.

5.3.1 Projection onto an orthonormal basis

The ilr transformation is the series of a projections onto an orthonormal basis in $\mathcal{S}^d$. If $\mathbf{M}$ is a $k \times k$ symmetric matrix of real numbers, then all the eigenvalues of $\mathbf{M}$ are real numbers and there exists an orthonormal basis $\mathbb{R}^k$ consisting of eigenvectors of $\mathbf{M}$. We can exploit this by defining the Aitchison dot product in terms of the clr (isometric) transformation. If $\mathbf{a} = \text{clr}(\mathbf{x})$ and $\mathbf{b} = \text{clr}(\mathbf{y})$, then the Aitchison dot product of the column vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ satisfying the sum

to zero constraint, can be expressed as

$$\begin{aligned}\langle \mathbf{x}, \mathbf{y} \rangle_a = \mathbf{a}^T \mathbf{b} &= \frac{1}{d} \sum_{i<j} \log\left(\frac{x_i}{x_j}\right) \log\left(\frac{y_i}{y_j}\right) \\ &= \frac{1}{d} \sum_{i<j} (a_i - a_j)(b_i - b_j) = \frac{1}{d} \mathbf{a}^T \mathbf{M} \mathbf{b}. \end{aligned} \tag{5.3.2}$$

The $d \times d$ symmetric matrix $\mathbf{M}$ is

$$\begin{pmatrix} d-1 & -1 & -1 & ... & -1 \\ -1 & d-1 & -1 & ... & -1 \\ -1 & -1 & d-1 & ... & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & -1 & ... & d-1 \end{pmatrix}. \tag{5.3.3}$$

The matrix $\mathbf{M}$ is degenerate, with $d-1$ non-zero eigenvalues d and one 0 (Appendix 5.4.3). The eigenspace with the eigenvalue of d, is the linear subspace space spanned by the vectors defined by the condition $\sum_{i=1}^{d} a_i = 0$. Consequently $\mathbf{M}\text{clr}(\mathbf{y})$ is a column eigenvector,

$$\mathbf{M}\text{clr}(\mathbf{y}) = \text{clr}(\mathbf{y}) d. \tag{5.3.4}$$

In order to obtain an orthonormal basis of the linear subspace associated with the eigenvalue d, a set of $d-1$ linearly independent vectors ($\in \mathbb{R}^d$) are selected from the subspace. The independent vectors $\mathbf{v}_1, ..., \mathbf{v}_{d-1}$ are defined as

$$\mathbf{v}_i = [0, .., 0, 1, -1, 0, ..., 0], \tag{5.3.5}$$

the first non element being placed in the ith column. These vectors are independent and sum to 0 (just as the eigenvectors). Applying the Gram-Schmidt procedure obtains the orthonormal vectors $\mathbf{u}_i \in \mathbb{R}^d, i = 1, 2, ..., d-1$, constituting an orthonormal basis of $(d-1)$-dimensional linear

subspace $V_{\mathcal{S}}$

$$\mathbf{u}_i = \sqrt{\frac{i}{i+1}}\left[\underbrace{\frac{1}{i}, \frac{1}{i}, ..., \frac{1}{i}}_{i \text{ elements}} - 1, ..., 0, 0\right]. \tag{5.3.6}$$

These vectors can then be back transformed into the simplex space using $\text{clr}^{-1}(\mathbf{u}_i)$ to give the orthonormal basis in $\mathcal{S}^d$. Thus

$$\mathbf{e}_i = \mathcal{C}\left[\exp\left(\sqrt{\frac{i}{i+1}}\left[\underbrace{\frac{1}{i}, \frac{1}{i}, ..., \frac{1}{i}}_{i \text{ elements}}, -1, ..., 0, 0\right]\right)\right]. \tag{5.3.7}$$

The ilr transformation for any composition $\mathbf{x} \in \mathcal{S}^d$ associated to an Aitchison orthonormal basis in $\mathcal{S}^d$, $\mathbf{e}_i$ $i = 1, 2, ..., d-1$, is the transformation from $\mathcal{S}^d$ to $\mathbb{R}^{d-1}$ given by

$$\mathbf{y} = \text{ilr}(\mathbf{x}) = [\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, ..., \langle \mathbf{x}, \mathbf{e}_{d-1} \rangle_a]. \tag{5.3.8}$$

In the case of $\mathbf{e}_i$ in (5.3.7) for $i = 1, ..., d-1$, the ilr is

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{e}_i \rangle_a &= \langle \text{clr}(\mathbf{x}), \mathbf{u}_i \rangle \\
&= \log\left(\frac{x_1}{g(x)}\right)\frac{\sqrt{i}}{i\sqrt{i+1}} + ... + \log\left(\frac{x_i}{g(x)}\right)\frac{\sqrt{i}}{i\sqrt{i+1}} - \log\left(\frac{x_{i+1}}{g(x)}\right)\sqrt{\frac{i}{i+1}} \\
&= \frac{\sqrt{i}}{i\sqrt{(i+1)}}\log(x_1 ... x_i) - \log(g(x))\sqrt{\frac{i}{i+1}} - \log(x_i + 1)\sqrt{\frac{i}{i+1}} + \log(g(x))\sqrt{\frac{i}{i+1}} \\
&= \sqrt{\frac{i}{i+1}}\log\left(\frac{g(x_1, ..., x_i)}{x_i + 1}\right),
\end{aligned} \tag{5.3.9}$$

where $g(x)$ is the geometric mean. The transformation has the benefit of persevering all the compositional geometry operations in the transformed space, without a constraint on the components (as with clr).

The inverse ilr transformation corresponds to the expression of $\mathbf{x}$ in the reference basis of $\mathcal{S}^d$

$$\mathbf{x} = \text{ilr}^{-1}(\mathbf{y}) = \bigoplus_{i=1}^{d-1}\left(\langle \mathbf{y}, \overrightarrow{\mathbf{e}_i} \rangle \otimes \mathbf{e}_i\right), \tag{5.3.10}$$

where $\langle \mathbf{y}, \overrightarrow{\mathbf{e}_i} \rangle = \langle \mathbf{x}, \mathbf{e}_i \rangle_a = y_i$ (and $\langle\rangle$ is the Euclidean dot product).

5.3.2 Transformation with respect to a group of parts and their balances

As there is no obvious canonical basis in $\mathcal{S}^d$, the choice of $\mathbf{u}_i$ which is back transformed, can be adjusted in order to reflect some subcomposition whilst still forming an orthonormal basis. Egozcue and Pawlowsky-Glahn (2005) introduce special orthonormal bases associated with a sequential binary partition of a compositional vector. To define the ilr transformation in this context (Egozcue et al. (2003), Egozcue and Pawlowsky-Glahn (2005)), a general vector $\mathbf{h}$ (this is the choice of vectors which generate the orthonormal basis $\in \mathcal{S}^d$) is used. First, a non-normalised vector

$$\mathbf{h}_i^* = \mathcal{C}\left[\exp\left(\left[\underbrace{0,...,0}_{j \text{ elements}}, \underbrace{\frac{1}{r},...,\frac{1}{r}}_{r \text{ elements}}, \underbrace{-\frac{1}{s},...,-\frac{1}{s}}_{s \text{ elements}}, \underbrace{0,...,0}_{t \text{ elements}}\right]\right)\right] \quad j+r+s+t=d. \tag{5.3.11}$$

is expressed.

The vector $\mathbf{h}_i^* \in \mathbf{S}^d$ is scaled to be of unit length, to form the orthonormal basis for the *balances*. The scaling is performed by transforming the vector with a clr transformation, scaling, then transforming back to the simplex space. As the norm $||\mathbf{h}_i^*||_a = \sqrt{\text{clr}(\mathbf{h}_i^*) \cdot \text{clr}(\mathbf{h}_i^*)}$, the normalised vector

$$\begin{aligned}\mathbf{h}_i &= \mathbf{h}_i^* \otimes ||\mathbf{h}_i^*||_a^{-1} \\ &= \text{clr}^{-1}\left(\text{clr}(\mathbf{h}_i^*) \times \sqrt{\frac{rs}{s+r}}\right)\end{aligned}$$

is thus,

$$\mathbf{h}_i = \mathcal{C}\left[\exp\left(\left[\underbrace{0,...,0}_{k \text{ elements}}, \underbrace{\sqrt{\frac{s}{r(s+r)}},...,\sqrt{\frac{s}{r(s+r)}}}_{r \text{ elements}}, \underbrace{-\sqrt{\frac{r}{s(s+r)}},...,-\sqrt{\frac{r}{s(s+r)}}}_{s \text{ elements}}, \underbrace{0,...,0}_{t \text{ elements}}\right]\right)\right],$$

where $k+r+s+t=d$.

To obtain the ilr transformation, a dot product operation is performed in the Aitchison space by (5.3.2), which equates to the Euclidean dot product after clr transformation. Here i is determined by the choices of k, r, s and t in $\mathbf{h}_i$. The closure operation $\mathcal{C}$, which scales the vector to 1 (and amounts to dividing the elements by a normalising constant) can be ignored, as this constant cancels in the clr transformation.

$$\begin{aligned}
\text{ilr}(\mathbf{x})_i =& \langle \mathbf{x}, \mathbf{h}_i \rangle_a \\
=& 0 + ... + 0 + \log\left(\frac{x_{k+1}}{g(\mathbf{x})}\right)\sqrt{\frac{s}{r(s+r)}} + ... + \log\left(\frac{x_{k+r}}{g(\mathbf{x})}\right)\sqrt{\frac{s}{r(s+r)}} + \\
& - \log\left(\frac{x_{k+r+1}}{g(\mathbf{x})}\right)\sqrt{\frac{r}{s(s+r)}} - ... - \log\left(\frac{x_{k+r+s}}{g(\mathbf{x})}\right)\sqrt{\frac{r}{s(s+r)}} + 0 + ... + 0 \\
=& \sqrt{\frac{s}{r(s+r)}} \sum_{i=k+1}^{k+r} \log(x_i) - r\sqrt{\frac{s}{r(s+r)}} \log(g(\mathbf{x})) + \\
& - \sqrt{\frac{r}{s(s+r)}} \sum_{i=k+r+1}^{k+r+s} \log(x_i) + s\sqrt{\frac{r}{s(s+r)}} \log(g(\mathbf{x}))
\end{aligned}$$

The elements in $\mathbf{h}_i$ ensure that $g(\mathbf{x}) = 1$, so the $\log(g(\mathbf{x}))$ terms are 0. Thus,

$$\begin{aligned}
\text{ilr}(\mathbf{x})_i =& \sqrt{\frac{sr}{(s+r)}} \log(g(x_{k+1}...x_{k+r})) - \sqrt{\frac{sr}{(s+r)}} \log(g(x_{k+r+1}...x_{k+r+s})) \\
=& \sqrt{\frac{rs}{r+s}} \log\left(\frac{g(x_{k+1}, ..., x_{k+r})}{g(x_{k+r+1}, ..., x_{k+r+s})}\right).
\end{aligned}$$

The ilr transformation is therefore

$$\text{ilr}(\mathbf{x})_i = \sqrt{\frac{rs}{r+s}} \log\left(\frac{g(x_{k+1}, ..., x_{k+r})}{g(x_{k+r+1}, ..., x_{k+r+s})}\right), \tag{5.3.12}$$

where each transformation coordinate i depends on the orthonormal basis $\mathbf{h}_i$.

As the choice of $\mathbf{h}_i$ (5.3.11) determines the ith transformation, Egozcue and Pawlowsky-Glahn (2005) introduce sequential binary partition to give an intuitive meaning to the orthogonal projections. The compositional vectors are partitioned into relevant non overlapping sets, where we separate parts $x_{k+1}, ..., x_{k+r}$ (r parts) from $x_{k+r+1}, ..., x_{k+r+s}$ (s parts) to define the i-order binary

partition (the orthonormal vector $\mathbf{h}_i \in \mathcal{S}^d$) called the *balancing element* as

$$\mathbf{e}_i = \mathcal{C}\left[\exp\left(\left[\underbrace{0,...,0}_{k \text{ elements}}, \underbrace{a,...,a}_{r \text{ elements}}, \underbrace{b,...,b}_{s \text{ elements}}, \underbrace{0,...,0}_{t \text{ elements}}\right]\right)\right] \quad k+t+s+r=d, \tag{5.3.13}$$

where

$$a = \sqrt{\frac{s}{r(r+s)}} \qquad b = -\sqrt{\frac{r}{s(r+s)}}. \tag{5.3.14}$$

The corresponding projections are the normalised log ratios of the geometric mean of each group of parts

$$\begin{aligned} \text{ilr}(\mathbf{x})_i &= \langle \mathbf{x}, \mathbf{e}_i \rangle_a \\ &= \log\left(\frac{(x_{k+1}...x_{k+r})^a}{(x_{k+r+1}...x_{k+r+s})^b}\right) = \sqrt{\frac{rs}{r+s}} \log\left(\frac{(x_{k+1}...x_{k+r})^{1/r}}{(x_{k+r+1}...x_{k+r+s})^{1/s}}\right), \end{aligned} \tag{5.3.15}$$

or the log contrasts between the groups. These are called balances as the expression is a ratio of geometric means which measures the relative weight of each group. The logarithm provides the appropriate scale and the square root coefficient is a normalising constant allowing a comparison of numerically different balances. A positive balance means that the group of parts in the numerator has more weight in the composition than the group in the denominator (and conversely for negative balances).

5.3.3 Relationship between transformations

As the clr transformation is isometric, the clr of the Aitchison dot product is

$$\text{clr}((\alpha \otimes \mathbf{x}_1) \oplus (\beta \otimes \mathbf{x}_2)) = \alpha \cdot \text{clr}(\mathbf{x}_1) + \beta \cdot \text{clr}(\mathbf{x}_2), \tag{5.3.16}$$

with $\mathbf{x}_1, \mathbf{x}_2$ in $\mathcal{S}^d$ and $\alpha, \beta \in \mathbb{R}^1$. If the row vector $\mathbf{x} \in S^d$ and $\text{ilr}(\mathbf{x}) = [y_1, ..., y_{d-1}]$, then we can construct $\mathbf{x}$ with the Aitchison operations as

$$\mathbf{x} = \bigoplus_{k=1}^{d-1} (y_k \otimes \mathbf{e}_k). \tag{5.3.17}$$

If we perform the clr transformation on $\mathbf{x}$

$$\text{clr}(\mathbf{x}) = \sum_{k=1}^{d-1} y_k \text{clr}(\mathbf{e}_k) = \sum_{k=1}^{d-1} y_k \mathbf{u}_k = \text{ilr}(\mathbf{x})\mathbf{U}, \tag{5.3.18}$$

where we use the construction of the orthonormal basis from the clr transformation (so $\text{clr}(\mathbf{e}_k) = \mathbf{u}_k$) . The $d-1 \times d$ matrix $\mathbf{U}$ has the orthonormal vectors $\text{clr}(\mathbf{e}_i)$ as row vectors. Thus

$$\text{clr}(\mathbf{x}) = \text{ilr}(\mathbf{x})\mathbf{U}. \tag{5.3.19}$$

The relationship between alr and clr is given by (Aitchison, 1986)

$$\text{alr}(\mathbf{x}) = \text{clr}(\mathbf{x})\mathbf{F}, \quad \mathbf{F}^T = [\mathbf{I}_{d-1} : -\mathbf{1}_{d-1}^T] \tag{5.3.20}$$

where $\mathbf{I}_{d-1}$ is the identity matrix of dimension $(d-1)$ and $\mathbf{1}_{d-1}$ is a $(d-1)$ row vector of units. The inverse relationship between *clr* and *alr* can be expressed as

$$\text{clr}(\mathbf{x}) = \text{alr}(\mathbf{x})\mathbf{A} \tag{5.3.21}$$

where the $(d-1) \times d$ matrix $\mathbf{A}$ is the pseudo inverse of matrix $\mathbf{F}$

$$\mathbf{A} = \frac{1}{d} \begin{pmatrix} d-1 & -1 & -1 & .. & -1 \\ -1 & d-1 & -1 & .. & -1 \\ -1 & -1 & d-1 & .. & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & .. & d-1 & -1 \end{pmatrix}. \tag{5.3.22}$$

Therefore we have the following relationships

$$\text{alr}(\mathbf{x}) = \text{ilr}(\mathbf{x})\mathbf{U}\mathbf{F} \quad \text{ilr}(\mathbf{x}) = \text{alr}(\mathbf{x})\mathbf{A}\mathbf{U}^T. \tag{5.3.23}$$

The inverse transformation of the coordinates $\text{ilr}(\mathbf{x}) = \mathbf{y}$ is

$$\text{ilr}^{-1}(\mathbf{y}) = \mathbf{x} = \bigoplus_{i=1}^{d-1} y_i \otimes \mathbf{e}_i, \quad y_i = \text{ilr}(x_i) = \langle \mathbf{x}, \mathbf{e}_i \rangle_a. \tag{5.3.24}$$

However a much easier approach is to transform ilr coordinates to clr coordinates using the orthonormal basis matrix

$$\begin{aligned} \text{clr}(\mathbf{x}) &= \text{ilr}(\mathbf{x})\mathbf{U} \\ &= \mathbf{y}\mathbf{U} \end{aligned} \tag{5.3.25}$$

A simple algorithm to recover $\mathbf{x}$ from its coordinates $\text{ilr}(\mathbf{x})$ consists of the following steps:

1. Construct the contrast matrix of the basis $\mathbf{U}$.
2. Compute the matrix product $\mathbf{y}\mathbf{U}$.
3. Apply $\text{clr}^{-1}(\mathbf{y}\mathbf{U})$.

5.3.4 Isometric-log-ratio and the balance interpretation

There are multiple ways to define orthonormal bases in the simplex. The main criterion for the selection of an orthonormal basis, is that it enhances the interpretability of the representation in coordinates. For instance, when performing principal component analysis an orthogonal basis is selected so that the first coordinate (principal component) represents the direction of maximum variability. As outlined in Section 5.3.2, Egozcue and Pawlowsky-Glahn (2005) link the bases for ilr transformation to a sequential binary partition of the compositional vector, so that they are easily interpreted in terms of grouped parts of the composition. The Cartesian coordinates of

a composition in such a basis are called *balances* (or ilr($\mathbf{x}$)) and the compositions of the basis *balancing elements*. A sequential binary partition is a hierarchy of the parts of a composition. In the first order of the hierarchy, all parts are split into two groups.

Table 5.1: Example of sign matrix, used to encode a sequential binary partition and build an orthonormal basis. The lower part of the table shows the matrix $\mathbf{U}$ of the basis, each vector is referred to as a *balancing element*.

order	x_1	x_2	x_3	x_4	x_5	x_6	r	s
1	+1	+1	-1	-1	+ 1	+1	4	2
2	+1	-1	0	0	-1	-1	1	3
3	0	+1	0	0	-1	-1	1	2
4	0	0	0	0	+1	-1	1	1
5	0	0	-1	+1	0	0	1	1
1	$+\frac{1}{\sqrt{12}}$	$+\frac{1}{\sqrt{12}}$	$-\frac{1}{\sqrt{3}}$	$-\frac{1}{\sqrt{3}}$	$+\frac{1}{\sqrt{12}}$	$+\frac{1}{\sqrt{12}}$		
2	$+\frac{\sqrt{3}}{2}$	$-\frac{1}{\sqrt{12}}$	0	0	$-\frac{1}{\sqrt{12}}$	$-\frac{1}{\sqrt{12}}$		
3	0	$+\frac{\sqrt{2}}{\sqrt{3}}$	0	0	$-\frac{1}{\sqrt{6}}$	$-\frac{1}{\sqrt{6}}$		
4	0	0	0	0	$+\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$		
5	0	0	$+\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	0	0		

In the following steps, each group is in turn split into two groups, and the process continues until all groups have a single part, as illustrated in Table 5.1. For each order of the partition, one can define the *balance* between the two sub-groups formed at that level: if $i_1, i_2, ..., i_r$ are the r parts of the first sub-group (coded by +1), and $j_1, j_2, ..., j_s$ the s parts of the second (coded by -1), the *balance* is defined as the normalised logratio of the geometric mean of each group of parts:

$$b_{order} = \sqrt{\frac{rs}{r+s}} \log\left(\frac{(x_{i1}...x_{ir})^{1/r}}{(x_{j1}...x_{js})^{1/s}}\right) = \log\left(\frac{(x_{i1}...x_{ir})^{a_+}}{(x_{j1}...x_{js})^{a_-}}\right) \tag{5.3.26}$$

where

$$a_+ = \frac{1}{r}\sqrt{\frac{rs}{r+s}} \quad a_- = -\frac{1}{s}\sqrt{\frac{rs}{r+s}} \quad a_0 = 0. \tag{5.3.27}$$

The vector $\mathbf{h}_i$ is called the *balancing element* and each ilr transformation $x_i^* = b_{order}$ is called a *balance* (5.3.26). This can be expressed in terms of a linear combination of the logarithms of the

parts in which coefficients add to zero

$$b_{order} = \sum_i \exp(a_+) \log(x_{ir}) + \sum_j \exp(a_-) \log(x_{js}), \tag{5.3.28}$$

hence a *balance* is a log contrast.

The interpretation of balances (5.3.26) relays on some of its properties. The geometric means are central values of the parts in each group of parts; its ratio measures the relative weight of each group; the logarithm provides the appropriate scale; and the square root coefficient is a normalising constant which allows us to compare numerically different balances.

The *balance* also has an intuitive interpretation. Imagine a political election where the parties are divided into two groups, either left and or right wing (with more than one party in each wing). If you only have the percentages within each group, you are unable to know which party and the respective wing, has won the election. The balance between the two wings will complete the information on the actual state of the election. The balance is the remaining relative information about the elections, once the information within the two wings has been removed.

For example, suppose that the composition of the votes for the six parties who contest the election is $\mathbf{x} \in \mathcal{S}^6$. The left wing consists of 4 parties represented by the group of parts $\{x_1, x_2, x_5, x_6\}$ and the right wing the remaining parts $\{x_3, x_4\}$. Consider the sequential binary partition in Table 5.1. The first partition just separates the two wings and thus the balance informs us about the equilibrium between the left and right.

A variety of questions regarding compositions are easily handled using the balances. If we are only interested in the relationships between the parties within the left wing we may wish to remove the information on the right wing. A traditional approach to this is to remove parts x_3 and x_4 and then close the remaining subcomposition. However, this is equivalent to projecting the composition of 6 parts orthogonally onto the subspace associated with the left wing, which is easily done by

setting $b_5 = 0$. The obtained projected composition is

$$\mathbf{x}_{proj} = \mathcal{C}[x_1, x_2, g(x_3, x_4), g(x_3, x_4), x_5, x_6], \quad g(x_3, x_4) = (x_3 x_4)^{1/2}, \tag{5.3.29}$$

where each part in the right wing has been substituted by the geometric mean within the right wing. This composition still contains the information on the left-right balance, b_1. If we are also interested in removing it ($b_1 = 0$), the remaining information will be only that within the left-wing subcomposition which is represented by the orthogonal projection

$$\mathbf{x}_{left} = \mathcal{C}[x_1, x_2, g(x_1, x_2, x_5, x_6), g(x_1, x_2, x_5, x_6), x_5, x_6]. \tag{5.3.30}$$

5.4 Appendix

The Euclidean geometry is not a proper geometry for compositional data. For example, consider the compositions $[5, 65, 30]$, $[10, 60, 30]$, $[50, 20, 30]$ and $[55, 15, 30]$. Intuitively, the difference between $[5, 65, 30]$ and $[10, 60, 30]$ is not the same as the difference between $[50, 20, 30]$ and $[55, 15, 30]$. The Euclidean distance is the same, as there is a difference of 5 units both between the first and the second respective components. But in the first case, the proportion in the first component is doubled, while in the second case, the relative increase is about 10%. The unit simplex structure has its own geometry and specific operators to account for these compositional characteristics, introduced by Aitchison (1986). This first two sections of this Appendix provide a brief summary of this geometric space.

5.4.1 Vector space structure

Given any d-part compositions $\mathbf{x}$, $\mathbf{y} \in \mathcal{S}^d$ their perturbation is

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1 \ x_2 y_2 \ ... \ x_d y_d], \tag{5.4.1}$$

where $\mathcal{C}$ is the closure or normalizing operation in which the elements of a positive vector are divided by their sum. The power transformed composition, where a is a real number is

$$a \otimes \mathbf{x} = \mathcal{C}[x_1^a \; x_2^a \; ... \; x_d^a] \tag{5.4.2}$$

The operations of perturbation $\oplus$ and power $\otimes$ play roles in the geometry of $\mathcal{S}^d$ analogous to translation and scalar multiplication in $\mathbb{R}^d$.

The simplex with perturbation and powering, $(\mathcal{S}^d, \oplus, \otimes)$, is a vector space. Thus, the following properties hold (Pawlowsky-Glahn et al., 2015);

Property 1: $(\mathcal{S}^d, \oplus)$ is a commutative group structure for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}^d$, it holds

1. Commutative property: $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$.
2. Associative property: $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$.

By analogy with standard operations in real space, $\mathbf{x} \oplus \mathbf{y}^{-1} = \mathbf{x} \ominus \mathbf{y}$ for the perturbation difference.

Property 2: Powering satisfies the properties of an external product. For $\mathbf{x}, \mathbf{y} \in \mathcal{S}^d, a, b \in \mathbb{R}^1$, it holds

1. associative property: $a \otimes (b \otimes \mathbf{x}) = (ab) \otimes \mathbf{x}$,
2. distributive property 1: $a \otimes (\mathbf{x} \oplus \mathbf{y}) = (a \otimes \mathbf{x}) \oplus (a \otimes \mathbf{y})$,
3. distributive property 2: $(a + b) \otimes \mathbf{x} = (a \otimes \mathbf{x}) \oplus (b \otimes \mathbf{x})$.

The closure operation cancels out any constant allowing us to omit the closure in intermediate steps of any computation without problem. This property can be expressed, for $\mathbf{z} \in \mathbb{R}_+^d$ and $\mathbf{x} \in \mathcal{S}^d$, as

$$\mathbf{x} \oplus (a \otimes \mathbf{z}) = \mathbf{x} \oplus (a \otimes \mathcal{C}(\mathbf{z})). \tag{5.4.3}$$

Nevertheless, one should be always aware that the closure constant is very important for the interpretation of the units. Therefore, controlling for the right units should be the last step in any

computation.

5.4.2 Inner product, norm and distance

To obtain a Euclidean vector space structure, we take the following inner product, with associated norm and distance (the subindex a stands for Aitchison).

The Aitchison inner product of $\mathbf{x}, \mathbf{y} \in \mathcal{S}^d$ is

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j}. \tag{5.4.4}$$

The Aitchison norm of $\mathbf{x} \in \mathcal{S}^d$ is

$$||\mathbf{x}||_a = \sqrt{\frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} \left(\log \frac{x_i}{x_j} \right)^2}. \tag{5.4.5}$$

The Aitchison distance between $\mathbf{x}$ and $\mathbf{y} \in \mathcal{S}^d$ is

$$d_a(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} \ominus \mathbf{y}||_a = \sqrt{\frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} \left(\log \frac{x_i}{x_j} - \log \frac{y_i}{y_j} \right)^2}. \tag{5.4.6}$$

The algebraic-geometric structure of $\mathcal{S}^d$ satisfies standard properties, such as compatibility of the distance with perturbation and powering for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}^d$ and $a \in \mathbb{R}^1$.

$$d_a(\mathbf{z} \oplus \mathbf{x}, \mathbf{z} \oplus \mathbf{y}) = d_a(\mathbf{x}, \mathbf{y}), \qquad d_a(a \otimes \mathbf{x}, a \otimes \mathbf{y}) = |a| d_a(\mathbf{x}, \mathbf{y}). \tag{5.4.7}$$

The Aitchison distance is subcompositionally coherent, as perturbation (5.4.1), powering (5.4.2), and inner product (5.4.4) induce the same linear vector space structure in the subspace corresponding to a subcomposition.

5.4.3 Determinant and eigenvalues for the matrix M

The $d \times d$ symmetric matrix $\mathbf{M}$ is

$$\mathbf{M} = \begin{pmatrix} d-1 & -1 & -1 & ... & -1 \\ -1 & d-1 & -1 & ... & -1 \\ -1 & -1 & d-1 & ... & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & -1 & ... & d-1 \end{pmatrix}. \tag{5.4.8}$$

The matrix $\mathbf{M}$ is degenerate with two different eigenvalues, 0 and d. These properties can be proved using the determinant lemma, where $\mathbf{A}$ is an invertible square matrix and $\mathbf{u}, \mathbf{v}$ are column vectors

$$\det(\mathbf{A} + \mathbf{u}\mathbf{v}^T) = (1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u})\det(\mathbf{A}). \tag{5.4.9}$$

Thus, $\mathbf{M}$ can be defined as

$$\mathbf{M} = \text{diag}(d) + \mathbf{1}_d\mathbf{1}_d^T(-1), \tag{5.4.10}$$

where $\text{diag}(d)$ is a diagonal matrix $(d \times d)$ and $\mathbf{1}_d$ is d dimensional vector of 1's. The determinant of $\mathbf{M}$, using (5.4.9) is thus

$$\begin{aligned} \det(\mathbf{M}) &= \left(1 - \mathbf{1}_d^T\text{diag}(d^{-1})\mathbf{1}_d\right) d^d \\ &= 0, \end{aligned}$$

the matrix is singular.

The eigenvalues are found by solving $\det(\mathbf{M} - \lambda\mathbf{I}) = 0$,

$$\begin{aligned} \det(\mathbf{M} - \lambda\mathbf{I}) &= \text{diag}(d - \lambda) + \mathbf{1}_d\mathbf{1}_d^T(-1) \\ &= \left(1 - \mathbf{1}^T\text{diag}((d-\lambda)^{-1})\mathbf{1}\right)\det(\text{diag}(d - \lambda)) \\ &= -\lambda(d-\lambda)^{d-1}. \end{aligned}$$

Setting this expression to zero and solving

$$0 = -\lambda(d - \lambda)^{d-1},$$

gives $d - 1$ eigenvalues of d and 1 eigenvalue of 0.

CHAPTER 6

Bayesian Compositional Regression with Microbiome Features via Variational Inference

6.1 Abstract

The microbiome plays a key role in the health of the human body. Interest often lies in finding features of the microbiome, alongside other covariates, which are associated with a phenotype of interest. One important property of microbiome data, which is often overlooked, is its compositionality as it can only provide information about the relative abundance of its constituting components. Typically, these proportions vary by several orders of magnitude in datasets of large dimensions. To address these challenges we develop a Bayesian hierarchical linear log-contrast model which is estimated by mean field Monte-Carlo co-ordinate ascent variational inference (CAVI-MC). We use novel priors which account for the large differences in scale and constrained parameter space associated with the compositional covariates. A Reversible Jump Monte Carlo Markov Chain guided by the data through univariate approximations of the variational posterior probability of

inclusion, with proposal parameters informed by approximating variational densities via auxiliary parameters, is used to estimate intractable marginal expectations. We demonstrate that our proposed method outperforms standard methods of variable selection applied to compositional data. We then apply the CAVI-MC to the analysis of real data exploring the relationship of the gut microbiome to body mass index.

Key words: Compositional, variational inference, microbiome, singular multivariate normal, Markov chain Monte Carlo.

6.2 Introduction

The human microbiome is the combined genome of the microorganisms that live in the human body. It has been estimated that these microbes make up to 10 trillion cells, equivalent to the number of human cells (Sender et al., 2016). Advances in genome sequencing technologies has enabled scientists to study these microbes and their function and to research microbiome–host interactions both in health and disease. The decreasing cost and increasing accessibility of nucleotide sequencing means it is the primary tool used to study the microbiome (Franzosa et al., 2015). Any microbiome dataset is compositional (Gloor et al., 2017) as the magnitude of a single operational taxonomic unit (OTU) depends on the sum of all the OTUs counts, and only provides information about the relative magnitudes of the compositional components. This means that the standard methods of analysis such as linear regression are not applicable to microbiome data (Li, 2015), unless a transformation is performed.

The large dimensions of these datasets often present a problem in variable selection where the number of covariates p exceeds the number of observations n ($p >> n$) and the space of possible combinations of significant variables is large, imposing a high computational burden. Sparse variable selection of the p covariates is expected, where just a few microbes are associated with the response. Bayesian variable selection approaches have the advantage of being able to include prior knowledge and simultaneously incorporate many sources of variation. Shrinkage priors encourage

the majority of regression coefficients to be shrunk to very small values when an estimator is applied identifying associations (Park and Casella, 2008). Alternatively, introducing latent variables produces posterior distributions of model inclusion and parameter values which enable model choice and a probabilistic understanding of the strength and nature of the association (Guan and Stephens, 2011). The different approaches within explicit variable selection are characterised by the location of the latent variable and its relationship with the covariates (George and McCulloch (1993), Kuo and Mallick (1998), Dellaportas et al. (2002)).

To model compositional data, a transformation must be performed to transfer the compositional vectors into Euclidean space. Various log ratio transformations have been proposed including additive log-ratio (alr), centred log-ratio (clr) (Aitchison, 1982) and more recently isometric log-ratio (ilr) (Egozcue et al., 2003). The ilr transformation defines balances proportional to the log difference between two groups which are scale invariant. Only the first coordinate can be interpreted as it represents all the relevant information about the compositional part.

The alr transformation, which constrains the associated parameter space to sum to 0, has proved to be useful in frequentist regression problems (Aitchison and Bacon-Shone, 1984), allowing a direct inference between selected covariates and the compositional data set. Lin et al. (2014) propose an adaptive l_1 regularisation regression for sparsity with the constraint imposed by the log contrasts. This has been extended to multiple linear constraints for sub-compositional coherence across predefined groups of predictors (Shi et al., 2016). A general approach to convex optimisation, where the model has been extended to the high-dimensional setting via regularization has recently been proposed by Combettes and Müller (2021). In the Bayesian framework Zhang et al. (2020) introduce a generalised transformation matrix on the parameters rather than the covariates, as a function of a tuning parameter c, similar to the generalized lasso. This ensures parameter estimates remain in the p space and as c reaches infinity the sum to zero constraint is imposed. By incorporating the matrix into conjugate prior and avoiding any singular distributions by not strictly imposing the zero sum constraint, a Gibbs sampler for the marginal posterior of the selection parameter can be derived. Alternative Bayesian approaches treat the the microbiome predictors as random, parameterised by a multivariate count model. Koslovsky et al.

(2020) combine this with the ilr transformation in a predictive model which identifies correlations across the microbiome. Li et al. (2019) cluster on a categorical covariate via a Gaussian mixture model in an ANOVA type model, but both approaches do not allow a direct inference between the compositional predictors and the response.

The abundances of features in microbiome data often differ by orders of magnitude. As far as we know this has not been explicitly accounted for in the current literature. In the Bayesian lasso (Park and Casella, 2008) separate scale parameters can have a hierarchical prior placed on them rather than this component being marginalised over which results in the Laplace prior. In the regularisation case, the choice of hyperprior defines how the parameters are shrunk to zero. This model is easily extended to the adaptive lasso (Leng et al., 2014) by positing independent exponential priors on each scale parameter, and then augmenting each tuning parameter with additional hyperpriors.

Typically, model selection is performed using Markov chain Monte Carlo (MCMC) methods. Various stochastic search based methods have been used to explore the model space in a computationally efficient manner (Lamnisos et al. (2013), Nott and Kohn (2005), Dellaportas et al. (2002)). Despite this body of work, MCMC can still be considered too slow in practice for sufficiently large scale problems. Variational inference is an alternative technique which uses optimisation to achieve computational savings by approximating the marginal posterior densities. Its success in machine learning problems has led to concerted efforts in the literature to encourage its use by statisticians (Blei et al. (2017), Ormerod and Wand (2010)). The speed of variational inference gives it an advantage, particular for exploratory regression, where a very large model is fitted to gain an understanding of the data and identify a subset of the microbiome which can be explored in more detail.

Approximate solutions arise in variational inference by restricting the family of densities which can be used as a proxy for the exact conditional density. Typically, the mean field variational family is used where independence is assumed across the factors. Thus by specifying conjugate priors, approximate marginal posteriors are members of the exponential family (Carbonetto and

Stephens, 2012). However, many models of interest such as logistic regression and non conjugate topic models, do not enjoy the properties required to exploit this algorithm. Using variational inference in these settings require algorithms to be adjusted to for the specific model requirement. A variety of strategies have been explored including alternative bounds (Jaakkola and Jordan (1997), Bishop and Svensen (2003)), numerical quadrature (Honkela and Valpola, 2005) and Monte Carlo approximation (Ye et al., 2020).

We propose a Bayesian hierarchical linear log-contrast model for compositional data which is estimated by mean field Monte Carlo co-ordinate ascent variational inference. We use the alr transformation proposed by Lin et al. (2014), because it is symmetric and removes the need to specify a reference category. Sparse variable selection is performed through novel priors within a hierarchical prior framework which account for the constrained parameter space associated with the compositional covariates and the different orders of magnitude in the taxon abundances. As our constrained priors are not conjugate, Monte Carlo expectations are used to approximate intractable integrals. These expectations are obtained via a reversible jump Monte Carlo Markov chain (RJMCMC) (Green, 1995), which is guided by the data through univariate approximations of the intractable variational posterior probability of inclusion. We exploit the nested nature of variational inference by proposing parameters from approximated variational densities via auxiliary parameters. Model averaging over all the explored models can be performed and shrunk estimates of the regression coefficient (by the model uncertainty) are available. The approach accommodates high dimensional microbial data and offers the potential to be scaled up for models with multiple responses.

We compare the performance of the proposed modelling approach with lasso, group lasso and Ordinary Least Squares (OLS) regressions on simulated data. The methods are then applied to a subset of the "Know Your Heart" cross-sectional study of cardiovascular disease (Cook et al., 2018) in order to examine the association of the gut microbiome with body mass index (BMI). The study was conducted in two Russian cities Novosibirsk and Arkhangelsk, enrolling 4542 men and women aged between 35-69 years recruited from the general population. A health check questionnaire was completed, providing information on smoking, weight and levels of alcohol consumption. We

analyse the microbiome of 515 subjects from the Arkhangelsk region at the phylum and genus level, as the 16S rRNA sequencing of faecal samples was only performed for these participants, alongside age and health covariates.

6.3 Methods

6.3.1 Microbiome Model

The microbiome data begins as raw counts for each taxon. Any zeros are replaced by a small pseudo-count (typically 0.5), before each row is standardised to sum to 1. The sample space of a vector of components is a simplex for each data point, where the rows of each vector make up the design matrix $\boldsymbol{Q}_{n\times d}$. The set of compositional explanatory variables can be transformed onto the unconstrained sample space $\mathbb{R}^{d-1}$ using the alr transformation

$$alr(\boldsymbol{q}_i) = \Big[\log\left(\frac{q_{i1}}{q_{id}}\right), \log\left(\frac{q_{i2}}{q_{id}}\right), ..., \log\left(\frac{q_{id-1}}{q_{id}}\right)\Big], \tag{6.3.1}$$

where $\boldsymbol{q}_i$ is the ith row of $\boldsymbol{Q}$ and the ratios have been arbitrarily chosen to involve the division of each of the first $d-1$ components by the final component. The log linear model, with the alr transformed variables as proposed by Aitchison and Bacon-Shone (1984), can be expressed as

$$y_i = \alpha \mathbf{1}_n + alr(\boldsymbol{q}_i)\tilde{\boldsymbol{\theta}} + \epsilon_i \tag{6.3.2}$$

where $\tilde{\boldsymbol{\theta}} = (\theta_1, ..., \theta_{d-1})^T$ is the corresponding $(d-1)$ vector of regression coefficients and ϵ_i is independent noise distributed as $N(0, \sigma^2)$. Although convenient, the interpretation of the model depends on the arbitrary choice of the reference category. If we expand the dot product $alr(\boldsymbol{q}_i)\cdot\tilde{\boldsymbol{\theta}}$ and set

$$\theta_d = -\sum_{j}^{d-1} \tilde{\theta}_j \tag{6.3.3}$$

the linear model can be conveniently expressed in matrix form (Lin et al., 2014) as

$$\mathbf{y} = \alpha \mathbf{1}_n + \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad \text{subject to } \sum_{j=1}^{d} \theta_j = 0 \tag{6.3.4}$$

where $\boldsymbol{Z} = (\log \boldsymbol{q}_1, ..., \log \boldsymbol{q}_d)$ is the $n \times d$ compositional design matrix and $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)^T$ is a d-vector of regression coefficients constrained to the affine hyperplane.

This likelihood is used by Zhang et al. (2020) who specify a d dimensional multivariate normal distribution on θ within a "spike-and-slab" prior,

$$\boldsymbol{\theta}|\sigma^2, \psi, \mathbf{V} \sim N_d(\mathbf{0}, \sigma^2 \psi \mathbf{V}), \qquad \mathbf{V} = \mathbf{I}_d - \frac{c^2}{1 + c^2 d}\mathbb{J}_d \tag{6.3.5}$$

where $\mathbb{J}_d$ is a matrix of ones and $\mathbf{V}$ is the generalised transformation matrix which incorporates the tuning parameter c to constrain the $\boldsymbol{\theta}$ parameter space and takes the form in (6.3.5) for the alr transformation. This approach allows the probability distribution to remain in the d dimensional space as $\mathbf{V}$ is a matrix of full rank, facilitating conjugate updates, as the sum to zero constraint is not imposed exactly.

Interest often lies in assessing the association of unconstrained data, in the form of categorical or continuous covariates against the response, alongside the microbiome. Two additional design matrices are added to the likelihood, $\boldsymbol{X}$ which comprises the scaled continuous covariates and $\boldsymbol{W}$ which contains the dummy variables for the $g = 1, ..., G$ categorical variables coded to indicate the m_g levels with respect to the intercept. The likelihood for our model is thus expressed as

$$\mathbf{y} = \alpha \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{\zeta} + \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad \text{subject to } \sum_{j=1}^{d} \theta_j = 0. \tag{6.3.6}$$

6.3.2 Compositional Priors

The linear constraint on the unconstrained vector can be expressed in matrix form as

$$\mathbf{T} = (\mathbf{I}_d - (1/d)\mathbb{J}_d) \tag{6.3.7}$$

where $\mathbf{T}$ is an idempotent matrix of rank $d-1$. If we originally parametrise $\theta_j \sim N(\mu_j, \psi_j)$, where the large differences in the order of magnitude of each row of the $\boldsymbol{Z}$ design matrix are accounted for by allowing each parameter θ_j to have a separate variance parameter ψ_j, then the constrained random variables associated with the compositional explanatory variables are from a singular multivariate normal distribution

$$\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\psi} \sim SMVN_d(\mathbf{T}\boldsymbol{\mu}, \mathbf{T}\text{diag}(\boldsymbol{\psi})\mathbf{T}^T). \tag{6.3.8}$$

with $\boldsymbol{\psi}$ a vector of scale parameters. This prior respects the sum to zero constraint imposed by the reparametrisation of the likelihood in (6.3.6). The distribution is degenerate, the transformation matrix $\mathbf{T}$ means the covariance matrix is singular, and will assign 0 values to all sets in the d dimensional space. Zhang et al. (2020) treat the constraint as a tuning parameter, restricting the values that $\boldsymbol{\theta}$ can take whilst still remaining in the d dimensional space so that the marginal posterior can be obtained in closed form. Our approach imposes the constraint exactly. The singular multivariate normal prior for the compositional data can be considered to be at the unobtainable limit of c in the alr transformation approach (6.3.5), when the tuning parameter creates a singular matrix where the standard normal prior is no longer appropriate.

We augment the prior on $\boldsymbol{\theta}$ with dependent latent indicator variables from a product of Bernoulli distributions which have been truncated to account for the alr transformation which prevents the selection of a single taxon into the model

$$p(\boldsymbol{\xi}|\kappa) \propto \prod_{j=1} \kappa^{\xi_j}(1-\kappa)^{1-\xi_j}\mathrm{I}\Big[\sum_j \xi_j \neq 1\Big], \tag{6.3.9}$$

where I is the indicator function. This truncation is particularly important in the presence of sparsity. The full singular multivariate normal spike-and-slab prior for $p(\boldsymbol{\theta}|\boldsymbol{\xi}) = p(\boldsymbol{\theta}_\xi|\boldsymbol{\xi})p(\boldsymbol{\theta}_{\bar{\xi}}|\boldsymbol{\xi})$, where $\boldsymbol{\theta}_\xi$ and $\boldsymbol{\theta}_{\bar{\xi}}$ are subvectors of $\boldsymbol{\theta}$ such that

$$p(\boldsymbol{\theta}_\xi|\Sigma, \boldsymbol{\xi}) = \frac{1}{(\det{}^*(2\pi\Sigma_\xi^+))^{(-1/2)}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}_\xi\Sigma_\xi^+\boldsymbol{\theta}_\xi\right) \quad \text{and} \quad p(\boldsymbol{\theta}_{\bar{\xi}} = 0|\boldsymbol{\xi}) = 1, \tag{6.3.10}$$

Σ_ξ^+ denotes the Moore-Penrose pseudo inverse of the matrix $\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi$ defined by $A^+ = VS^+U^T$ if $A = USV^T$ is the singular value decomposition of A and S^+ is the diagonal matrix which has the same entries as S and where $S_i^+ i = 1/S_{ii}$ for the nonzero diagonal entries. $\det^*$ the psuedo-determinant defined as the product of the nonzero eigenvalues of the matrix and $\boldsymbol{\xi}$ is a vector of zeros and ones. The $\boldsymbol{\theta}_\xi$ parameters are dependent (the covariance for unit scale is equal to the fraction $-1/d_\xi$ and for the case of $d_\xi = 2$ the correlation is 1). This prior implies a univariate spike-and-slab on the diagonal of the covariance matrix in (6.3.10).

$$p(\boldsymbol{\psi}|\boldsymbol{\xi}) = \prod_{j=1}^{d}\left[\frac{b_\psi^{a_\psi}}{\Gamma(a_\psi)}(\psi_j)^{-a_\psi-1}\exp\{-b_\psi\psi_j^{-1}\}\right]^{\xi_j}\delta_0(\psi_j)^{1-\xi_j} \quad \psi_j > 0 \;\forall\; j. \tag{6.3.11}$$

A beta distribution is placed on the sparsity parameter κ and the hyperparameter b_ψ is given a gamma prior. This approach can be interpreted as replacing the continuous mixing density in the Bayesian lasso, which can have either hierarchical structure (Leng et al., 2014) or be marginalised over (Park and Casella, 2008), with a discrete mixture. This set of explicit variable selection priors on the compositional data ensures that the marginal posterior of variable ξ_j represents the inclusion of the jth taxon in the model.

6.3.3 Priors

The choice of the remaining prior distributions is partly down to convenience. The prior distributions and likelihood are semi-conjugate pairs which means the optimal form for the mean field variational density is in the same exponential family form.

We employ a variable selection spike-and-slab prior George and McCulloch (1997) for β_s associ-

ated with the continuous variables in the design matrix $\boldsymbol{X}$, where each s parameter is independent. The spike is a point mass at 0 (Dirac distribution) with probability $1 - p(\gamma_s) = 1 - \omega$ and the slab is a zero centred Gaussian with variance w which requires the variables to be standardised. The binary latent indicator variable γ_s represents the inclusion of the sth covariate in the model.

In the case of the categorical data matrix, we are interested in selecting the group of variables associated with the response into the model, rather than a particular level. Each factor variable (or group) $g = 1, .., G$ has $j = 1, ..., m_g, m_{g+1}$ levels which are coded as dummy variables in $\boldsymbol{W}$ with reference to the intercept. Motivated by the Bayesian group lasso (Xu and Ghosh, 2015) who introduce binary indicators to perform selection both between and within the groups levels, we employ a variable selection spike-and-slab prior on the vector $\boldsymbol{\zeta}_g$ with dimension m_g. The spike is a point mass at 0 (Dirac distribution) with probability $1 - p(\chi_g) = 1 - \varrho$ and the slab is a zero centred Gaussian with variance v. The binary latent indicator variable χ_g represents the inclusion of the gth categorical variable into the model. In the case where there factors have just 2 levels, the prior reduces to the same form as its unrestricted continuous counterpart, with a different scale parameter.

Hierarchical priors are also included to fully incorporate the uncertainty surrounding these parameters. The probability that a given covariate in the design matrices of $\boldsymbol{X}$ and $\boldsymbol{W}$ affects the response is modelled by the parameters ω and ϱ, with beta priors. Inverse gamma distributions with gamma (shape and scale) hyperpriors on their respective scales are placed on the prior variance parameters w and v.

6.3.4 Variational Inference

We employ coordinate ascent variational inference (CAVI) (Blei et al., 2017) as our estimation procedure, rather than relying entirely on MCMC which often requires substantial computing resources when the dimensionality of the problem is large. We use the *mean field variational family*, but allow dependencies within each member (block), where the latent variables are mutually independent and each governed by a distinct factor in the variational density. We define the

blocks to ensure the dependency between the latent indicator variable(s) and their associated parameter(s) is captured. An example of a block is the joint q approximating density for the prior parameters $q(\beta_s, \gamma_s)$ directly associated with the design matrix $\boldsymbol{X}$. The full mean field approximation distribution $q(\boldsymbol{\vartheta})$ is defined in the Supplementary Materials.

6.3.5 Unconstrained Updates

The variational inference updates are available analytically for all unconstrained parameters and hyperparameters in the model. Derivations are given in the Supplementary Material. The updates involve a combination of univariate and multivariate calculations. The regression parameters directly associated with the $\boldsymbol{X}$ and $\boldsymbol{W}$ design matrices have joint updates in the same spike-and-slab form as their priors. The conjugate update for $q(\beta_s, \gamma_s)$ is

$$q(\beta_s|\gamma_s, \mathbf{y}) = \mathcal{N}(\mu_{\beta_s}, \sigma^2_{\beta_s})^{\gamma_s}\delta_0(\beta_s)^{1-\gamma_s} \qquad q(\gamma_s|\mathbf{y}) = Bern((\gamma_s)^{(1)}).$$

with free parameters

$$\begin{aligned}
\sigma^2_{\beta_s} &= \left(\|X_s\|^2(\sigma^{-2})^{(1)} + (w^{-1})^{(1)}\right)^{-1}, \\
\mu_{\beta_s} &= (\sigma^{-2})^{(1)}\sigma^2_{\beta_s}X_s^T\left(\mathbf{y} - (\alpha)^{(1)}\mathbf{1}_n - \sum_{k\neq s} X_k(\beta_k)^{(1)} - \sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_g)^{(1)} - \boldsymbol{Z}(\boldsymbol{\theta}_\xi)^{(1)}\right), \\
(\gamma_s)^{(1)} &= \left[1 + \exp\left\{(\log(1-\omega))^{(1)} - (\log\omega)^{(1)} - \frac{1}{2}\left((\log w^{-1})^{(1)} - \mu^2_{\beta_s}\sigma^{-2}_{\beta_s} - \log\left(\sigma^2_{\beta_s}\right)\right)\right\}\right]^{-1},
\end{aligned}$$

where $(\cdot)^{(1)}$ denotes the q expectation. The conjugate update for $q(\boldsymbol{\zeta}_g, \chi_g)$ is

$$q(\boldsymbol{\zeta}_g|\chi_g, \mathbf{y}) = \mathcal{N}_{m_g}(\boldsymbol{\mu}_{\zeta_g}, \Sigma_{\zeta_g})^{\chi_g}\delta_0(\boldsymbol{\zeta}_g)^{1-\chi_g} \qquad q(\chi_g|y) = Bern((\chi_g)^{(1)}),$$

where the free parameters for $\boldsymbol{\zeta}_g$ are updated by the multivariate extension of the previous univariate update,

$$
\begin{aligned}
\Sigma_{\zeta_g} &= \left[(\sigma^{-2})^{(1)}\boldsymbol{W}_g^T\boldsymbol{W}_g + (v^{-1})^{(1)}\right]^{-1}, \\
\boldsymbol{\mu}_{\zeta_g} &= (\sigma^{-2})^{(1)}\Sigma_{\zeta_g}\boldsymbol{W}_g^T\left(\mathbf{y} - (\alpha)^{(1)}\mathbf{1}_n - \sum_s X_s(\beta_s)^{(1)} - \sum_{k\neq g}\boldsymbol{W}_k(\boldsymbol{\zeta}_k)^{(1)} - \boldsymbol{Z}(\boldsymbol{\theta})^{(1)}\right), \\
(\chi_g)^{(1)} &= \Bigg[1 + \exp\Bigg\{(\log(1-\varrho))^{(1)} - (\log\varrho)^{(1)} - \frac{m_g}{2}(\log v^{-1})^{(1)} - \frac{1}{2}\boldsymbol{\mu}_{\zeta_g}^T\Sigma_{\zeta_g}^{-1}\boldsymbol{\mu}_{\zeta_g} + \\
&\quad - \frac{1}{2}\log\big(\det(\Sigma_{\zeta_g})\big)\Bigg\}\Bigg]^{-1}.
\end{aligned}
$$

The marginal expectation of $\boldsymbol{\zeta}_g$ and β_s is the mean of the conditional density when the parameter is included in the model, shrunk by the probability of being included in the model. The nested q density update for each free parameter(s) is the expectation of the log joint distribution with respect to all the other factors. Thus, any update involving a marginal expectation from a parameter with a spike and slab prior involves a form of regularisation.

The selection of the spike-and-slab priors for β_s, $\boldsymbol{\zeta}_g$ and $\boldsymbol{\theta}$ with sparsity inducing hyperparameters for variable selection, shrinks the parameters estimates in the variational updates rather then performing explicit variable selection as in MCMC. These estimates are a useful proxy for the final model effects, but as opposed to a model with regularisation priors, the expectation of the model indicator parameters gives us the probability of a covariate being associated with the response. In the case of $\boldsymbol{\zeta}_g$, which is associated with the gth categorical covariate, the parameterisation has a convenient interpretation. Each element in the vector is free to vary but all elements are shrunk by the same value. Thus the expectation $(\chi_g)^{(1)}$ is the probability of the categorical covariate (rather than the individual levels) being included in the model.

6.3.6 CAVI-MC

The conditional vector update $q(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{\xi})$ is available analytically and takes the form

$$q(\boldsymbol{\theta}_\xi|\boldsymbol{\xi}, \mathbf{y}) = SMVN_{d_\xi}(\mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi}, \mathbf{T}_\xi \Sigma_{\theta_\xi} \mathbf{T}_\xi), \quad q(\boldsymbol{\theta}_{\bar{\xi}}|\boldsymbol{\xi}, \mathbf{y}) = \delta_0(\boldsymbol{\theta}_{\bar{\xi}}), \tag{6.3.12}$$

where δ_0 is the Dirac distribution on the subvector $\boldsymbol{\theta}_{\bar{\xi}}$ with updates

$$\boldsymbol{\mu}_{\theta_\xi} = \Sigma_{\theta_\xi}(\sigma^{-2})^{(1)} \boldsymbol{Z}_\xi^T (\mathbf{y} - (\alpha)^{(1)} \mathbf{1}_n - \sum_s X_s (\beta_s)^{(1)} - \sum_g \boldsymbol{W}_g (\boldsymbol{\zeta}_g)^{(1)}) \tag{6.3.13}$$

$$\Sigma_{\theta_\xi} = \left((\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi) \mathbf{T}_\xi)^+ + (\sigma^{-2})^{(1)} \boldsymbol{Z}_\xi^T \boldsymbol{Z}_\xi \right)^{-1} \tag{6.3.14}$$

The truncated Bernoulli prior distributions for $\boldsymbol{\xi}$ and unique scale parameter ψ_j for each element in $\boldsymbol{\theta}$, prevents a conjugate posterior update for the joint block $q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi})$. All other updates are available analytically.

The difficult to compute joint $q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi})$ update is performed by inserting a Monte Carlo step within the mean field variational inference approach. We take advantage of the structure of the target density $p(\boldsymbol{\vartheta}, \mathbf{y}) \equiv f(\boldsymbol{\vartheta})$ (the data $\mathbf{y}$ is omitted for notational purposes as its fixed) which has the form

$$f(\boldsymbol{\vartheta}) = h(\boldsymbol{\vartheta}) \exp(\langle \boldsymbol{\eta}, T(\boldsymbol{\vartheta}) \rangle - A(\boldsymbol{\eta})), \quad \boldsymbol{\vartheta} \in S_p \tag{6.3.15}$$

for r-dimensional constant vector $\boldsymbol{\eta}$, vector function $T(\boldsymbol{\vartheta})$ and relevant scalar functions $h > 0$. In our case this admits the factorisation

$$h(\boldsymbol{\vartheta}) = h_{q(\vartheta_j)}(\vartheta_j) h_{q(\boldsymbol{\vartheta}_{-j})}(\boldsymbol{\vartheta}_{-j}), \qquad T_l(\boldsymbol{\vartheta}) = T_{l,j}(\vartheta_j) T_{l,-j}(\boldsymbol{\vartheta}_{-j}), \quad 1 \leq l \leq r, \text{ for all } j \notin \mathcal{J},$$

where $\mathcal{J}$ is the set of all analytically available updates. This allows us to avoid generating and storing the samples from the approximating densities which would involve considerable computational cost, by using the q marginal expectations in the Monte Carlo estimate for $q(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{\xi})$. Ye et al. (2020) show that, under regularity conditions, an MC-CAVI recursion will get arbitrarily

close to a maximiser of the evidence lower bound with any given high probability.

The MCMC approach involves two move types, within-model moves where the samples are generated from a Metropolis-Hastings sampler and between-model moves which are sampled from a RJMCMC. The samplers involve using some form of the joint approximating posterior $q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi}|\boldsymbol{y}) \propto \; q(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{y})q(\boldsymbol{\xi}, \boldsymbol{\psi}|\boldsymbol{y})$ which is simplified as $q(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{y})$ has the conjugate spike-and-slab form (6.3.12).

Randomly choose either a between-model move which consists of sequentially updating $\boldsymbol{\xi}, \boldsymbol{\psi}|\boldsymbol{\xi}$ and $\boldsymbol{\theta}|\boldsymbol{\psi}$, $\boldsymbol{\xi}$ or a within-model move where $\boldsymbol{\xi}$ is not updated. This naturally leads to questions regarding the proposals for $\boldsymbol{\psi}$ which has a constrained support and $\boldsymbol{\xi}$ which has the potential to be a very large binary space.

Between-model RJMCMC - Approximating $q(\boldsymbol{\xi}, \boldsymbol{\psi}|\boldsymbol{y})$ to $p(\boldsymbol{\xi}|\vartheta)$ for the proposal distribution $j_m(\boldsymbol{\xi}, \boldsymbol{\xi}')$

The choice of priors for the parameters associated with microbiome features, the indicator vector $\boldsymbol{\xi}$ and set of scale parameters $\boldsymbol{\psi}_\xi$, prevents a conjugate update for $q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi})$. An MCMC step is introduced to sample from the intractable q approximating posterior. To search the binary space we use a RJMCMC where the proposal for ψ_j conditional on $\xi_j = 1$ is from the q approximating density of the auxiliary parameter Ω_j

$$\pi(\psi_j|\xi_j = 1) = IG_q(a^*_{\Delta_j}, b^*_{\Delta_j}), \tag{6.3.16}$$

where the calculation of the free parameters $a^*_{\Delta_j}$ and $b^*_{\Delta_j}$ is explained in the next section. $\boldsymbol{\theta}$ is generated directly from the singular multivariate normal target distribution (6.3.12).

There is considerable research in sampling high-dimensional binary vectors. Lamnisos et al. (2009) propose a general model for the proposal which combines local moves with global ones by changing blocks of variables. They find that the acceptance rates for Metropolis-Hastings samplers that include, exclude or swap a single variable improves. Lamnisos et al. (2013) extend

their model with adaptive parameters which change during the mixing of the MCMC. Motivated by incorporating information from data into the proposal parameters, we use the variational inference posterior distribution $q(\boldsymbol{\xi}, \boldsymbol{\psi}|\mathbf{y})$ which is only available up to a constant of proportionality

$$
\begin{aligned}
q(\boldsymbol{\xi}, \boldsymbol{\psi}|\mathbf{y}) \propto \ \exp\Bigg(&\frac{1}{2}(\boldsymbol{\mu}^T_{\theta_{(\xi,\psi)}}\mathbf{T}_\xi(\mathbf{T}_\xi^T\Sigma_{\theta_{(\xi,\psi)}}\mathbf{T}_\xi)^+\mathbf{T}_\xi\boldsymbol{\mu}_{\theta_{(\xi,\psi)}}) + \frac{1}{2}\log\Big(\det{}^*(\mathbf{T}_\xi\Sigma_{\theta_{(\xi,\psi)}}\mathbf{T}_\xi)\Big) + \\
&\sum_j \xi_j(\log\kappa)^{(1)} - \frac{1}{2}\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)) + \sum_j(1-\xi_j)(\log(1-\kappa))^{(1)} + \\
&-(a_\psi+1)\sum_j \xi_j\log(\psi_j) - b_\psi\sum_j\xi_j\psi_j^{-1} + (a_\psi\log(b_\psi) - \log(\Gamma(a_\psi))\sum_j\xi_j\Bigg),
\end{aligned}
\tag{6.3.17}
$$

to obtain a univariate approximation relative to the jth element to guide the RJMCMC. These normalised probabilities are used to obtain our proposal probabilities in a birth-death and swap sampling scheme. Similar to adaptive parameters in MCMC, these selection probabilities are updated at each iteration of the CAVI.

The pseudo determinant in (6.3.17) is approximated by removing the constraints $\mathbf{T}_\xi$ and taking the MCMC expectation conditional on $\xi_j = 1$. So for the jth element the approximation is

$$
\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)) \approx \{\log(\psi_j)\}^{\{1\}}_{\emptyset} \tag{6.3.18}
$$

where the curly brackets $\{\}$ denote an MCMC expectation and $\emptyset$ defines an expectation over all non-zero values. A similar approach can be used to approximate the determinant containing Σ_{θ_ξ}

$$
\log\big(\det{}^*(\mathbf{T}_\xi\Sigma_{\theta_\xi}\mathbf{T}_\xi)\big) \approx \log\Big(\bar{\sigma}^2_{\theta_j}\Big).
$$

where $\bar{\sigma}^2_{\theta,tj}$ is the non-zero variance average over the MCMC iterations, obtained by extracting the diagonal from $\Sigma_{\theta_{(\xi,\psi)}}$ at each iteration. If the jth term has not been included in the model the term is approximated by

$$
\log\big(\det{}^*(\mathbf{T}_\xi\Sigma_{\theta_\xi}\mathbf{T}_\xi)\big) \approx \log\Big(\big[\|Z_j\|^2(\sigma^{-2})^{(1)}\big]^{-1}\Big) \tag{6.3.19}
$$

After approximating Σ_{θ_ξ} to a scalar for each jth element the matrix dot product reduces to

$$\boldsymbol{\mu}_{\theta_\xi}^T \mathbf{T}_\xi (\mathbf{T}_\xi^T \Sigma_{\theta_\xi} \mathbf{T}_\xi)^+ \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi} \approx \bar{\sigma}_{\theta_j}^2 \Big(\sum_j (1 - 1/d_\xi) \mu_{\theta_{\xi_j}}^2 - 2 \sum_{j<j'} (\mu_{\theta_{\xi_{j'}}} \mu_{\theta_{\xi_j}} / d_\xi) \Big). \tag{6.3.20}$$

To account for the cross product terms which contains the elements of $\boldsymbol{\xi}$ not equal to j and the associated $\boldsymbol{\mu}_\theta$ terms, a combination of conditional expectations and marginal expectations which shrink the values in proportion to its probability of being zero, is used. As ξ_j can not be separated from the sum in the numerator d_ξ, two approximations of the matrix dot product are used conditional on the expectation from the previous chain.

Defining the expectations with respect to the parameter currently being updated from the previous MCMC by a curly bracket as:

- $\{\mu_{\theta_j}\}_{\not\emptyset}^{\{1\}}$: Conditional expectation $\xi_j = 1$. Weighted average of the nonzero terms from previous chain,
- $\{\mu_{\theta_j}\}^{\{1\}}$: Expectation wrt q from the previous chain,
- $\{d_\xi\}^{\{1\}}$: Expectation wrt q from the previous chain,

the approximation of the dot product $(\mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi})^T \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi}$ is thus

$$\begin{aligned} &\bar{\sigma}_{\theta,j}^{-2} \left(\sum_j (1 - 1/\{d_\xi\}^{\{1\}}) \xi_j (\{\mu_{\theta_j}\}_{\not\emptyset}^{\{1\}})^2 - \frac{2}{\{d_\xi\}^{\{1\}}} \sum_{j<j'} \xi_j \{\mu_{\theta_{\xi_j}}\}_{\not\emptyset}^{\{1\}} \{\mu_{\theta_{\xi_{j'}}}\}^{\{1\}} \right) && \{d_\xi\}^{\{1\}} > 2 \\ &\bar{\sigma}_{\theta,j}^{-2} \sum_j \xi_j (\{\mu_{\theta_j}\}_{\not\emptyset}^{\{1\}})^2 && \{d_\xi\}^{\{1\}} < 2. \end{aligned}$$

Although $\{d_\xi \in \mathbb{N}_0 | d_\xi \leq d, d_\xi \neq 1\}$, the support of the MCMC expectation $\{d_\xi\}^{\{1\}}$ is the positive real line so we threshold on 2. When $\{d_\xi\}^{\{1\}} > 2$ the probabilities used in the proposal distribution

for the RJMCMC, derived from approximating Equation (6.3.17) and normalising is

$$\tilde{p}(\xi_j = 1|\boldsymbol{\vartheta}) \equiv \Bigg[\exp\Bigg\{ (\log(1-\kappa))^{(1)} - \frac{1}{2\bar{\sigma}^2_{\theta,j}} \Bigg((1 - 1/\{d_\xi\}^{\{1\}})(\{\mu_{\theta_j}\}^{\{1\}}_{\emptyset})^2 + \tag{6.3.21}$$

$$- \frac{2}{\{d_\xi\}^{\{1\}}} \{\mu_{\theta_{\xi_j}}\}^{\{1\}}_{\emptyset} \sum_{j' \neq j} \{\mu_{\theta_{\xi_{j'}}}\}^{\{1\}} \Bigg) - \frac{1}{2} \log\big(\bar{\sigma}^2_{\theta,j}\big) + \frac{1}{2} (\log \psi_j)^{\{1\}}_{\emptyset} - (\log \kappa)^{(1)} +$$

$$(\log \Gamma(a_\psi) - a_\psi \log b_\psi) + + (a_\psi + 1)(\log \psi_j)^{\{1\}}_{\emptyset} + b_\psi (\psi_j^{-1})^{\{1\}}_{\emptyset} \Bigg\} + 1 \Bigg]^{-1},$$

which contains the variational expectations and an MCMC conditional expectation from the previous iterations. This is then used to propose the various move types in the RJMCMC.

Pseudo Updates for MCMC proposals

A conjugate update for the parameters associated with the microbiome features $q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi})$ is prevented by the choice of priors for the indicator vector $\boldsymbol{\xi}$ and set of scale parameters $\boldsymbol{\psi}_\xi$. Samples from the intractable q approximating posterior are simulated from an MCMC step instead. The move types in the RJMCMC for $\boldsymbol{\xi}$ use an element-wise approximation of the joint $q(\boldsymbol{\xi})$ density (6.3.21). For the proposal distribution of $\boldsymbol{\psi}$, we use the model likelihood and an unconstrained approximation to the constrained priors. In order to do this we define auxiliary parameters (upper case Greek letters) which are unconstrained versions of the constrained parameters. We derive pseudo variational updates from an unconstrained model with a simpler prior parametrisation, then use the q approximating distribution of the relevant auxiliary parameter as our proposal for $\boldsymbol{\psi}$. We can think of the auxiliary parameters as introducing an alternative directed acyclic graph (DAG) which is updated first, helping us to approximate the model in order to guide the MCMC step. These updates are refined by the full variational inference updates which account for the constraint at each iteration. The parameter κ and the hyperparameters a_Δ, b_Δ which are set to a_ψ, b_ψ provide a link back to the constrained model.

The series of pseudo variational updates are determined from a simple prior parametrisation where the parameters associated with the compositional covariates are not constrained to sum to

0. This unconstrained model has the following prior parametrisation

$$p(\Omega_j|\Delta_j,\Upsilon_j) = N(\Omega_j|0,\Delta_j)^{\Upsilon_j}\delta_0(\Omega_j)^{1-\Upsilon_j} \qquad p(\Delta_j|\Upsilon_j) = IG(\Delta_j|a_\Delta,b_\Delta)^{\Upsilon_j}\delta_0(\Delta_j)^{1-\Upsilon_j}$$
$$p(\Upsilon_j) = Bern(\Upsilon_j|\kappa).$$

Where $\boldsymbol{\Omega}$ are the unconstrained version of the $\boldsymbol{\theta}$ parameters, $\boldsymbol{\Delta}$ are the variance parameters for $\boldsymbol{\Omega}$ which are both dependent on the model selection parameters $\boldsymbol{\Upsilon}$. The prior for the model selection parameter Υ_j is a simple Bernoulli distribution. The remaining priors and likelihood take the form defined in the initial prior parametrisation. The introduction of independence across each univariate $(\Omega_j,\Delta_j,\Upsilon_j)$ block, (where the data is being treated as unconstrained) ensures the q expectations are all available in closed form (derived in the Supplementary Section).

Despite the similarities of the prior parametrisation to (6.3.5), the addition of a separate scale parameter Δ_j for Ω_j prevents a joint conjugate update on the $(\Omega_j,\Delta_j,\Upsilon_j)$ block. Instead we update $q(\Omega_j,\Upsilon_j)$ (for $j=1,...,d$) before updating $q(\Delta_j|\Upsilon_j)$. Both require expectations conditional on Υ_j as well as the typical marginal expectations. The $q(\Omega_j,\Upsilon_j)$ update is

$$q(\Omega_j,\Upsilon_j) \propto N(\Omega_j|\mu_{\Omega_j},\sigma^2_{\Omega_j})^{\Upsilon_j}\delta_0(\Omega_j)^{1-\Upsilon_j} \tag{6.3.22}$$
$$\Big\{\exp\Big(\frac{1}{2}\log\sigma^2_{\Omega_j} + (\log\kappa)^{(1)} - \frac{1}{2}\mathbb{E}_q(\log\Delta_j|\Upsilon_j) + \frac{1}{2}\mu^2_{\Omega,j}\sigma^{-2}_{\Omega,j} + a_\Delta\log(b_\Delta) + \tag{6.3.23}$$
$$-\log(\Gamma(a_\Delta)) - (a_\Delta+1)\mathbb{E}_q(\log\Delta_j|\Upsilon_j) - b_\Delta\mathbb{E}_q[\Delta_j^{-1}|\Upsilon_j]\Big)\Big\}^{\Upsilon_j}\Big\{1-\kappa)^{(1)} + \delta_0(\Delta_j)\Big\}^{1-\Upsilon_j}$$

The binary form of the pseudo update for Ω_j and Υ_j enables us to determine the values for the conditional expectations. In Equation (6.3.22) we have under q, where we condition on the value of Υ_j

$$q(\Omega_j|\Upsilon_j=1,\mathbf{y}) = \mathcal{N}(\mu_{\Omega,j},\sigma^2_{\Omega,j}) \quad q(\Omega_j|\Upsilon_j=0,\mathbf{y}) = \delta_0(\Omega_j), \tag{6.3.24}$$

which allows us to set the expectations in the normal variance update as $\mathbb{E}_q[\Delta_j^{-1}|\Upsilon_j = 1]$

$$\sigma^2_{\Omega,j} = \left(\|Z_j\|^2(\sigma^{-2})^{(1)} + \mathbb{E}_q[\Delta_j^{-1}|\Upsilon_j = 1]\right)^{-1} \tag{6.3.25}$$

$$\mu_{\Omega,j} = \sigma^2_{\Omega,j} Z_j^T \left\{ (\sigma^{-2})^{(1)} \left(\mathbf{y} - \sum_{k \neq j} Z_k (\Omega_k)^{(1)} - \sum_s X_s (\beta_s)^{(1)} \right) \right\}. \tag{6.3.26}$$

The conditional expectation prevents us averaging over Υ_j which shrinks the marginal expectation, creating an update which has the same form as (6.3.5). Using the form of (6.3.23) to determine the conditional expectation and normalising gives the probability of inclusion

$$(\Upsilon_j)^{(1)} = \Bigg[\exp \Bigg\{ \frac{\log\left(\sigma^{-2}_{\Omega,s}\right)}{2} + (\log(1-\kappa))^{(1)} - (\log \kappa)^{(1)} + \frac{\mathbb{E}_q(\log \Delta_j | \Upsilon_j = 1)}{2} + \log \Gamma(a_\Delta)$$
$$- \frac{1}{2}\mu^2_{\Omega,j}\sigma^{-2}_{\Omega,j} - a_\Delta \log(b_\Delta) + (a_\Delta + 1)\mathbb{E}_q(\log \Delta_j | \Upsilon_j = 1) + b_\Delta \mathbb{E}_q[\Delta_j^{-1}|\Upsilon_j = 1] \Bigg\} + 1 \Bigg]^{-1}.$$

The univariate approximation of $q(\boldsymbol{\xi}, \boldsymbol{\psi}|\boldsymbol{y})$ (6.3.21) can be interpreted as a refinement of $(\Upsilon_j)^{(1)}$ using MCMC expectations and information on all elements of $\boldsymbol{\xi}$ to partially account for the constraint in the probability of inclusion.

The spike-and-slab form of the pseudo update for $q(\Delta_j|\Upsilon_j)$ allows us to again back out the conditioning in the conditional expectation of $\mathbb{E}_q[\Omega_j^2|\Upsilon_j]$ in $b^*_{\Delta_j}$.

$$q(\Delta_j|\Upsilon_j = 1, \mathbf{y}) = IG\left(\Delta_j \middle| \frac{1}{2} + a_\psi, \frac{(\sigma^2_{\Omega,j} + \mu^2_{\Omega,j})}{2} + b_\psi\right), \quad q(\Delta_j|\Upsilon_j = 0, \mathbf{y}) = \delta_0(\Delta_j)$$

As the update Δ_j is conditional on Υ_j, the free parameters in the proposal distributions are not a function of shrunken estimates. The $q(\Delta_j|\Upsilon_j, \mathbf{y})$ auxiliary approximating density is then used to propose scale parameters with the appropriate support, which are informed by the data, for $\boldsymbol{\psi}_\xi$ in the MCMC move.

6.3.7 RJMCMC moves and model proposals

This section explains the RJMCMC moves in detail. In the RJMCMC the proposal for $\psi_j|\xi_j = 1$ is from the q approximating density of the auxiliary parameter Ω_j, where the free parameters are obtained from the pseudo updates. As $q(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{\xi})$ is available in closed form, we are able to sample directly from it. Since the proposals do not depend on their current values, this leads to a reverse move which is a random function and thus a Jacobian which is equal to 1.

The RJMCMC involves the following steps:

- Select a birth-death or swap move with probability ϕ, $1 - \phi$.
- Propose a new model $\boldsymbol{\xi}'$ with probability $j(\boldsymbol{\xi}, \boldsymbol{\xi}')$ explained in the next section.
- Generate $\mathbf{u}$ from our proposal density $g(\mathbf{u}|a^*_\Delta, b^*_\Delta, \boldsymbol{\xi}', \boldsymbol{\psi}') \sim q(\boldsymbol{\theta}'|\boldsymbol{\psi}', \boldsymbol{\xi}') \prod_j \pi(\boldsymbol{\psi}'_j|a^*_{\Delta_j}, b^*_{\Delta_j}, \boldsymbol{\xi}')$.
- Set $(\boldsymbol{\theta}'_{(\xi',\psi')}, \boldsymbol{\psi}'_{\xi'}, \mathbf{u}') = h(\boldsymbol{\theta}_{(\xi,\psi)}, \boldsymbol{\psi}_\xi, \mathbf{u})$ where h is a specified invertible mapping function.
- Accept the proposed move to model $\boldsymbol{\xi}'$ with probability

$$\alpha_b = \min\left\{1, \frac{\Big[q(\boldsymbol{\theta}'|\mathbf{y}, \boldsymbol{\xi}', \boldsymbol{\psi}')q(\boldsymbol{\psi}', \boldsymbol{\xi}'|\mathbf{y})\Big] j_m(\boldsymbol{\xi}', \boldsymbol{\xi}) g'(\mathbf{u}'|a^*_\Delta, b^*_\Delta, \boldsymbol{\xi}, \boldsymbol{\psi})}{\Big[q(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\xi}, \boldsymbol{\psi})q(\boldsymbol{\psi}, \boldsymbol{\xi}|\mathbf{y})\Big] j_m(\boldsymbol{\xi}, \boldsymbol{\xi}') g(\mathbf{u}|a^*_\Delta, b^*_\Delta, \boldsymbol{\xi}', \boldsymbol{\psi}')} \left|\frac{\partial h(\boldsymbol{\theta}_{(\xi,\psi)}, \boldsymbol{\psi}_\xi, \mathbf{u})}{\partial(\boldsymbol{\theta}_{(\xi,\psi)}, \boldsymbol{\psi}_\xi, \mathbf{u})}\right|\right\}. \tag{6.3.27}$$

  where the target is in the square parenthesis.

The acceptance probability for the RJMCMC between-model move, as the Jacobian is equal to 1, simplifies to

$$\alpha_b = \min\left\{1, \frac{q(\boldsymbol{\xi}', \boldsymbol{\psi}'|\mathbf{y}) j_m(\boldsymbol{\xi}', \boldsymbol{\xi}) \pi(\boldsymbol{\psi}|\boldsymbol{\xi})}{q(\boldsymbol{\xi}, \boldsymbol{\psi}|\mathbf{y}) j_m(\boldsymbol{\xi}, \boldsymbol{\xi}') \pi(\boldsymbol{\psi}'|\boldsymbol{\xi}')}\right\} \tag{6.3.28}$$

where $j_m(\boldsymbol{\xi}, \boldsymbol{\xi}')$ is the proposal probability for the latent variable selection parameter $\boldsymbol{\xi}'$ (which

depends on the move type and the data) and

$$\begin{aligned}\log q(\boldsymbol{\psi},\boldsymbol{\xi}|\mathbf{y},.) \propto & \frac{1}{2}\boldsymbol{\mu}^T_{\theta_{(\xi,\psi)}}\mathbf{T}_\xi(\mathbf{T}_\xi^T\Sigma_{\theta_{(\xi,\psi)}}\mathbf{T}_\xi)^+\mathbf{T}_\xi\boldsymbol{\mu}_{\theta_{(\xi,\psi)}} + \frac{1}{2}\log\Big(\det{}^*(\mathbf{T}_\xi\Sigma_{\theta_{(\xi,\psi)}}\mathbf{T}_\xi)\Big)+ \\ & - \frac{1}{2}\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)) + \sum_j \xi_j(\log\kappa)^{(1)} + \sum_j(1-\xi_j)(\log(1-\kappa))^{(1)}+ \\ & - (a_\psi+1)\sum_j \xi_j\log(\psi_j) - b_\psi\sum_j\xi_j\psi_j^{-1} + (a_\psi\log(b_\psi) - \log(\Gamma(a_\psi))\sum_j\xi_j.\end{aligned} \tag{6.3.29}$$

As described in the main paper, a univariate approximation is used to calculate $j(\boldsymbol{\xi},\boldsymbol{\xi}')$ in the birth-death or swap move of the RJMCMC.

Birth-death and swap moves

To guide the RJMCMC over a large binary space, we use a univariate approximation $\tilde{p}(\xi_j = 1|\boldsymbol{\vartheta})$ of the joint approximating density $q(\boldsymbol{\psi},\boldsymbol{\xi})$ relative to the jth element. The probability of a new model $j_m(\boldsymbol{\xi},\boldsymbol{\xi}')$ is a function of this approximation and the move type.

Each time a variable is selected for (or removed from) the model, the remaining approximate probabilities proposal for all elements outside of the model must be renormalised. The normalised probabilities for a variable h to be selected for the model, the birth move is

$$b_h(\boldsymbol{\vartheta}) = \frac{\tilde{p}_h(\xi_h = 1|\boldsymbol{\vartheta})}{\sum_{j\notin\mathcal{M}}\tilde{p}_j(\xi_j=1|\boldsymbol{\vartheta})}, \tag{6.3.30}$$

where any $\tilde{p}(\xi_j = 1|\boldsymbol{\vartheta})$ below a small threshold ε_b (set at 1×10^{-30}) is replaced by ε_b to avoid zero probabilities. The normalised probabilities to remove a variable h from the model $\mathcal{M}$, the death move is

$$d_h(\boldsymbol{\vartheta}) = \frac{1-\tilde{p}_h(\xi_h=1|\boldsymbol{\vartheta})+\varepsilon_d}{\sum_{j\in\mathcal{M}}(1-\tilde{p}_j(\xi_j=1|\boldsymbol{\vartheta})+\varepsilon_d)} \tag{6.3.31}$$

as we select the variables to remove with probability inversely proportional to the approximate probability of inclusion. ε_d guarantees that the probabilities are comparable when they are close to the limit of their domain. The difference between the groups is relative to the size of ε_d.

If i is the current iteration, define $\sum_j (\xi_j)^{[i]} = (d_\xi)^{[i]}$ the size of the current model in the MCMC, the proposal is generated in the following way:

Sample (birth-death) and swap with probability ϕ and $1-\phi$ respectively if $2 \leq (d_\xi)^{[i]} < d$:

- (Birth-Death) Sample uniformly birth or death:
  - (Birth): If $(d_\xi)^{[i]} = 0$ add 2 variables else add 1.

$$(d_\xi)^{[i]} \neq 0 \text{ (Birth)}: \quad \frac{j_m(\boldsymbol{\xi}', \boldsymbol{\xi})}{j_m(\boldsymbol{\xi}, \boldsymbol{\xi}')} = \frac{\phi(0.5) d(\boldsymbol{\vartheta})}{\phi(0.5) b(\boldsymbol{\vartheta})} \tag{6.3.32}$$

$$(d_\xi)^{[i]} = 0 \text{ (Birth)}: \quad \frac{j_m(\boldsymbol{\xi}', \boldsymbol{\xi})}{j_m(\boldsymbol{\xi}, \boldsymbol{\xi}')} = \frac{\phi(0.5)}{\phi(0.5)(b_{(h)}(\boldsymbol{\vartheta}) b_{(l)}(\boldsymbol{\vartheta}) + b_{(l)}(\boldsymbol{\vartheta}) b_{(h)}(\boldsymbol{\vartheta}))} \tag{6.3.33}$$

  - (Death): If $d_\xi = 2$ remove 2 variables else remove 1.

$$(d_\xi)^{[i]} = 2 \text{ (Death)}: \quad \frac{j_m(\boldsymbol{\xi}', \boldsymbol{\xi})}{j_m(\boldsymbol{\xi}, \boldsymbol{\xi}')} = \frac{\phi(0.5)(b_{(h)}(\boldsymbol{\vartheta}) b_{(l)}(\boldsymbol{\vartheta}) + b_{(l)}(\boldsymbol{\vartheta}) b_{(h)}(\boldsymbol{\vartheta}))}{\phi(0.5)} \tag{6.3.34}$$

$$(d_\xi)^{[i]} \notin \{0, 2\} \text{ (Death)}: \quad \frac{j_m(\boldsymbol{\xi}', \boldsymbol{\xi})}{j_m(\boldsymbol{\xi}, \boldsymbol{\xi}')} = \frac{\phi(0.5) b(\boldsymbol{\vartheta})}{\phi(0.5) d(\boldsymbol{\vartheta})} \tag{6.3.35}$$

When we add two elements h and l the order is not important. As the probability of selecting each element is not the same, we have to add the probabilities so that

$$b_{(h)}(\boldsymbol{\vartheta}) b_{(l)}(\boldsymbol{\vartheta}) + b_{(l)}(\boldsymbol{\vartheta}) b_{(h)}(\boldsymbol{\vartheta}) \tag{6.3.36}$$

is the probability of choosing element h first and element l second plus the probability of choosing element l first and element h second (the order is in the bracket).

- (Swap):
  - Sample a variable included in the model h and swap with one outside l.

$$\text{(Swap)}: \quad \frac{j_m(\boldsymbol{\xi}', \boldsymbol{\xi})}{j_m(\boldsymbol{\xi}, \boldsymbol{\xi}')} = \frac{(1-\phi) d_l(\boldsymbol{\vartheta}) b_h(\boldsymbol{\vartheta})}{(1-\phi) d_h(\boldsymbol{\vartheta}) b_l(\boldsymbol{\vartheta})}. \tag{6.3.37}$$

Within-model moves

Within-model samples are included so that both $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ are sampled sufficiently. This enables the calculation of q expectations within the ELBO and the free parameter updates for $q(\sigma^2)$. Its is particularly important when estimating $||u||^{(2)}$ as the calculation has to be split into its component parts, because the latent variables which perform variable selection need to be incorporated for the expectations. If $\boldsymbol{\theta}|\boldsymbol{\xi}, \boldsymbol{\psi}$ has not been sampled sufficiently to estimate $\mathbb{E}_q[\boldsymbol{\theta}_\xi^T \boldsymbol{Z}_\xi^T \boldsymbol{Z}_\xi \boldsymbol{\theta}_\xi]$, then the cross product terms may not be sufficiently large enough to prevent the dot product from having a negative value.

The within-model move is performed after a successful between-model move and for a random subset of the total number of iterations. Conditional on $\boldsymbol{\xi}$, propose ψ_j for each j element in the model

$$\pi(\psi_j|\xi_j = 1) \sim IG(\psi_j|a^*_{\Delta_j}, b^*_{\Delta_j}) \tag{6.3.38}$$

and then propose the vector θ directly from the target distribution

$$\pi(\boldsymbol{\theta}|\boldsymbol{\xi} \in \{1\}, \psi) \sim SMVN_{d_\xi}(\boldsymbol{\theta}_\xi|\boldsymbol{\mu}_{\theta_{(\xi,\psi)}}, \Sigma_{\theta_{(\xi,\psi)}}). \tag{6.3.39}$$

The acceptance probability simplifies to

$$\alpha_w = \min\left\{1, \frac{q(\boldsymbol{\psi}'|\mathbf{y}, \boldsymbol{\xi})\pi(\boldsymbol{\psi}|\boldsymbol{\xi})}{q(\boldsymbol{\psi}|\mathbf{y}, \boldsymbol{\xi})\pi(\boldsymbol{\psi}'|\boldsymbol{\xi})}\right\} \tag{6.3.40}$$

where $\log q(\boldsymbol{\psi}|\boldsymbol{\xi}, \mathbf{y})$ is proportional to (6.3.29).

Algorithm

CAVI is performed by iterating through the analytical variational updates, maximising the evidence lower bound (ELBO) with respect to each coordinate direction whilst fixing the other coordinate values. For the $q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi})$ block an MCMC is implemented to obtain Monte Carlo estimates of the intractable marginal expectations of the approximating densities. The proposal

probabilities for the sampling scheme are a function of the data and the free parameters, and are updated at each iteration of the CAVI.

For each run we compute the ELBO (Section 1 of the Supplementary Material), with the updated free parameters, until this converges to the local optimum. The ELBO is no longer monotonically increasing because of the Monte Carlo variability, but we are able to declare convergence when the random fluctuations are small around a fixed point. The implementation of the overall approach is described in Algorithm 5, with the MCMC move detailed in 6.

It is computationally inefficient to start with a large number of iterations m, when the current variational distribution can be far from the maximiser. The software allows the user to specify a smaller number of iterations to begin with before increasing the number of iterations as the algorithm becomes more stable, improving the accuracy of the Monte Carlo estimates.

Algorithm 5: MC - CAVI for variable selection.

Input : A model $p(\mathbf{y}, \boldsymbol{\vartheta})$, a data set $\mathbf{y}$. Number of Monte Carlo samples m.
Output : Variational densities $q(\boldsymbol{\vartheta}_{-(\theta,\psi,\xi)}) = \prod_v q_v(\vartheta_v)$ and Monte Carlo expectations.
Intialize: First and second order raw moments of the variational factors, prior hyperparameters.

for *k = 1,..,K* **do**

  for *v = 1,...,V* **do**

    Set $q_v(\vartheta_v) \propto \exp\{\mathbb{E}_{-v}[\log p(\vartheta_v|\boldsymbol{\vartheta}_{-v}, \mathbf{y})]\}$

  end

  Calculate the arguments for proposal distribution for $\boldsymbol{\psi}$ from the psuedo variational updates.

$$a^*_{\Delta_j} = \frac{1}{2} + a_\Delta \qquad b^*_{\Delta_j} = \frac{1}{2}(\mu^2_{\Omega_j} + \sigma^2_{\Omega_j}) + b_\Delta$$
$$\psi_j \sim IG(a^*_{\Delta_j}, b^*_{\Delta_j})$$

  Calculate the probabilites $\tilde{p}(\boldsymbol{\xi}|\boldsymbol{\vartheta})$ for the $\boldsymbol{\xi}$ proposal (by approximating $q(\boldsymbol{\xi}|\boldsymbol{y})$ and normalising) in the RJMCMC.

$$\tilde{p}(\xi_j = 1|\boldsymbol{\vartheta}) \equiv \Bigg[\exp\Bigg\{ (\log(1-\kappa))^{(1)} - \frac{1}{2}\log\big(\bar{\sigma}^2_{\theta,j}\big) + \frac{1}{2}(\log\psi_j)^{\{1\}}_{\not\emptyset} - (\log\kappa)^{(1)} +$$
$$+ (\log\Gamma(a_\psi) - a_\psi \log b_\psi) + (a_\psi + 1)(\log\psi_j)^{\{1\}}_{\not\emptyset} + b_\psi(\psi_j^{-1})^{\{1\}}_{\not\emptyset} \Bigg\} + 1 +$$
$$- \frac{1}{2\bar{\sigma}^2_{\theta,j}} \left((1 - 1/\{d_\xi\}^{\{1\}})(\{\mu_{\theta_j}\}^{\{1\}}_{\not\emptyset})^2 - \frac{2}{\{d_\xi\}^{\{1\}}} \{\mu_{\theta_{\xi_j}}\}^{\{1\}}_{\not\emptyset} \sum_{j' \neq j} \{\mu_{\theta_{\xi_{j'}}}\}^{\{1\}} \right) \Bigg]^{-1}$$

  Perform MCMC step Algorithm:

  return $\mathbb{E}_q(\boldsymbol{\xi}|\mathbf{y})^{[k]}$, $\mathbb{E}_q(\boldsymbol{\psi}|\mathbf{y})^{[k]}$, $\mathbb{E}_q(\boldsymbol{\theta}|\mathbf{y})^{[k]}$, $\mathbb{E}_q(\boldsymbol{\theta}_\xi^T \boldsymbol{Z}_\xi^T \boldsymbol{Z}_\xi \boldsymbol{\theta}_\xi|\mathbf{y})^{[k]}$ *and cross product terms in the ELBO calculation*

  Compute ELBO.

end

return $q(\boldsymbol{\vartheta}_{-(\theta,\psi,\xi)})$, $\mathbb{E}_q(\boldsymbol{\xi}|\mathbf{y})$, $\mathbb{E}_q(\boldsymbol{\psi}|\mathbf{y})$, $\mathbb{E}_q(\boldsymbol{\theta}|\mathbf{y})$.

Algorithm 6: MCMC step for CAVI-MC.

Input: k current loop of CAVI-MC, q expectations, pseudo VB updates, $\tilde{q}(\boldsymbol{\xi}|\boldsymbol{\vartheta})$.

for *i = 1,...,m* **do**

if *Between-model move proposed* **then**

Given the current position of the variational samples $\boldsymbol{\xi}$, $\boldsymbol{\psi}_\xi$ and $\boldsymbol{\theta}_{(\psi,\xi)}$, propose either a birth-death move or swap move. Propose a new model with probability

$$j_m(\boldsymbol{\xi}, \boldsymbol{\xi}') \propto \tilde{p}(\boldsymbol{\xi}|(\kappa)^{(1)}, (\boldsymbol{\mu}_{\theta_\xi})_{\not\emptyset}^{\{1\}[k-1]}, (\boldsymbol{\psi}^{-1})^{\{1\}[k-1]}, (\log \boldsymbol{\psi})^{\{1\}[k-1]}, (\bar{\sigma}_\theta^2), (d_\xi)^{\{1\}[k-1]})$$

Draw $\boldsymbol{\psi}'$ proposals for all the nonzero elements in $\boldsymbol{\xi}'$ with probability

$$\pi(\boldsymbol{\psi}'|\boldsymbol{\xi}', a^*_{\Delta_j}, b^*_{\Delta_j}) = \prod_j \left[IG\Big(\psi_j|\frac{1}{2} + a_\Delta, \frac{1}{2}(\mu^2_{\Omega_j} + \sigma^2_{\Omega_j}) + b_\Delta\Big) \right]^{\xi'_j}$$

Draw the $\boldsymbol{\theta}'$ proposal

$$\boldsymbol{\mu}'_{\theta_{(\xi,\psi)}} = \Sigma_{\theta_{(\xi,\psi)}} (\sigma^{-2})^{(1)} \boldsymbol{Z}_\xi^T (\mathbf{u}_{\not j})^{(1)} \qquad \Sigma'_{\theta_{(\xi,\psi)}} = \Big((\mathbf{T}_\xi \mathrm{diag}(\boldsymbol{\psi}'_\xi)\mathbf{T}_\xi)^+ + (\sigma^{-2})^{(1)} \boldsymbol{Z}_\xi^T \boldsymbol{Z}_\xi \Big)^{-1}$$
$$\boldsymbol{\theta}'_{(\psi,\xi)} \sim \mathrm{SMVN}_{d'_\xi}((\mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi})', (\mathbf{T}_\xi \Sigma_{\theta_\xi} \mathbf{T}_\xi)' | \boldsymbol{\psi}', \boldsymbol{\xi}', \boldsymbol{Z}, (\mathbf{u}_{\not j})^{(1)}, (\sigma^{-2})^{(1)})$$

The acceptance probability is

$$\alpha_b = \min\left\{ \frac{q(\boldsymbol{\psi}', \boldsymbol{\xi}'|\mathbf{y}) j_m(\boldsymbol{\xi}', \boldsymbol{\xi}) \pi(\boldsymbol{\psi}|\boldsymbol{\xi}, , a^*_\Delta, b^*_\Delta)}{q(\boldsymbol{\psi}, \boldsymbol{\xi}|\mathbf{y}) j_m(\boldsymbol{\xi}, \boldsymbol{\xi}') \pi(\boldsymbol{\psi}'|\boldsymbol{\xi}', a^*_\Delta, b^*_\Delta)}, 1 \right\}$$

with the target density simplifed to:

$$\begin{aligned} q(\boldsymbol{\xi}, \boldsymbol{\psi}|\mathbf{y}) \propto \exp\Bigg(& \frac{1}{2}(\boldsymbol{\mu}^T_{\theta_{(\xi,\psi)}} \mathbf{T}_\xi (\mathbf{T}_\xi^T \Sigma_{\theta_{(\xi,\psi)}} \mathbf{T}_\xi)^+ \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_{(\xi,\psi)}}) + \frac{1}{2}\log\big(\det{}^*(\mathbf{T}_\xi \Sigma_{\theta_{(\xi,\psi)}} \mathbf{T}_\xi)\big) + \\ & - \frac{1}{2}\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)) + \sum_j (1-\xi_j)(\log(1-\kappa))^{(1)} - (a_\psi + 1)\sum_j \xi_j \log(\psi_j) + \\ & + \sum_j \xi_j (\log\kappa)^{(1)} - b_\psi \sum_j \xi_j \psi_j^{-1} + (a_\psi \log(b_\psi) - \log(\Gamma(a_\psi)) \sum_j \xi_j \Bigg) \end{aligned}$$

for *l=1,...,L* **do**

Perform within-model moves: Given the current position of the variational samples $\boldsymbol{\xi}$, $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ draw proposals $\boldsymbol{\psi}'|\boldsymbol{\xi}$ and $\boldsymbol{\theta}'|\boldsymbol{\psi}', \boldsymbol{\xi}$ using the same distributions as the between-model move. Proposed moved accepted with probability

$$\alpha_w = \min\left\{ \frac{q(\boldsymbol{\psi}', \boldsymbol{\xi}|\mathbf{y}) \pi(\boldsymbol{\psi}|\boldsymbol{\xi}, a^*_\Delta, b^*_\Delta)}{q(\boldsymbol{\psi}, \boldsymbol{\xi}|\mathbf{y}) \pi(\boldsymbol{\psi}'|\boldsymbol{\xi}, a^*_\Delta, b^*_\Delta)}, 1 \right\}.$$

end

else

for *l=1,...,L* **do**

Perform within-model moves with probability α_w.

end

end

end

6.4 Simulation Study

We validate the performance of our variational inference model against two frequentist variable selection approaches, ordinary least squares (OLS) (when $n >> p$) and group lasso regression which have software freely available on CRAN (R, 2017). Importantly, both of these approaches ignore the sum to zero constraint on the associated vector of parameters $\boldsymbol{\theta}$ after the columns of the compositional design matrix $\boldsymbol{Q}$ have been logged.

We generate the covariate data using an approach which is similar to Lin et al. (2014). An $n \times d$ data matrix $\mathbf{O} = (o_{ij})$ is drawn from a multivariate normal distribution $N_p(\boldsymbol{\mu}_o, \Sigma_o)$, and then the compositional covariate matrix $\mathbf{Q} = (q_{ij})$ is obtained via the transformation $q_{ij} = \exp(\tau o_{ij}) / \sum_{k=1}^{d} \exp(\tau o_{ik})$. The covariates thus follow a logistic normal distribution (Aitchison and Shen, 1980). To account for the differences in the order of magnitudes of the components, we fix $\tau = 2$ and let $\mu_{oj} = \log(d \times 0.5)$ for j = 1,...,5 and $\mu_{oj} = 0$ otherwise. As the correlations between the abundances of features in the microbiome can vary quite considerably according to the taxonomy class, we choose three settings for Σ_o: $\Sigma_o = \mathbf{I}$, $(\rho^{|i-j|})$ with $\rho = 0.2$ or 0.5. We vary the number of compositional features from 45 $(n = 100, d = 45)$ to 100 $(n = 100, d = 100)$ and $(n = 200, d = 100)$, but keep the total number of continuous covariates $p = 20$ and categorical covariates $G = 4$ with associated levels (3, 5, 5 and 5) fixed. Two scenarios are simulated from model (6.3.6), non-zero $\boldsymbol{\theta}$ elements only with $\boldsymbol{\theta} = (1, -1.3, 0.7, 0, 0, -1, 1.3, -0.7, 0, 0, ..., 0)$ ("simple" scenario) and additional non-zero elements of $\boldsymbol{\beta} = (1, -0.8, 0.6, -1.5, 0, 0, ..., 0)$ and the second categorical covariate with reference to the intercept $\boldsymbol{\zeta} = (1, -0.8, 0.6, 1.5)$ as the first level is included in the intercept ("mixed" scenario).

Fast OLS backward selection via Akaike information criterion is performed using "fastbw" (Harrell, 2021), where factors rather than columns are removed from the design matrix. A complete model is fitted and the approximate Wald statistics are computed via restricted maximum likelihood, assuming multivariate normality of estimates. The regularisation paths of the group lasso penalised learning for a sequence of regularization parameters are fitted by "gglaso" (Yang et al., 2020). Group lasso is used so that selection, as in the OLS approach, is performed on the cate-

gorical group rather than the individual levels within the factor. The penalty parameter selection is performed using cross validation over a grid of values and the mean squared error loss function. For the CAVI-MC model, vague priors are placed on the hyperparameters and initial q expectations are randomly sampled from the prior distributions. 30 variational inference iterations are performed (although the algorithm typically converges after approximately 8 iterations) for each run. The initial number of between-model MCMC iterations is set to 5000, before 10000 iterations are performed after the 5th set of variational inference updates.

We define the signal to noise ratio (SNR) as SNR = mean $|\boldsymbol{\beta}_\gamma + \boldsymbol{\zeta}_\chi + \boldsymbol{\theta}_\xi|/\sigma$. To generate the data with SNR of 0.5, 1 and 5 the SNR expression is solved for σ and 100 simulations for each setting are preformed. To assess the performance of the approaches we use metrics which evaluate the ability to select the correct variables and estimate the appropriate effects. We compute the l_2 loss $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} + \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} + \hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}||_2$ to assess the accuracy of the coefficient estimates, where the approximate posterior mean is used for the parameter estimate of the Bayesian model. To asses the accuracy of the variable selection, the true positive rate (TPR or sensitivity) and false positive rate (FPR or 1 - specificity) is reported, where positives and negatives in the context of the frequentist approaches refer to non-zero and zero coefficients respectively. Variable selection is performed by thresholding the marginal approximate posterior distributions $\mathbb{E}[q(\gamma_j|y)]$, $\mathbb{E}[q(\chi_j|y)]$ and $\mathbb{E}[q(\xi_j|y)]$ at 0.5. When there is a mixture of different parameters in the true model, the TPR and FPR are also decomposed in to the TPR($\boldsymbol{\theta}$) and FPR($\boldsymbol{\theta}$) for the compositional covariates and TPR($\boldsymbol{\beta}, \boldsymbol{\zeta}$) and FPR($\boldsymbol{\beta}, \boldsymbol{\zeta}$) for the unconstrained covariates.

The proposed CAVI-MC method performs much better than the existing methods in terms of estimation with low to moderate dimensionality. When the signal is moderate or strong the CAVI-MC approach provides a more accurate estimation of the model, both in terms of a lower false positive rate (FPR) and L2 loss. The approach works well even in the presence of high correlation with sufficient signal. This can be seen in Table 6.1 for the “mixed" scenario with a SNR of 1, and in the full table of results in the Supplementary material.

The lasso approach fails to capture the sparsity of the true model in each of the scenarios. This

characteristic is particularly obvious when $n >> p$. In Table 6.2, where the SNR is 1, $n = 100$ and $\rho = 0$, the FPR of the compositional covariates for the group lasso is 35%. For $\rho = 0.2$, the FPR is approximately 70%. The presence of correlation between the compositional covariates appears to make this problem worse.

When the true model contains both types of covariates, the two alternative approaches which fail to account for the compositional nature are easily outperformed by the CAVI-MC. The lasso methods suffer from high FPR even when the SNR is high and the correlation is low. The OLS approach struggles to identify the correct unconstrained covariates. This maybe due to the much larger variability in the true β compared with θ, despite similar means.

Each of the methods perform poorly when the SNR is low and the correlation is high. Where as the lasso approaches are inclined to include unnecessary variables in the model (leading to a very high FPR), the OLS and the CAVI-MC tend to exclude relevant variables resulting in low TPR, whilst maintaining low FPR. This increases the l_2 loss as the non zero parameter estimates shrink to zero. High correlation tends to magnify the problems with low SNR. The between-model moves in the CAVI-MC rely on a RJMCMC which is guided by independent pseudo updates. These are analogous to the OLS regression model, which tends to drop true positive variables from the model when the signal reduces and the correlation increases. When this happens the low signal is coupled with a poor guide for searching the large binary space for ξ parameter. This may explain why in Table 6.1 for $n = 100, d = 100$, the CAVI-MC has a TPR for θ below that of the group lasso approach.

A snapshot of the failings of all three approaches is provided by the plot of the ROC curves for a SNR of 0.5 in the “simple" scenario (Figure 6.4.1) where the red and green dots and blue cross represent the TPR and FPR of the CAVI-MC, lasso and OLS approach respectively. When the correlation increases from 0.2 to 0.5, the green dot shifts to the right as the FPR increases, where as the blue cross and red drop down as the TPR decreases. The CAVI-MC outperforms the two alternative approaches easily in the first two scenarios by combination of a very high TP and very low FP. When $\rho = 0.5$ the TPR of 0.72 for the CAVI-MC is not as large as the lasso but the FPR

of 0.01 is two orders of magnitude lower than the lasso. Despite the lower TPR for $\rho = 0.5$ the parameter estimation of the CAVI-MC remains far more accurate, with a considerably lower L2 loss than the lasso.

Tables

Table 6.1: Subset of the results from the "mixed" scenario with SNR 1 for $d = 100$ compositional covariates, $G = 24$ categorical covariates, for the variational Bayes (VB) and group lasso approach. The true positive and false positive rates for the unconstrained and constrained covariates are reported alongside the L2 loss of the estimated parameters (2 decimal places).

n	ρ	Method	TPR	FPR	TPR(θ)	FPR(θ)	TPR(β, ζ)	FPR(β, ζ)	L2
100	0	VB	0.99	0.00	1.00	0.00	0.98	0.01	0.94
		GLasso	0.77	0.20	1.00	0.20	0.60	0.19	5.71
100	0.2	VB	0.99	0.00	1.00	0.00	0.98	0.01	0.99
		GLasso	0.74	0.65	0.96	0.71	0.57	0.58	2.79
100	0.5	VB	0.36	0.00	0.26	0.00	0.48	0.00	9.69
		GLasso	0.68	0.27	0.89	0.21	0.53	0.21	4.28
200	0	VB	1.00	0.00	1.00	0.00	1	0.00	0.37
		OLS	0.68	0.00	1.00	0.00	0.43	0.00	4.57
		GLasso	1.00	0.30	1.00	0.32	1.00	0.23	4.06
200	0.2	VB	1.00	0.00	1.00	0.00	1.00	0.01	0.40
		OLS	0.67	0.00	1.00	0.00	0.42	0.00	4.65
		GLasso	0.99	0.35	1.00	0.37	0.98	0.29	2.53
200	0.5	VB	1.00	0.00	1.00	0.00	1.00	0.00	0.02
		OLS	0.68	0.00	1.00	0.00	0.44	0.00	5.16
		GLasso	1.00	0.33	1.00	0.33	1.00	0.30	2.74

Table 6.2: Table of true positive rate, false positive rate and the L2 loss of the estimated parameters for the true model with elements $\boldsymbol{\theta}$ as the only significant parameter for the VB approach, OLS and group lasso for a SNR of 1. The total number of compositional, continuous and categorical covariates are represented by d, p and G respectively.

$(n, d, p+G)$	ρ	Method	TPR	FPR	L2 loss
(100, 45, 24)	0	VB	1.00	0.00	0.08
		OLS	0.94	0.08	2.32
		GLasso	0.98	0.35	3.86
(100, 45, 24)	0.2	VB	1.00	0.01	0.04
		OLS	0.97	0.16	2.13
		GLasso	0.99	0.68	3.63
(100, 45, 24)	0.5	VB	0.94	0.00	0.39
		OLS	1.00	0.16	2.41
		GLasso	1.00	0.62	3.84
(200, 100, 24)	0	VB	1.00	0.00	0.03
		OLS	0.99	0.00	0.23
		GLasso	1.00	0.22	0.16
(200, 100, 24)	0.2	VB	1.00	0.00	0.03
		OLS	1.00	0.00	0.13
		GLasso	1.00	0.15	0.13
(200, 100, 24)	0.5	VB	1.00	0.00	0.02
		OLS	1.00	0.00	0.88
		GLasso	1.00	0.23	0.25

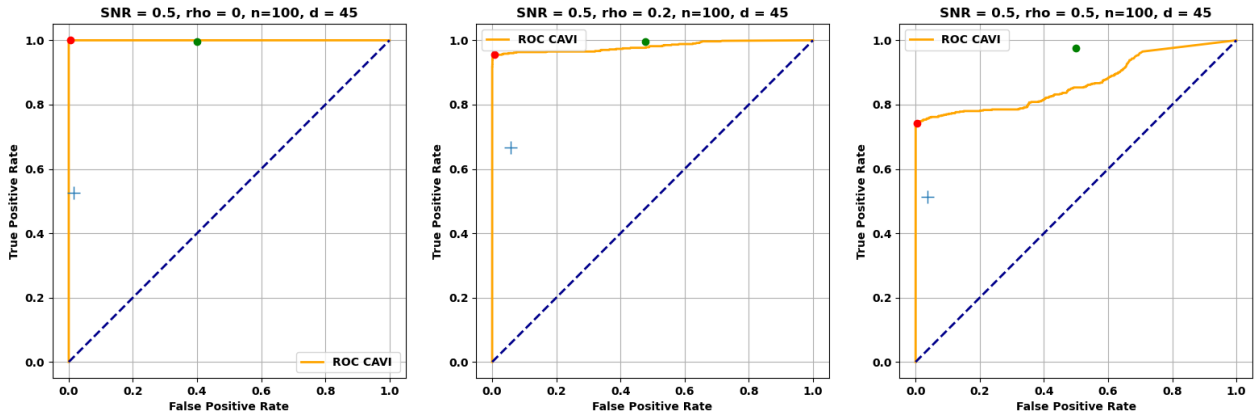

Figure 6.4.1: Plot of the ROC curves for the CAVI-MC from the "simple" scenario for a SNR of 0.5. The red and green dots and blue cross represent the TPR and FPR of the CAVI-MC, lasso and OLS respectively.

6.5 Data

We apply our proposed method to a subset of the main study in Arkhangelsk, containing 515 men and womem aged between 35-69 years recruited from the general population, from the "Know your Heart" cross-sectional study of cardiovascular disease (Cook et al., 2018). As part of the study, participants were asked to volunteer faecal samples for analysis of the gut microbiome. The relative abundances of the microbes were then determined by 16S rRNA sequencing (using the variable regions V3-V4) followed by taxonomic classification using a Naive Bayes Classifier (Bokulich et al., 2018). A baseline questionnaire captured unconstrained covariate information on age, sex and smoking status. Information on alcohol consumption from the questionnaire and biomarker data was used to derive a categorical factor with four levels on alcohol use.

The gut microbiome plays an important role in energy extraction and obesity (Tseng and Wu, 2019), which we illustrate by regressing BMI against the microbiome at the phylum and genus level alongside the unconstrained covariates. The counts are transformed into relative abundances after adding a small constant of 0.5 to replace the zero counts and then log transformed. BMI is also log transformed and the continuous age covariate is standardised. The same CAVI-MC VI set up described in the simulation study is applied to each regression model and the CAVI is

monitored to confirm convergence. Four separate CAVI-MC runs are performed at different initial starting points for the q expectations.

Thresholding the marginal expectation of the approximate posterior distributions at 0.5, we find an increase in Firmicutes (which has a -0.8 correlation with Bacteroidetes) and a decrease in Synergistetes is associated with an increase of BMI at the phylum level. At the genus level, BMI is increased by an increase in *Roseburia* and a reduction in *Oscillospira*. The corresponding marginal expectation of the approximating posterior $\mathbb{E}[q(\boldsymbol{\xi}|y)]$ is plotted in Figure 6.5.1. We also find BMI to be positively associated with age. The corresponding CAVI for each model clearly indicates an optimum has been reached (Figure 6.5.2), with each run finding the same local optimum.

Our findings appear to be consistent with previous studies. The ratio of Firmicutes to Bacteroidetes at the phylum level is considered to be a biomarker for obesity (Armougom et al. (2009), Davis (2016)). Increases in physical training of rats has led to an increase in their levels of Synergistetes (de Oliveira Neves et al., 2020). At the genus level Yuan et al. (2021) identifies *Roseburia* to be positively correlated with obesity in children, and
Chen et al. (2020) determines *Oscillospira* to be negatively associated with BMI.

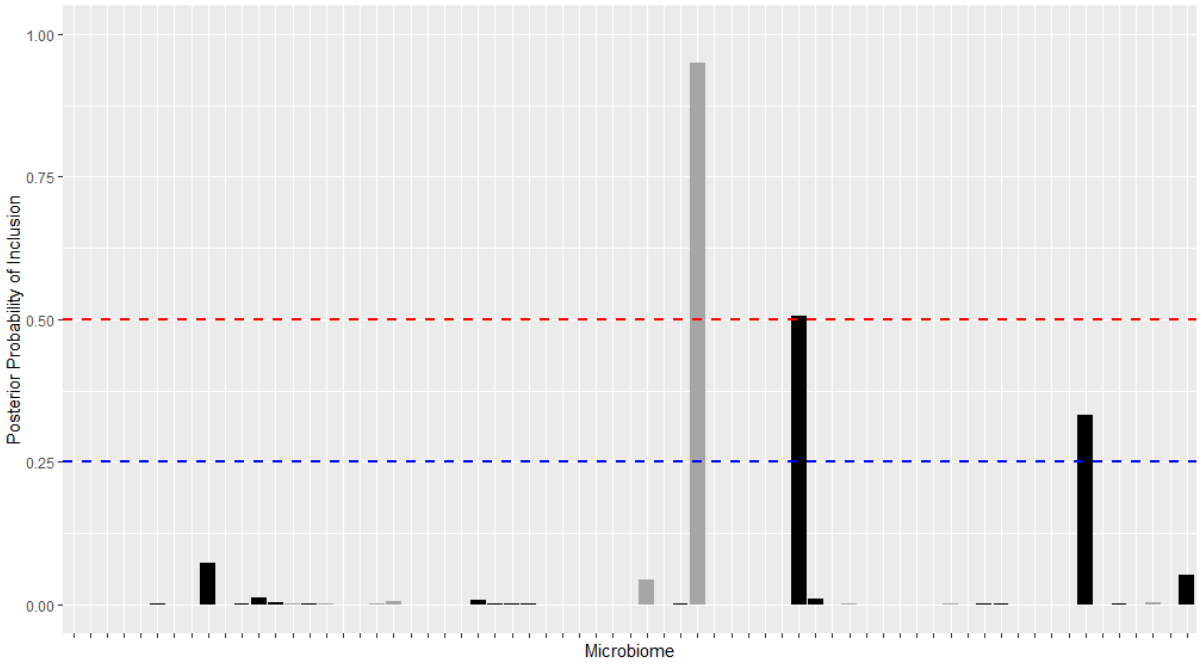

Figure 6.5.1: Plot of the marginal expectation of the approximating posterior $\mathbb{E}_q[p(\boldsymbol{\xi}|\mathbf{y})]$ at the genus level. The grey denotes a positive θ_j, black a negative θ_j. The bars above 0.25 probability of inclusion (blue dashed line) are *Roseburia*, *Oscillospira* and *Oxalobacter* respectively. The red dashed line at 0.5 probability of inclusion indicate the thresholding value used to determine a significant association.

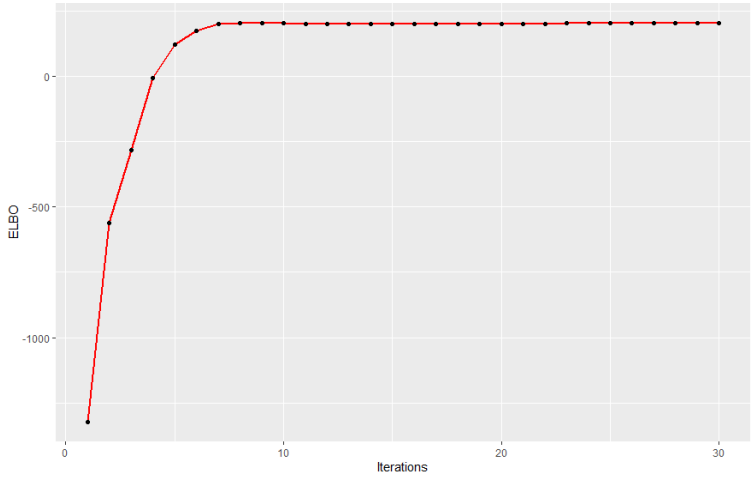

Figure 6.5.2: Plot of the ELBO against iterations for the CAVI-MC applied to the "Know Your Heart" data set with the microbiome grouped at the genus level. 30 iterations are performed, with 30,000 between state space moves by the RJMCMC after 4 iterations. The approximate straight line after only 7 iterations implies that the model has reached convergence. Despite the MCMC component removing the monotonic properties of the ELBO for a small number of iterations, this property is preserved in our case.

6.6 Discussion

Our Bayesian hierarchical linear log-contrast model estimated by mean field Monte Carlo coordinate variational inference improves regression modelling for compositional data. Sparse variable selection is performed through priors which fully account for the constrained parameter space associated with the compositional covariates. We introduce Monte Carlo expectations to approximate integrals which are not available in closed form. These expectations are obtained via RJMCMC with proposal parameters informed by approximating variational densities via auxiliary parameters with pseudo updates. As long as there is sufficient signal to guide the RJMCMC, the approach leads to an increase in the TPR and a reduction in the FPR.

The CAVI-MC suffers when the SNR is low and the correlation is high. Addressing the correlation by adapting the prior parameterisation may help to improve the model in these settings. One approach to address this issue is to use a Markov Random Field prior (Chen and Welling, 2012) which imposes a structure on the selection of ξ. Zhang et al. (2020) use this prior to incorporate the phylogenetic relationship among the bacterial taxa alongside a model which partially accounts for the constraint on the parameters. Alternatively, to avoid having to pre-define the structure of the taxa, a Dirichlet Process could be used to account for the correlation of the microbiome by clustering the covariates (Curtis and Ghosh, 2011) prior to the regression.

At the genus level, despite the CAVI-MC identifying associations between the BMI and *Roseburia* and *Oscillospira*, some of the other microbiome features which have been found to be associated with BMI were not detected. *Bifidobacterium* has been found to be negatively associated with BMI in children (Ignacio et al., 2016). This taxon was also found to be associated with BMI in adults, alongside a negative association between BMI and *Methanobrevibacter* (Schwiertz et al., 2010). However, associations between BMI and the gut microbiome at the genus level are subject to a high degree of variation across studies (Verdam et al., 2013). This maybe partly explained by the tools used to construct the microbiome datasets, which can identify quite different results from the same sample (Nearing et al., 2021).

As genetic sequencing becomes more widely available, interest grows in modelling the relationship

between the microbiome and a complex set of phenotypes such as blood concentrations of lipids or other metabolites. Bayesian Hierarchical models have been introduced for multiple outcomes (Ruffieux et al. (2017), Lewin et al. (2016)), which leverage shared information improving predictor selection. These approaches often use the simplifying assumption of conditionally independent residuals to allow different covariates to be associated with different responses. In future work, we would like to explore this multiple response extension to our model, using a hierarchical approach to allow information on the shared parameters to be pooled whilst incorporating correlation between the responses to aid variable selection.

Acknowledgments

This work was supported by the UK Medical Research Council grant MR/N013638/1 and, MR/M013138/1 "Methods and tools for structural models integrating multiple high-throughput omics data sets in genetic epidemiology". The approach is applied to data from the the Know Your Heart study, a component of International Project on Cardiovascular Disease in Russia (IPCDR) and funded by Wellcome Trust Strategic Award [100217], UiT The Arctic University of Norway (UiT), Norwegian Institute of Public Health, and Norwegian Ministry of Health and Social Affairs. The funding bodies had no role in the design of the study, data collection, analysis, interpretation of data, or in writing the manuscript. *Conflict of Interest*: None declared.

6.7 Supplementary Material

6.7.1 CAVI-MC Updates

This section contains all of the variation inference updates for the CAVI-MC.

Parameterisation

The full prior parameterisation is defined below. The likelihood and first level parameters are:

$$p(\mathbf{y}|\alpha,\boldsymbol{\beta},\boldsymbol{\zeta},\boldsymbol{\theta},\sigma^2) = \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{y}-\mathbf{1}_n\alpha-\mathbf{X}\boldsymbol{\beta}-\mathbf{W}\boldsymbol{\zeta}-\boldsymbol{Z}\boldsymbol{\theta}\right\|^2\right) \tag{6.7.1}$$

$$p(\alpha|w_\alpha) = (2\pi w_\alpha)^{-1/2}\exp\left(-\frac{1}{2w_\alpha}\alpha^2\right) \tag{6.7.2}$$

$$p(\beta_s|\gamma_s,w) = \left[(2\pi)^{-1/2}(w)^{-1/2}\exp\left\{-\frac{1}{2w}\left\|\beta_s\right\|^2\right\}\right]^{\gamma_s}\delta_0(\beta_s)^{1-\gamma_s} \quad \beta_s\in\mathbb{R}^1 \tag{6.7.3}$$

$$p(\gamma_s|\omega) = \omega^{\gamma_s}(1-\omega)^{1-\gamma_s} \quad \gamma_s\in\{0,1\} \tag{6.7.4}$$

$$p(\boldsymbol{\theta}|\boldsymbol{\xi},\boldsymbol{\psi},\mathbf{T}) = \frac{1}{\det{}^*(2\pi\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi^T)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{\theta}_\xi)^T(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi^T)^+(\boldsymbol{\theta}_\xi)\right)\delta_0(\boldsymbol{\theta}_{\bar{\xi}}) \tag{6.7.5}$$

$$p(\boldsymbol{\psi}|\boldsymbol{\xi}) = \prod_{j=1}^{d}\left[\frac{b_\psi^{a_\psi}}{\Gamma(a_\psi)}(\psi_j)^{-a_\psi-1}\exp\{-b_\psi\psi_j^{-1}\}\right]^{\xi_j}\delta_0(\psi_j)^{1-\xi_j} \quad \psi_j>0,\ \forall\ j \tag{6.7.6}$$

$$p(\boldsymbol{\zeta}_g|\chi_g,\upsilon) = \left(\frac{1}{(2\pi\upsilon)^{m_g/2}}\exp\left(-\frac{1}{2\upsilon}\boldsymbol{\zeta}_g^T\boldsymbol{\zeta}_g\right)\right)^{\chi_g}\delta_0(\boldsymbol{\zeta}_g)^{1-\chi_g} \tag{6.7.7}$$

$$p(\chi_g|\varrho) = \varrho^{\chi_g}(1-\varrho)^{1-\chi_g} \tag{6.7.8}$$

$$p(\sigma^2|\tau,\nu) = \frac{\nu^\tau}{\Gamma(\tau)}(\sigma^2)^{-\tau-1}\exp\{-\nu\sigma^{-2}\} \quad \sigma^2>0 \tag{6.7.9}$$

The hyperparameters are:

$$p(w_\alpha|a_\alpha, b_\alpha) = \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)}(w_\alpha)^{-a_\alpha-1}\exp\{-b_\alpha w_\alpha^{-1}\} \quad w > 0 \tag{6.7.10}$$

$$p(b_\alpha) = \frac{b_{b\alpha}^{a_{b\alpha}}}{\Gamma(a_{b\alpha})}(b_\alpha^{a_{b\alpha}-1})\exp\{-b_{b\alpha}b_\alpha\} \quad b_\alpha > 0 \tag{6.7.11}$$

$$p(\omega|a_\omega, b_\omega) = \frac{1}{B(a_\omega, b_\omega)}\omega^{a_\omega-1}(1-\omega)^{b_\omega-1} \quad 0 \leq \omega \leq 1 \tag{6.7.12}$$

$$p(w|a_w, b_w) = \frac{b_w^{a_w}}{\Gamma(a_w)}(w)^{-a_w-1}\exp\{-b_w w^{-1}\} \quad w > 0 \tag{6.7.13}$$

$$p(b_w) = \frac{b_b^{a_b}}{\Gamma(a_b)}(b_w^{a_b-1})\exp\{-b_b b_w\} \quad b_w > 0 \tag{6.7.14}$$

$$p(\nu) = \frac{b_\nu^{a_\nu}}{\Gamma(a_\nu)}(\nu^{a_\nu-1})\exp\{-\nu b_\nu\} \tag{6.7.15}$$

$$p(\boldsymbol{\xi}) \propto \prod_{j=1}^{d}\kappa^{\xi_j}(1-\kappa)^{1-\xi_j}\mathbf{I}\left[\sum_j \xi_j \neq 1\right] \tag{6.7.16}$$

$$p(\kappa) = \frac{1}{B(a_\kappa, b_\kappa)}\kappa^{a_\kappa-1}(1-\kappa)^{b_\kappa-1} \quad 0 \leq \kappa \leq 1 \tag{6.7.17}$$

$$p(\varrho) = \frac{1}{B(a_\varrho, b_\varrho)}\varrho^{a_\varrho-1}(1-\varrho)^{b_\varrho-1} \quad 0 \leq \varrho \leq 1 \tag{6.7.18}$$

$$p(v|a_v, b_v) = \frac{b_v^{a_v}}{\Gamma(a_v)}(w)^{-a_v-1}\exp\{-b_v v^{-1}\} \quad v > 0 \tag{6.7.19}$$

$$p(b_v) = \frac{b_{bv}^{a_{bv}}}{\Gamma(a_{bv})}(b_v^{a_{bv}-1})\exp\{-b_{bv}b_v\} \quad b_v > 0 \tag{6.7.20}$$

The prior parameterisation is defined above, where the indexes s, j, g assign unique variables per index where as α, λ, τ and b assign single parameters. The design matrix $\boldsymbol{X}$ contains the continuous covariates, $\boldsymbol{W}$ contains the categorical covariates as dummy variables with reference to an intercept and $\boldsymbol{Z}$ contains the log microbiome data.

By imposing a constraint on θ we introduce a covariance between the elements θ_j which we capture within the mean field family. The joint posterior is

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\vartheta}) =& p(\mathbf{y}|\mathbf{X}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\theta}, \sigma^2) \times \left\{\prod_s p(\beta_s|w, \gamma_s) \times \prod_s p(\gamma_s|\omega)\right\} \times \left\{\prod_g p(\boldsymbol{\zeta}_g, \chi_g) \times p(\chi_g|\varrho)\right\} \\ & \left\{p(\boldsymbol{\theta}|\Sigma(\mathbf{T}, \boldsymbol{\psi}), \boldsymbol{\xi}) \times p(\boldsymbol{\psi}|\boldsymbol{\xi}) \times p(\boldsymbol{\xi})\right\} \times p(\alpha|w_\alpha) \times p(w_\alpha|b_\alpha) \times p(b_\alpha) \\ & p(\omega) \times p(\kappa) \times p(\varrho) \times p(\sigma^2|\tau, \nu) \times p(w|b_w) \times p(b_w) \times p(\nu) \times p(v|b_v) \times p(b_v) \end{aligned}$$

Define the mean-field approximation distribution as

$$q(\boldsymbol{\vartheta}) = q(\alpha) \times \left\{ \prod_s q(\beta_s, \gamma_s) \right\} \times \left\{ \prod_g q(\boldsymbol{\zeta}_g, \chi_g) \right\} \times q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi}) \times q(\omega) \times q(\kappa) \times q(\varrho) \times$$
$$q(\sigma^2) \times q(w_\alpha) \times q(w) \times q(v) \times q(b_\alpha) \times q(b_w) \times q(b_v) \times q(\nu) \times q(\tau)$$

with $f(\boldsymbol{\vartheta})^{(j)}$ as the j-th moment of $f(\boldsymbol{\vartheta})$ with respect to $q(\boldsymbol{\vartheta})$, $\mathbb{E}_q\left[f(\boldsymbol{\vartheta})^j\right]$.

By defining a block in the mean field approximation as a multivariate density $q(\theta, \xi)$, this allows us to incorporate correlation between the elements in θ (and the corresponding elements in ξ) related to the compositional explanatory variables and the correlation between θ_j and ξ_j. Now the expectation is with respect to the vector.

CAVI updates

The CAVI update is proportional to

$$\begin{aligned}
\log q(\alpha) &\propto \mathbb{E}_{(-\alpha)}\left[\log p(\mathbf{y}|.) + \log p(\alpha|w_\alpha)\right] \\
&\propto \mathbb{E}_{(-\alpha)}\Bigg[-\frac{1}{2\sigma^2}\Big\|\mathbf{y} - \alpha\mathbf{1}_n - \sum_s X_s\gamma_s\beta_s - \sum_j Z_j\xi_j\theta_j - \sum_g W_g\chi_g\boldsymbol{\zeta}_g\Big\|^2 + \\
&\quad + \frac{1}{2}\log\big(w_\alpha^{-1}\big) - \frac{\alpha^2}{2w_\alpha}\Bigg] \\
&\propto -\frac{\alpha^2}{2(w_\alpha)^{(1)}} - \frac{1}{2(\sigma^2)^{(1)}}\Big(\alpha^2 n - 2\alpha\mathbf{1}_n^T\mathbf{y} - 2\alpha\mathbf{1}_n^T\sum_s X_s(\beta_s)^{(1)} + \\
&\quad - 2\alpha\mathbf{1}_n^T\sum_j Z_j(\theta_j)^{(1)} - 2\alpha\mathbf{1}_n^T\sum_g W_g(\boldsymbol{\zeta}_g)^{(1)}\Big)
\end{aligned}$$

By exponentiating and completing the square we have

$$q(\alpha) = N(\mu_\alpha, \sigma_\alpha^2)$$

with updates

$$\mu_\alpha = \sigma_\alpha^2\Big[(\sigma^{-2})^{(1)}\mathbf{1}_n^T\Big(\mathbf{y} - \sum_s X_s(\beta_s)^{(1)} - (\boldsymbol{Z}_\xi\boldsymbol{\theta}_\xi)^{(1)} - \sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_g)^{(1)}\Big)\Big] \tag{6.7.21}$$

$$\sigma_\alpha^2 = \Big(n(\sigma^{-2})^{(1)} + (w_\alpha^{-1})^{(1)}\Big)^{-1} \tag{6.7.22}$$

$$\begin{aligned}
\log q(\beta_s, \gamma_s) &= \mathbb{E}_{(\beta_s,\gamma_s)}\Big[\log p(\mathbf{y}|.) + \log p(\beta_s|\gamma_s, w) + \log p(\gamma_s|\omega_s)\Big] + cst \\
&= \mathbb{E}_{(\beta_s,\gamma_s)}\Big[-\frac{1}{2\sigma^2}\Big\|\mathbf{y} - \alpha\mathbf{1}_n - \sum_{k\neq s} X_k\beta_k - X_s\beta_s - \boldsymbol{Z}\boldsymbol{\theta} + \\
&\quad - \sum_g \boldsymbol{W}_g\boldsymbol{\zeta}_g\Big\|^2 - \frac{\gamma_s\beta_s^2}{2w} + \gamma_s\log(2\pi w)^{-1/2} + \gamma_s\log(\omega) + (1-\gamma_s)(\log(1-\omega))\Big] + cst
\end{aligned}$$

where cst is a constant with respect to β_s and γ_s. The spike-and-slab prior forces the latent selection variables into the likelihood component

$$\begin{aligned}
\log(\beta_s, \gamma_s) =& \mathbb{E}_{(\beta_s,\gamma_s)}\Big[-\frac{1}{2\sigma^2}\Bigg(\|X_s\|^2\gamma_s\beta_s^2 + 2X_s^T\gamma_s\beta_s\sum_{k\neq s}X_k\gamma_k\beta_k - 2X_s^T\gamma_s\beta_s\mathbf{y} + \\
&+ 2X_s^T\gamma_s\beta_s\boldsymbol{Z}_\xi\boldsymbol{\theta}_\xi + +2X_s^T\gamma_s\beta_s\mathbf{1}_n\alpha + 2\gamma_s\beta_s X_s^T\sum_g \boldsymbol{W}_g\boldsymbol{\zeta}_g\chi_g\Bigg) - \frac{\gamma_s\beta_s^2}{2w} + \\
&+ \gamma_s\log(2\pi w)^{-1/2} + \gamma_s\log(\omega) + (1-\gamma_s)(\log(1-\omega))\Big] + cst \\
\propto& -\frac{\gamma_s\beta_s^2}{2}\left(\frac{\|X_s\|^2}{(\sigma^2)^{(1)}} + \frac{1}{(w)^{(1)}}\right) + \gamma_s\beta_s\Bigg(\frac{X_s^T}{(\sigma^2)^{(1)}}\Bigg[\sum_{k\neq k}X_k(\beta_k)^{(1)} - \mathbf{y} + (\boldsymbol{Z}_\xi\boldsymbol{\theta}_\xi)^{(1)} + \\
&+ \mathbf{1}_n(\alpha)^{(1)} + \sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_g)^{(1)}\Bigg]\Bigg)\gamma_s\left(\frac{\log((w))^{(1)}}{2} + (\log\omega)^{(1)} - \frac{\log(2\pi)}{2}\right) + \\
&+ (1-\gamma_s)((\log(1-\omega))^{(1)} + \delta_0(\beta_s))
\end{aligned}$$

By exponentiating and completing the square we arrive at

$$q(\beta_s, \gamma_s|\mathbf{y}) = \left[(2\pi\sigma^2_{\beta,s})^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2_{\beta,s}} (\beta_s - \mu_{\beta,s})^2 \right\}\right]^{\gamma_s} \times \tag{6.7.23}$$
$$\times \left[\left\{\exp\left((\log w^{-1})^{(1)}\right)\sigma^2_{\beta,s}\right\}^{\frac{1}{2}} \exp\left\{\frac{1}{2}\mu_{\beta,s}\sigma^{-2}_{\beta,s}\right\} \exp\left\{(\log\omega)^{(1)}\right\}\right]^{\gamma_s} \times$$
$$\times \delta_0(\beta_s)^{1-\gamma_s} \exp\left\{(\log 1-\omega)^{(1)}\right\}^{1-\gamma_s}$$

With updates

$$\sigma^2_{\beta,s} = \left[\|X_s\|^2(\sigma^{-2})^{(1)} + (w^{-1})^{(1)}\right]^{-1} \tag{6.7.24}$$
$$\mu_{\beta,s} = \sigma^2_{\beta,s} X_s^T \left[(\sigma^{-2})^{(1)}\left(\mathbf{y} - (\alpha)^{(1)}\mathbf{1}_n - \sum_{k\neq s} X_k(\beta_k)^{(1)} - (\boldsymbol{Z}_\xi\boldsymbol{\theta}_\xi)^{(1)} - \sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_g)^{(1)}\right)\right]$$
$$= \sigma^2_{\beta,s}(\sigma^{-2})^{(1)} X_s^T(\boldsymbol{u}_{-s})^{(1)} \tag{6.7.25}$$

and thus by calling

$$(\gamma_s)^{(1)} = \left[1 + \sqrt{\sigma^{-2}_{\beta,s}}\exp\left\{(\log 1-\omega)^{(1)} - (\log\omega)^{(1)} - \frac{1}{2}(\log w^{-1})^{(1)} - \frac{1}{2}\mu^2_{\beta,s}\sigma^{-2}_{\beta,s}\right\}\right]^{-1} \tag{6.7.26}$$

we have under q

$$q(\beta_s|\gamma_s = 1, y) = \mathcal{N}(\mu_{\beta,s}, \sigma^2_{\beta,s}), \quad q(\beta_s|\gamma_s = 0, y) = \delta_0(\beta_s)$$
$$q(\gamma_s|y) \sim Bern((\gamma_s)^{(1)}).$$

Note that now

$$(\beta_s)^{(1)} = \mu_{\beta,s}(\gamma_s)^{(1)} \tag{6.7.27}$$
$$(\beta_s)^{(2)} = (\sigma^2_{\beta,s} + \mu^2_{\beta,s})(\gamma_s)^{(1)}. \tag{6.7.28}$$

The index g denotes the categorical factor groupings $g = 1, ..., G$ and m_g is the dimension of the vector $\boldsymbol{\zeta}_g$. As the categorical factors are coded with reference to the intercept, m_g is always 1 less than the levels in the categorical factor.

$$\begin{aligned}
\log q(\boldsymbol{\zeta}_g, \chi_g) =& \mathbb{E}_{(\boldsymbol{\zeta}_g, \chi_g)}\Big[\log p(\mathbf{y}|.) + \log p(\boldsymbol{\zeta}_g|\chi_g, v) + \log p(\chi_g|\varrho)\Big] + cst \\
=& \mathbb{E}_{(\zeta_g, \chi_g)}\Big[-\frac{1}{2\sigma^2}\Big\|\mathbf{y} - \alpha\mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta} - \sum_{k\neq g}\boldsymbol{W}_k\boldsymbol{\zeta}_k - \boldsymbol{W}_g\boldsymbol{\zeta}_g - \boldsymbol{Z}\boldsymbol{\theta}\Big\|^2 - \frac{\chi_g\boldsymbol{\zeta}_g^T\boldsymbol{\zeta}_g}{2v} + \\
&+ \chi_g \log(2\pi v)^{-1/2} + \chi_g\log(\varrho) + (1-\chi_g)(\log(1-\varrho))\Big] + cst
\end{aligned}$$

where cst is a constant with respect to $\boldsymbol{\zeta}_g$ and χ_g. The spike-and-slab prior forces the latent selection variables into the likelihood component

$$\begin{aligned}
\log q(\boldsymbol{\zeta}_g, \chi_g) \propto& \mathbb{E}_{(\boldsymbol{\zeta}_g, \chi_g)}\Big[-\frac{1}{2\sigma^2}\Bigg(\chi_g\boldsymbol{\zeta}_g^T\boldsymbol{W}_g^T\boldsymbol{W}_g\boldsymbol{\zeta}_g - 2\chi_g\boldsymbol{\zeta}_g^T\boldsymbol{W}_g^T(\mathbf{y} - \alpha\mathbf{1}_n - \sum_s X_s\gamma_s\beta_s - \boldsymbol{Z}_\xi\boldsymbol{\theta}_\xi + \\
&- \sum_k \boldsymbol{W}_k\boldsymbol{\zeta}_k\chi_k) - \frac{\chi_g\boldsymbol{\zeta}_g^T\boldsymbol{\zeta}_g}{2v} + \chi_g\log(2\pi v)^{-m_g/2} + \chi_g\log(\varrho) + (1-\chi_g)(\log(1-\varrho))\Big] \\
\propto& \chi_g\Bigg(-\frac{1}{2}\Bigg(\frac{1}{(v)^{(1)}}\boldsymbol{\zeta}_g^T\boldsymbol{\zeta}_g + \frac{1}{(\sigma^2)^{(1)}}\boldsymbol{\zeta}_g^T\boldsymbol{W}_g^T\boldsymbol{W}_g\boldsymbol{\zeta}_g - 2\frac{1}{(\sigma^2)^{(1)}}\boldsymbol{\zeta}_g^T\boldsymbol{W}_g^T(\boldsymbol{u}_{-g})^{(1)}\Bigg) + \\
&- \chi_g\frac{m_g}{2}(\log 2\pi)\chi_g(\log\varrho)^{(1)} + \frac{m_g}{2}(\log v^{-1})^{(1)}) + (1-\chi_g)(\log(1-\rho_g))^{(1)} + \delta_0(\boldsymbol{\zeta}_g)
\end{aligned}$$

defining

$$\Sigma_{\zeta_g} = \left[(\sigma^{-2})^{(1)}\boldsymbol{W}_g^T\boldsymbol{W}_g + (v^{-1})^{(1)}\mathbf{I}_{m_g}\right]^{-1} \tag{6.7.29}$$

$$\boldsymbol{\mu}_{\zeta_g} = (\sigma^{-2})^{(1)}\Sigma_{\zeta_g}\boldsymbol{W}_g^T(\boldsymbol{u}_{-g})^{(1)} \tag{6.7.30}$$

by exponentiating, completing the square we have

$$q(\boldsymbol{\zeta}_g, \chi_g|\mathbf{y}) = \left[\frac{1}{(2\pi)^{m_g/2}}\det(\Sigma_{\zeta_g})^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\zeta}_g - \boldsymbol{\mu}_{\zeta_g})^T\Sigma_{\zeta_g}^{-1}(\boldsymbol{\zeta}_g - \boldsymbol{\mu}_{\zeta_g})\right\}\right]^{\chi_g} \times \delta_0(\boldsymbol{\zeta}_g)^{1-\chi_g}$$
$$\left[\exp\left(\frac{1}{2}\boldsymbol{\mu}_{\zeta_g}^T\Sigma_{\zeta_g}^{-1}\boldsymbol{\mu}_{\zeta_g} + \frac{1}{2}\log\det(\Sigma_{\zeta_g}) + \frac{m_g}{2}(\log v^{-1})^{(1)} + (\log\varrho)^{(1)}\right)\right]^{\chi_g} \times$$
$$\left[\exp\big((\log(1-\varrho))^{(1)}\big)\right]^{1-\chi_g} \tag{6.7.31}$$

and thus by calling

$$(\chi_g)^{(1)} = \Big[1 + \exp\Big((\log 1 - \varrho)^{(1)} - (\log\varrho)^{(1)} - \frac{m_g}{2}(\log v^{-1})^{(1)} - \frac{1}{2}\boldsymbol{\mu}_{\zeta_g}^T\Sigma_{\zeta_g}^{-1}\boldsymbol{\mu}_{\zeta_g} +$$
$$- \frac{1}{2}\log\big(\det(\Sigma_{\zeta_g})\big)\Big)\Big]^{-1}$$

we have under q

$$q(\boldsymbol{\zeta}_g|\chi_g = 1, y) = \mathcal{N}_{m_g}(\boldsymbol{\mu}_{\zeta_g}, \Sigma_{\zeta_g}), \quad q(\boldsymbol{\zeta}_g|\chi_g = 0, y) = \delta_0(\boldsymbol{\zeta}_g)$$
$$q(\chi_g|y) \sim Bern((\chi_g)^{(1)}).$$

Note that now

$$(\boldsymbol{\zeta}_g)^{(1)} = \boldsymbol{\mu}_\zeta(\chi_g)^{(1)} \tag{6.7.32}$$
$$(\boldsymbol{\zeta}_g^T\boldsymbol{\zeta}_g)^{(1)} = (\mathrm{tr}(\Sigma_{\zeta_g}) + \boldsymbol{\mu}_{\zeta_g}^T\boldsymbol{\mu}_{\zeta_g})(\chi_g)^{(1)} \tag{6.7.33}$$
$$(\boldsymbol{\zeta}_g^T\boldsymbol{W}_g^T\boldsymbol{W}_g\boldsymbol{\zeta}_g)^{(1)} = (\mathrm{tr}(\boldsymbol{W}_g\Sigma_{\zeta_g}\boldsymbol{W}_g^T) + \boldsymbol{\mu}_{\zeta_g}^T\boldsymbol{W}_g^T\boldsymbol{W}_g\boldsymbol{\mu}_{\zeta_g})(\chi_g)^{(1)} \tag{6.7.34}$$

$$\log q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi}|\cdot) = \mathbb{E}_{(\theta,\psi,\xi)}\Big[\log p(\mathbf{y}|.) + \log p(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{\xi}) + \log p(\boldsymbol{\psi}|\boldsymbol{\xi}) + \log p(\boldsymbol{\xi})\Big] + cst$$

$$\log q(\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\xi}|\cdot) \propto \mathbb{E}_{-(\xi,\psi,\theta)}\Bigg[-\frac{1}{2}\left(\boldsymbol{\theta}_\xi^T(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)^+\boldsymbol{\theta}_\xi + \sigma^{-2}\left\|\mathbf{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}_\xi\boldsymbol{\theta}_\xi-\boldsymbol{W}\boldsymbol{\zeta}\right\|^2\right)+$$
$$-\frac{1}{2}(d_\xi-1)\log(2\pi)-\frac{1}{2}\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi))\Bigg]_{[I(\sum_j\theta_j=0)]}+$$
$$+\mathbb{E}_{-(\xi,\psi,\theta)}\Bigg[\sum_j\Big(\xi_j\log(\kappa)+(1-\xi_j)\log(1-\kappa)\Big)+\log\delta(\theta_{\bar{\xi}})+ \tag{6.7.35}$$
$$+\sum_j\xi_j(a_\psi\log(b_\psi))-\sum_j\xi_j\log(\Gamma(a_\psi))-\sum_j(a_\psi+1)\xi_j\log(\psi_j)+$$
$$-b_\psi\sum_j(1-\xi_j)\psi_j^{-1}\Bigg]$$

which we express as

$$\log p(\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\xi}|\mathbf{y},\cdot)\propto A+B \tag{6.7.36}$$

where each capital letter refers to the expression within the parenthesis of the expectations in equation (6.7.35).

$$A\propto -\frac{1}{2}(d_\xi-1)\log(2\pi)-\frac{1}{2}\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi))+$$
$$-\frac{1}{2}\Bigg(\boldsymbol{\theta}_\xi^T(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)^+\boldsymbol{\theta}_\xi+\sigma^{-2}\Big(\boldsymbol{\theta}_\xi^T\boldsymbol{Z}_\xi^T\boldsymbol{Z}_\xi\boldsymbol{\theta}_\xi-2\boldsymbol{\theta}_\xi^T\boldsymbol{Z}_\xi^T(\mathbf{y}-\alpha\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{W}\boldsymbol{\zeta})\Big)\Bigg) \tag{6.7.37}$$

define

$$\boldsymbol{u}_{\not\beta}=\mathbf{y}-\alpha\mathbf{1}_n-\sum_s X_s\gamma_s\beta_s-\sum_g\boldsymbol{W}_g\boldsymbol{\zeta}_g \tag{6.7.38}$$

and the vector $\boldsymbol{\mu}_{\theta_\xi}$ and matrix Σ_{θ_ξ}

$$\boldsymbol{\mu}_{\theta_\xi}=\Sigma_{\theta_\xi}(\sigma^{-2})^{(1)}\boldsymbol{Z}_\xi^T(\mathbf{u}_{\not\beta})^{(1)} \tag{6.7.39}$$
$$\Sigma_{\theta_\xi}=\left((\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)^+ + (\sigma^{-2})^{(1)}\boldsymbol{Z}_\xi^T\boldsymbol{Z}_\xi\right)^{-1} \tag{6.7.40}$$

Unlike in the β_s updates for the free variational parameters, these are still function of the vector $\boldsymbol{\xi}$. On completing the square we have

$$\boldsymbol{\theta}_\xi^T \Sigma_{\theta_\xi}^{-1} \boldsymbol{\theta}_\xi - 2\boldsymbol{\theta}_\xi^T (\Sigma_{\theta_\xi}^{-1}) \boldsymbol{\mu}_{\theta_\xi} = (\boldsymbol{\theta}_\xi - \boldsymbol{\mu}_{\theta_\xi})^T \Sigma_{\theta_\xi}^{-1} (\boldsymbol{\theta}_\xi - \boldsymbol{\mu}_{\theta_\xi}) - \boldsymbol{\mu}_{\theta_\xi}^T \Sigma_{\theta_\xi}^{-1} \boldsymbol{\mu}_{\theta_\xi}$$

$$\begin{aligned}
\log q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi} | \mathbf{y}, .) \propto & \Bigg[-\frac{1}{2} \log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi) \mathbf{T}_\xi)) - \frac{1}{2} \Bigg([\boldsymbol{\theta}_\xi - \boldsymbol{\mu}_{\theta_\xi}]^T \Sigma_{\theta_\xi}^{-1} [\boldsymbol{\theta}_\xi - \boldsymbol{\mu}_{\theta_\xi}] \Bigg) + \\
& - \frac{1}{2}(d_\xi - 1) \log 2\pi - \boldsymbol{\mu}_{\theta_\xi}^T \Sigma_{\theta_\xi}^{-1} \boldsymbol{\mu}_{\theta_\xi} \Bigg]_{[I(\sum_j \theta_{\xi_j} = 0)]} + \sum_j \xi_j (\log \kappa)^{(1)} + \\
& + (a_\psi \log(b_\psi) - \log(\Gamma(a_\psi)) \sum_j \xi_j - \sum_j (a_\psi + 1) \xi_j \log(\psi_j) - b_\psi \sum_j \xi_j \psi_j^{-1} \\
& + \sum_j (1 - \xi_j) (\log(1 - \kappa))^{(1)}
\end{aligned} \tag{6.7.41}$$

We can remove the index by adding the constraint on μ_{θ_ξ} and Σ_{θ_ξ} with the matrix T_ξ.

$$\begin{aligned}
\log q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi} | \boldsymbol{y}, .) \propto & -\frac{1}{2} \log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi) \mathbf{T}_\xi)) + \sum_j \xi_j (\log \kappa)^{(1)} + \sum_j (1 - \xi_j) (\log \kappa)^{(1)} + \\
& - \frac{1}{2}(d_\xi - 1) \log(2\pi) - \frac{1}{2} \Bigg([\boldsymbol{\theta}_\xi - \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi}]^T (\mathbf{T}_\xi \Sigma_{\theta_\xi} \mathbf{T}_\xi)^+ [\boldsymbol{\theta}_\xi - \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi}] \Bigg) + \\
& + \frac{1}{2} \boldsymbol{\mu}_{\theta_\xi}^T \mathbf{T}_\xi^T (\mathbf{T}_\xi^T \Sigma_{\theta_\xi} \mathbf{T}_\xi)^+ \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi} - \sum_j (a_\psi + 1) \xi_j \log(\psi_j) - b_\psi \sum_j \xi_j \psi_j^{-1} + \\
& + (a_\psi \log(b_\psi) - \log(\Gamma(a_\psi)) \sum_j \xi_j
\end{aligned} \tag{6.7.42}$$

We can then identify the singular multivariate normal density

$$\begin{aligned}\log q(\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\xi}|\boldsymbol{y},.) \propto & -\frac{1}{2}(d_\xi-1)\log(2\pi)-\frac{1}{2}\log\big(\det{}^*(\mathbf{T}_\xi\Sigma_{\theta_\xi}T_\xi)\big)+\frac{1}{2}\log\big(\det{}^*(\mathbf{T}_\xi\Sigma_{\theta_\xi}\mathbf{T}_\xi)\big)+\\ & -\frac{1}{2}\left([\boldsymbol{\theta}_\xi-\mathbf{T}_\xi\boldsymbol{\mu}_{\theta_\xi}]^T(\mathbf{T}_\xi\Sigma_{\theta_\xi}T_\xi)^+[\boldsymbol{\theta}_\xi-\mathbf{T}_\xi\boldsymbol{\mu}_{\theta_\xi}]\right)+\frac{1}{2}\boldsymbol{\mu}_{\theta_\xi}^T\mathbf{T}_\xi^T(\mathbf{T}_\xi^T\Sigma_{\theta_\xi}\mathbf{T}_\xi)^+\mathbf{T}_\xi\boldsymbol{\mu}_{\theta_\xi}+\\ & -\frac{1}{2}\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi))+\sum_j\xi_j(\log\kappa)^{(1)}+\sum_j(1-\xi_j)(\log(1-\kappa))^{(1)}+\\ & -\sum_j(a_\psi+1)\xi_j\log(\psi_j)-b_\psi\sum_j\xi_j\psi_j^{-1}+(a_\psi\log(b_\psi)-\log(\Gamma(a_\psi))\sum_j\xi_j\end{aligned}$$

which can be expressed as

$$q(\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\xi}|\boldsymbol{y},.) \propto \text{SMVN}_{d_\xi}(\mathbf{T}_\xi\boldsymbol{\mu}_{\theta_\xi},\mathbf{T}_\xi\Sigma_{\theta_\xi}\mathbf{T}_\xi)\delta(\bar{\xi})\times \tag{6.7.43}$$

$$\begin{aligned}\exp\Bigg(&\frac{1}{2}\boldsymbol{\mu}_{\theta_\xi}^T\mathbf{T}_\xi(\mathbf{T}_\xi^T\Sigma_{\theta_\xi}\mathbf{T}_\xi)^+\mathbf{T}_\xi\boldsymbol{\mu}_{\theta_\xi}+\frac{1}{2}\log\big(\det{}^*(\mathbf{T}_\xi\Sigma_{\theta_\xi}\mathbf{T}_\xi)\big)+ \qquad (6.7.44)\\ &-\frac{1}{2}\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi))+\sum_j\xi_j(\log\kappa)^{(1)}+\sum_j(1-\xi_j)(\log(1-\kappa))^{(1)}+\\ &-\sum_j(a_\psi+1)\xi_j\log(\psi_j)-b_\psi\sum_j\xi_j\psi_j^{-1}+(a_\psi\log(b_\psi)-\log(\Gamma(a_\psi))\sum_j\xi_j\Bigg)\end{aligned}$$

We can identify the singular multivariate normal density (6.7.43) which is a function of $\boldsymbol{\xi}$ and $\boldsymbol{\psi}$. The $\boldsymbol{\xi}$ and $\boldsymbol{\psi}$ component (6.7.44) contains terms which do not have a conjugate update. The first term

$$\boldsymbol{\mu}_{\theta_\xi}^T\mathbf{T}_\xi(\mathbf{T}_\xi^T\Sigma_{\theta_\xi}\mathbf{T}_\xi)^+\mathbf{T}_\xi\boldsymbol{\mu}_{\theta_\xi} \tag{6.7.45}$$

has dependencies on $\boldsymbol{\xi}$ in $\boldsymbol{\mu}_{\theta_\xi}$ and Σ_{θ_ξ} which are a function of $\boldsymbol{\psi}$ and the remaining q expectations.

Thus

$$q(\boldsymbol{\theta}_\xi|\boldsymbol{\psi},\boldsymbol{\xi}) = \text{SMVN}(\mathbf{T}_\xi\mu_{\theta_\xi},\mathbf{T}_\xi\Sigma_{\theta_\xi}\mathbf{T}_\xi) \quad \text{and} \quad q(\boldsymbol{\theta}_{\bar{\xi}}=0|\boldsymbol{\xi})=1, \tag{6.7.46}$$

or

$$q(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{\xi}) = \text{SMVN}(\mathbf{T}_\xi\mu_{\theta_\xi},\mathbf{T}_\xi\Sigma_{\theta_\xi}\mathbf{T}_\xi)\delta(\boldsymbol{\theta}_{\bar{\xi}}) \tag{6.7.47}$$

and

$$\log q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi}|\boldsymbol{y}, .) \propto \log\big(\text{SMVN}(\mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi}, \mathbf{T}_\xi \Sigma_{\theta_\xi} \mathbf{T}_\xi)\big) + \frac{1}{2}\mu_{\theta_\xi}^T \mathbf{T}_\xi (\mathbf{T}_\xi^T \Sigma_{\theta_\xi} \mathbf{T}_\xi)^+ \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi} +$$
$$\frac{1}{2}\log\big(\det^*(\mathbf{T}_\xi \Sigma_{\theta_\xi} \mathbf{T}_\xi)\big) - \frac{1}{2}\log(\det^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)) + \sum_j \xi_j (\log\kappa)^{(1)} +$$
$$- \sum_j (a_\psi + 1)\xi_j \log(\psi_j) + (a_\psi \log(b_\psi) - \log(\Gamma(a_\psi)) \sum_j \xi_j +$$
$$+ \sum_j (1 - \xi_j)(\log(1-\kappa))^{(1)} - b_\psi \sum_j \xi_j \psi_j^{-1}. \tag{6.7.48}$$

For w we have

$$\log q(w) = \mathbb{E}_{-w}\left[\sum_s \log p(\beta_s|w, \gamma_s) + \log p(w|a_w, b_w)\right] + cst$$

$$\begin{aligned} q(w) =& \mathbb{E}_{-w}\Big[\sum_s -\frac{\gamma_s}{2}\Big(\log w - w^{-1}\frac{\beta_s^2}{2}\Big)(-a_w - 1)\log w - b_w w^{-1}\Big] + cst \\ \propto& \log w\Big(-\frac{1}{2}\Big\{\sum_s (\gamma_s)^{(1)}\Big\} - a_w - 1\Big) - w^{-1}\Big(\frac{1}{2}\Big\{\sum_s (\beta_s)^{(2)}\Big\} + (b_w)^{(1)}\Big) \end{aligned} \tag{6.7.49}$$

thus

$$q(w) = Inv - Gamma(a_w^*, b_w^*) \tag{6.7.50}$$

with parameters

$$a_w^* = \frac{1}{2}\Big\{\sum_s (\gamma_s)^{(1)}\Big\} + a_w \tag{6.7.51}$$

$$b_w^* = \frac{1}{2}\Big\{\sum_s (\beta_s)^{(2)} + \Big\} + (b_w)^{(1)} \tag{6.7.52}$$

For v we have

$$\log q(v) = \mathbb{E}_{-v}\left[\sum_g \log p(\boldsymbol{\zeta}_g|v,\chi_g) + \log p(v|a_v,b_v)\right] + cst$$

$$\begin{aligned} q(v) =& \mathbb{E}_{-v}\Big[\sum_g \chi_g\Big(-\frac{m_g}{2}\log v - v^{-1}\frac{\boldsymbol{\zeta}_g^T\boldsymbol{\zeta}_g}{2}\Big) + (-a_v-1)\log v - b_v v^{-1}\Big] + cst \\ \propto& \log v\Big(-\frac{1}{2}\Big\{\sum_g m_g(\chi_g)^{(1)}\Big\} - a_v - 1\Big) - v^{-1}\Big(\frac{1}{2}\Big\{\sum_g (\chi_g\boldsymbol{\zeta}_g^T\boldsymbol{\zeta}_g)^{(1)}\Big\} + (b_v)^{(1)}\Big) \end{aligned} \tag{6.7.53}$$

thus

$$q(v) = Inv - Gamma(a_v^*, b_v^*) \tag{6.7.54}$$

with parameters

$$a_v^* = \frac{1}{2}\Big\{\sum_g m_g(\chi_g)^{(1)}\Big\} + a_v \tag{6.7.55}$$

$$b_v^* = \frac{1}{2}\Big\{\sum_g (\boldsymbol{\zeta}_g^T\boldsymbol{\zeta}_g)^{(1)}\Big\} + (b_v)^{(1)} \tag{6.7.56}$$

$$\log q(\omega) = \mathbb{E}_{-\omega}\left[\log\prod_s p(\gamma_s|\omega) + \log p(\omega)\right] + cst \tag{6.7.57}$$

$$\begin{aligned} \log q(\omega) =& \sum_s (\gamma_s)^{(1)}\log\omega + \sum_s (1-\gamma_s)^{(1)}\log(1-\omega) + (a_\omega - 1)\log\omega + (b_\omega - 1)\log(1-\omega) + cst \\ =& \left(a_\omega + \sum_s (\gamma_s)^{(1)} - 1\right)\log\omega + \left(b_{\omega,s} + p - \sum_s (\gamma_s)^{(1)} - 1\right)\log(1-\omega) + cst. \end{aligned}$$

which implies that

$$q(\omega) = Beta(a_w^*, b_w^*) \tag{6.7.58}$$

with parameters

$$a_\omega^* = a_\omega + \sum_s (\gamma_s)^{(1)} \tag{6.7.59}$$

$$b_\omega^* = b_\omega + p - \sum_s (\gamma_s)^{(1)} \tag{6.7.60}$$

where

$$\begin{aligned}
(\omega)^{(1)} &= a_\omega^* / \left(a_\omega^* + b_\omega^*\right) = a_\omega^* / \left(a_\omega + b_\omega + 1\right) \\
(\log \omega)^{(1)} &= \Psi(a_\omega^*) - \Psi(a_\omega^* + b_\omega^*) \\
(\log(1-\omega))^{(1)} &= \Psi(b_\omega^*) - \Psi(a_\omega^* + b_\omega^*)
\end{aligned} \tag{6.7.61}$$

where $\Psi(\cdot)$ is the digamma function.

$$\log q(w_\alpha) = \mathbb{E}_{-w_\alpha}\left[\log p(\alpha|, w_\alpha)) + \log p(w_\alpha|a_\alpha, b_\alpha)\right] + cst \tag{6.7.62}$$

$$\begin{aligned}
\log q(w_\alpha) &= \frac{1}{2}\log\left(w_\alpha^{-1}\right) - \frac{w_\alpha^{-1}}{2}\alpha^2 + (a_\alpha + 1)\log\left(w_\alpha^{-1}\right) - w_\alpha^{-1}(b_\alpha)^{(1)} + cst \\
&= \left(a_\alpha + \frac{1}{2}\right)\log\left(w_\alpha^{-1}\right) - w_\alpha^{-1}\left((b_\alpha)^{(1)} + \frac{1}{2}(\alpha)^{(2)}\right) + cst
\end{aligned}$$

Thus we have

$$q(w_\alpha) = Inv - Gamma(a_\alpha^*, b_\alpha^*) \tag{6.7.63}$$

with parameters

$$a_\alpha^* = a_\alpha + \frac{1}{2} \tag{6.7.64}$$

$$b_\alpha^* = (b_\alpha)^{(1)} + \frac{1}{2}(\alpha)^{(2)} \tag{6.7.65}$$

where

$$(w_\alpha^{-1})^{(1)} = a_\alpha^*/b_\alpha^* \tag{6.7.66}$$

$$(\log w_\alpha^{-1})^{(1)} = \Psi(a_\alpha^*) - \log b_\alpha^*. \tag{6.7.67}$$

$$\log q(b_w) = \mathbb{E}_{-b_w}\Big[\log p(w|a_w, b_w) + \log p(b_w|a_b, b_b)\Big] \tag{6.7.68}$$

$$\begin{aligned}\log q(b_w) &= \mathbb{E}_{-b_w}\Big[a_w \log b_w - b_w w^{-1} + (a_b - 1)\log b_w - b_b b_w\Big] + cst \\ &= a_w \log b_w - b_w (w)^{(-1)} + (a_b - 1)\log b_w - b_b b_w + cst \\ &= \log b_w (a_w + a_b - 1) - b_w((w)^{(-1)} + b_b) + cst \end{aligned} \tag{6.7.69}$$

thus

$$q(b_w) = Gamma(a_b^*, b_b^*)$$

with parameters

$$a_b^* = a_w + a_b \tag{6.7.70}$$

$$b_b^* = (w)^{(-1)} + b_b \tag{6.7.71}$$

where

$$(b_w)^{(1)} = a_b^*/b_b^* \tag{6.7.72}$$

$$(\log b_w)^{(1)} = \Psi(a_b^*) - \log b_b^* \tag{6.7.73}$$

$$\log q(b_\alpha) = \mathbb{E}_{-b_\alpha}\Big[\log p(w_\alpha|a_\alpha, b_\alpha) + \log p(b_\alpha|a_{b,\alpha}, b_{b,\alpha})\Big] \tag{6.7.74}$$

$$\begin{aligned}\log q(b_\alpha) &= \mathbb{E}_{-b_\alpha}\Big[a_\alpha \log b_\alpha - b_\alpha w_\alpha^{-1} + (a_{b,\alpha} - 1)\log b_\alpha - b_{b,\alpha} b_\alpha\Big] + cst \\ &= \log b_\alpha(a_\alpha + a_{\alpha,b} - 1) - b_\alpha((w_\alpha)^{(-1)} + b_{\alpha,b}) + cst \end{aligned} \tag{6.7.75}$$

thus

$$q(b_\alpha) = Gamma(a^*_{b,\alpha}, b^*_{b,\alpha})$$

with parameters

$$a^*_{b,\alpha} = a_\alpha + a_{b,\alpha} \tag{6.7.76}$$

$$b^*_{b,\alpha} = (w_\alpha)^{(-1)} + b_{b,\alpha} \tag{6.7.77}$$

where

$$(b_\alpha)^{(1)} = a^*_{b,\alpha}/b^*_{b,\alpha} \tag{6.7.78}$$

$$(\log b_\alpha)^{(1)} = \Psi(a^*_{b,\alpha}) - \log b^*_{b,\alpha} \tag{6.7.79}$$

$$\log q(b_v) = \mathbb{E}_{-b_v}\Big[\log p(v|a_v, b_v) + \log p(b_v|a_{bv}, b_{bv})\Big] \tag{6.7.80}$$

$$\begin{aligned}\log q(b_v) &= \mathbb{E}_{-b_v}\left[a_v \log b_v - b_v v^{-1} + (a_{bv}-1)\log b_v - b_{bv} b_v\right] + cst \\ &= a_v \log b_v - b_v (v^{-1})^{(1)} + (a_{bv}-1)\log b_v - b_{bv} b_v + cst \\ &= \log b_v (a_v + a_{bv} - 1) - b_v (v^{-1})^{(1)} + b_{bv}) + cst \end{aligned} \tag{6.7.81}$$

thus

$$q(b_v) = Gamma(a_v^*, b_v^*)$$

with parameters

$$a_{bv}^* = a_v + a_{bv} \tag{6.7.82}$$

$$b_{bv}^* = (v^{-1})^{(1)} + b_{bv} \tag{6.7.83}$$

where

$$(b_v)^{(1)} = a_{bv}^* / b_{bv}^* \tag{6.7.84}$$

$$(\log b_v)^{(1)} = \Psi(a_{bv}^*) - \log b_{bv}^* \tag{6.7.85}$$

$$\log q(\sigma^2) = \mathbb{E}_{-\sigma^2}[\log p(\sigma^2|\tau,\nu)] + \mathbb{E}_{-\sigma^2}[\log p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\zeta},\sigma^2)] + cst$$

Using $\mathbb{E}_q[\boldsymbol{Z}_\xi \boldsymbol{\theta}_\xi] = \boldsymbol{Z}(\boldsymbol{\theta})^{(1)}$ and

$$\begin{aligned}\mathbb{E}_q[\boldsymbol{\zeta}_g^T \boldsymbol{W}_g^T \boldsymbol{W}_g \boldsymbol{\zeta}_g \chi_g | \chi_g] &= \mathbb{E}_q[\boldsymbol{\zeta}_g^T \boldsymbol{W}_g^T \boldsymbol{W}_g \boldsymbol{\zeta}_g | \chi_g]\chi_g \\ &= (\mathrm{tr}(\boldsymbol{W}_g \Sigma_{\zeta_g} \boldsymbol{W}_g) + \mathbb{E}_q[\boldsymbol{\zeta}_g^T|\chi_g]\boldsymbol{W}_g^T \boldsymbol{W}_g \mathbb{E}_q[\boldsymbol{\zeta}_g|\chi_g])\chi_g\end{aligned}$$

so $\mathbb{E}_g[\boldsymbol{\zeta}_g^T \boldsymbol{W}_g^T \boldsymbol{W}_g \boldsymbol{\zeta} \chi_g]$ referred to as $(\boldsymbol{\zeta}_g^T \boldsymbol{W}_g^T \boldsymbol{W}_g \boldsymbol{\zeta})^{(1)}$ and

$$\begin{aligned}\mathbb{E}_q[\mathbb{E}_q[\boldsymbol{\zeta}_g^T \boldsymbol{W}_g^T \boldsymbol{W}_g \boldsymbol{\zeta} \chi_g | \chi_g]] &= \left(\text{tr}(\boldsymbol{W}_g \Sigma_{\zeta_g} \boldsymbol{W}_g^T) + \boldsymbol{\mu}_{\zeta_g}^T \boldsymbol{W}_g^T \boldsymbol{W}_g \boldsymbol{\mu}_{\zeta_g}\right)(\chi_g)^{(1)} \\ &= (\boldsymbol{\zeta}_g^T \boldsymbol{W}_g^T \boldsymbol{W}_g \boldsymbol{\zeta})^{(1)}\end{aligned}$$

$$\begin{aligned}||\boldsymbol{u}||^{(2)} =& ||\mathbf{y}||^2 + n(\alpha)^{(2)} + \sum_s ||X_s||^2 (\beta_s)^{(2)} + \sum_g (\boldsymbol{\zeta}_g^T \boldsymbol{W}_g^T \boldsymbol{W}_g \boldsymbol{\zeta}_g)^{(1)} + \mathbb{E}_q[\boldsymbol{\theta}_\xi^T \boldsymbol{Z}_\xi^T \boldsymbol{Z}_\xi \boldsymbol{\theta}_\xi] \\ &- 2\sum_s y^T X_s (\beta_s)^{(1)} - 2\mathbf{y}^T \boldsymbol{Z}(\boldsymbol{\theta}_\xi)^{(1)} - 2\sum_g \mathbf{y}^T \boldsymbol{W}_g (\boldsymbol{\zeta}_g)^{(1)} - 2(\alpha)^{(1)} \mathbf{1}_n^T \mathbf{y} + \\ &+ 2\sum_{s \neq s', s<s'} X_s^T X_{s'} (\beta_s)^{(1)} (\beta_{s'})^{(1)} + 2(\boldsymbol{Z}(\boldsymbol{\theta})^{(1)})^T (\sum_s X_s (\beta_s)^{(1)}) + \\ &+ 2(\boldsymbol{Z}(\boldsymbol{\theta})^{(1)})^T (\sum_g \boldsymbol{W}_g (\boldsymbol{\zeta}_g)^{(1)}) + 2\sum_{g \neq g', g<g'} (\boldsymbol{\zeta}_g)^{(1)T} \boldsymbol{W}_g^T \boldsymbol{W}_{g'} (\boldsymbol{\zeta}_{g'})^{(1)} + \\ &+ 2(\alpha)^{(1)} \mathbf{1}_n^T \sum_s X_s (\beta_s)^{(1)} + 2(\alpha)^{(1)} \mathbf{1}_n^T \boldsymbol{Z}(\boldsymbol{\theta})^{(1)} + 2(\alpha)^{(1)} \mathbf{1}_n^T \sum_g \boldsymbol{W}_g (\boldsymbol{\zeta}_g)^{(1)} \\ &+ 2\sum_s \sum_g (\beta_s)^{(1)} X_s^T \boldsymbol{W}_g (\boldsymbol{\zeta}_g)^{(1)}\end{aligned} \tag{6.7.86}$$

$$\begin{aligned}\log q(\sigma^2) =& \frac{n}{2} \log \sigma^{-2} - \frac{\sigma^{-2}}{2} \mathbb{E}_{-\sigma^2} \left[\|\mathbf{y} - \alpha \mathbf{1}_n - \sum_s X_s \gamma_s \beta_s - \boldsymbol{Z}_\xi \boldsymbol{\theta}_\xi - \sum_g \boldsymbol{W}_g \boldsymbol{\zeta}_g\|^2 \right] + \\ &+ (\tau + 1) \log \sigma^{-2} - (\nu)^{(1)} \sigma^{-2} + cst \\ =& \log \sigma^{-2} \left(\frac{n}{2} + \tau + 1 \right) + \sigma^{-2} \left(\frac{\|\boldsymbol{u}\|^{(2)}}{2} + (\nu)^{(1)} \right) + cst\end{aligned}$$

$$q(\sigma^2) = Inv - Gamma(\nu^*, \tau^*)$$

$$\nu^* = \frac{n}{2} + \tau \tag{6.7.87}$$

$$\tau^* = \frac{\|\boldsymbol{u}\|^{(2)}}{2} + (\nu)^{(1)} \tag{6.7.88}$$

where

$$(\sigma^{-2})^{(1)} = \frac{\nu^*}{\tau^*} \tag{6.7.89}$$

$$(\log \sigma^{-2})^{(1)} = \Psi(\nu^*) - \log \tau^* \tag{6.7.90}$$

$$\begin{aligned}
\log q(\kappa) =& \mathbb{E}_{-\kappa}\Big[\log \prod_j p(\xi_j|\kappa) + \log p(\kappa)\Big] + cst \\
=& \mathbb{E}_{-\kappa}\Big[\Big(\sum_j \xi_j \log(\kappa) + \sum_j (1-\xi_j)\log(1-\kappa)\Big) \mathrm{I}\Big[\sum_j \xi_j \neq 1\Big] + (a_j - 1)\log(\kappa) + \\
&+ (b_j - 1)\log(1-\kappa)\Big] + cst
\end{aligned}$$

As the update for ξ from the construction of the MCMC and the SMVN is

$$\mathbb{E}_q[\xi] = \mathbb{E}_q\left[\xi \ \mathrm{I}\Big[\sum_j \xi_j \neq 1\Big]\right] = (\xi)^{(1)} \tag{6.7.91}$$

the update can be solved in closed form, using the MCMC marginal expectations.

$$\log q(\kappa) = \left(\sum_j (\xi_j)^{(1)} + a_j - 1\right)\log(\kappa) + \left(d - \sum_j (\xi_j)^{(1)} + b_j - 1\right)\log(1-\kappa) + cst$$

$$q(\kappa) = Beta(a_\kappa^*, b_\kappa^*) \tag{6.7.92}$$

with parameters

$$a_\kappa^* = a_\kappa + \sum_j (\xi_j)^{(1)} \tag{6.7.93}$$

$$b_\kappa^* = b_\kappa + d - \sum_j (\xi_j)^{(1)} \tag{6.7.94}$$

where

$$
\begin{aligned}
(\kappa)^{(1)} &= a_\kappa^* / \left(a_\kappa^* + b_\kappa^*\right) = a_\kappa^* / \left(a_\kappa + b_\kappa + 1\right) \\
(\log \kappa)^{(1)} &= \Psi(a_\kappa^*) - \Psi(a_\kappa^* + b_\kappa^*) \\
(\log(1-\kappa))^{(1)} &= \Psi(b_\kappa^*) - \Psi(a_\kappa^* + b_\kappa^*)
\end{aligned}
\tag{6.7.95}
$$

where $\Psi(\cdot)$ is the digamma function.

The update for $q(\varrho)$ is

$$
\begin{aligned}
\log q(\varrho) &= \mathbb{E}_{-\varrho}\left[\log p(\chi_g|\varrho) + \log p(\varrho)\right] + cst \\
&= \mathbb{E}_{-\varrho}\left[\chi_g \log(\varrho) + (1-\chi_g)\log(1-\varrho) + (a_\varrho - 1)\log(\varrho) + (b_\varrho - 1)\log(1-\varrho)\right] + cst \\
&= \left((\chi_g)^{(1)} + a_\varrho - 1\right)\log(\varrho) + \left(1 - (\chi_g)^{(1)} + b_\varrho - 1\right)\log(1-\varrho) + cst
\end{aligned}
$$

$$
q(\varrho) = Beta(a_\varrho^*, b_\varrho^*) \tag{6.7.96}
$$

with parameters

$$
a_\varrho^* = a_\varrho + \sum_g (\chi_g)^{(1)} \tag{6.7.97}
$$

$$
b_\varrho^* = b_\varrho + G - \sum_g (\chi_g)^{(1)} \tag{6.7.98}
$$

where

$$
\begin{aligned}
(\varrho)^{(1)} &= a_\varrho^* / \left(a_\varrho^* + b_\varrho^*\right) = a_\varrho^* / \left(a_\varrho + b_\varrho + 1\right) \\
(\log \varrho)^{(1)} &= \Psi(a_\varrho^*) - \Psi(a_\varrho^* + b_\varrho^*) \\
(\log(1-\varrho))^{(1)} &= \Psi(b_\varrho^*) - \Psi(a_\varrho^* + b_\varrho^*)
\end{aligned}
\tag{6.7.99}
$$

where $\Psi(\cdot)$ is the digamma function.

$$\begin{aligned}\log q(\nu) &= \mathbb{E}_{-\nu}\left[\log p(\sigma^2|\tau,\nu) + \log p(\nu)\right] + cst \\ &= \tau\log\nu - \nu(\sigma^{-2})^{(1)} + (a_\nu - 1)\log\nu - \nu b_\nu \\ &= (\tau + a_\nu - 1)\log\nu - ((\sigma^{-2})^{(1)} + b_\nu)\nu\end{aligned}$$

$$q(\nu) = Inv - Gamma(a_\nu^*, b_\nu^*)$$

$$a_\nu^* = \tau + a_\nu \tag{6.7.100}$$

$$b_\nu^* = (\sigma^{-2})^{(1)} + b_\nu \tag{6.7.101}$$

where

$$(\nu)^{(1)} = \frac{a_\nu^*}{b_\nu^*} \tag{6.7.102}$$

$$(\log\nu)^{(1)} = \Psi(a_\nu^*) - \log b_\nu^* \tag{6.7.103}$$

Pseudo updates

The pseudo updates are derived in full. The prior parameterisation is

$$p(\Omega_j|\Delta_j,\Upsilon_j) = \left[\frac{1}{(2\pi\Delta_j)^{(-1/2)}}\exp\left(-\frac{1}{2\Delta_j}\Omega_j^2\right)\right]^{\Upsilon_j}\delta_0(\Omega_j)^{1-\Upsilon_j} \tag{6.7.104}$$

$$p(\Delta_j|\Upsilon_j) = \left[\frac{b_\Delta^{a_\Delta}}{\Gamma(a_\Delta)}(\Delta_j)^{-a_\Delta-1}\exp\{-b_\Delta\Delta_j^{-1}\}\right]^{\Upsilon_j}\delta_0(\Delta_j)^{1-\Upsilon_j} \tag{6.7.105}$$

$$P(\Upsilon_j) = (\kappa)^{\Upsilon_j}(1-\kappa)^{1-\Upsilon_j} \tag{6.7.106}$$

The joint update $q(\Omega_j, \Upsilon_j)$ is

$$q(\Omega_j, \Upsilon_j) \propto \mathbb{E}_{(-\Omega_j, \Upsilon_j)}\Big[\log p(y|.) + \log p(\Omega_j|\Delta_j, \Upsilon_j) + p(\Delta_j|\Upsilon_j) + p(\Upsilon_j)\Big] \tag{6.7.107}$$

$$\begin{aligned} q(\Omega_j, \Upsilon_j) \propto &\Big[N(\Omega_j|\mu_{\Omega_j}, \sigma^2_{\Omega_j})\Big]^{\Upsilon_j}\,[\delta_0(\Omega_j)]^{1-\Upsilon_j} \\ &\Big[\exp\Big(\frac{1}{2}\log\sigma^2_{\Omega_j} + (\log\kappa)^{(1)} - \frac{1}{2}\mathbb{E}_q(\log\Delta_j|\Upsilon_j) + \frac{1}{2}\mu^2_{\Omega,j}\sigma^{-2}_{\Omega,j} + a_\Delta\log(b_\Delta) + \\ &- \log(\Gamma(a_\Delta)) - (a_\Delta+1)\mathbb{E}_q(\log\Delta_j|\Upsilon_j) - b_\Delta\mathbb{E}_q[\Delta_j^{-1}|\Upsilon_j]\Big)\Big]^{\Upsilon}\Big[(1-\kappa)^{(1)} + \delta_0(\Delta_j)\Big]^{1-\Upsilon_j} \end{aligned}$$

$$\begin{aligned} \sigma^2_{\Omega,j} &= \big[\|Z_j\|^2(\sigma^{-2})^{(1)} + \mathbb{E}_q[\Delta_j^{-1}|\Upsilon_j]\big]^{-1} \\ \mu_{\Omega,j} &= \sigma^2_{\Omega,j}Z_j^T\left[(\sigma^{-2})^{(1)}\left(y - (\alpha)^{(1)}\mathbf{1}_n - \sum_{k\neq j}Z_k(\Omega_k)^{(1)} - \sum_s X_s(\beta_s)^{(1)} - \sum_g W_g(\boldsymbol{\zeta}_g)^{(1)}\right)\right] \end{aligned}$$

$$\pi(\Omega_j|\Upsilon_j = 1, y) = \mathcal{N}(\mu_{\Omega,j}, \sigma^2_{\Omega,j}), \quad q(\Omega_j|\Upsilon_j = 0, y) = \delta_0(\Omega_j) \tag{6.7.108}$$

which gives us the update

$$\begin{aligned} \sigma^2_{\Omega,j} &= \big[\|Z_j\|^2(\sigma^{-2})^{(1)} + \mathbb{E}_q[\Delta_j^{-1}|\Upsilon_j = 1]\big]^{-1} \\ \mu_{\Omega,j} &= \sigma^2_{\Omega,j}Z_j^T\left[(\sigma^{-2})^{(1)}\left(\mathbf{y} - (\alpha)^{(1)}\mathbf{1}_n - \sum_{k\neq j}Z_k(\Omega_k)^{(1)} - \sum_s X_s(\beta_s)^{(1)} - \sum \boldsymbol{W}_g(\boldsymbol{\zeta}_g)^{(1)}\right)\right]. \end{aligned}$$

The terms in the $q(\Upsilon_j)$, using $\Delta_j = 0$ when $\Upsilon_j = 0$, are proportional to

$$\begin{aligned} p(\Upsilon_j = 1) \propto &\exp\Big(\frac{1}{2}\log\sigma^2_{\Omega_j} + (\log\kappa)^{(1)} - (a_\Delta + 3/2)\mathbb{E}_q(\log\Delta_j|\Upsilon_j = 1) + \frac{1}{2}\mu^2_{\Omega,j}\sigma^{-2}_{\Omega,j} + \\ &+ a_\Delta\log(b_\Delta) - \log(\Gamma(a_\Delta)) - b_\Delta\mathbb{E}_q[\Delta_j^{-1}|\Upsilon_j = 1]\Big) \\ p(\Upsilon_j = 0) \propto &(\log(1-\kappa))^{(1)}. \end{aligned}$$

Which after normalisation is

$$(\Upsilon_j)^{(1)} = \left[1 + \exp\left\{\frac{1}{2}\log\left(\sigma_{\Omega,s}^{-2}\right) + (\log(1-\kappa))^{(1)} - (\log\kappa)^{(1)} + \frac{1}{2}\mathbb{E}_q(\log\Delta_j|\Upsilon_j = 1) - \frac{1}{2}\mu_{\Omega,j}^2\sigma_{\Omega,j}^{-2} + \right.\right.$$
$$\left.\left. - a_\Delta\log(b_\Delta) + \log(\Gamma(a_\Delta)) + (a_\Delta+1)\mathbb{E}_q(\log\Delta_j|\Upsilon_j=1) + b_\Delta\mathbb{E}_q[\Delta_j^{-1}|\Upsilon_j=1]\right\}\right]^{-1}$$

Note that now

$$(\Omega_j)^{(1)} = \mu_{\Omega,j}(\Upsilon_j)^{(1)} \tag{6.7.109}$$
$$(\Omega_j)^{(2)} = (\sigma_{\Omega,j}^2 + \mu_{\Omega,j}^2)(\Upsilon_j)^{(1)}. \tag{6.7.110}$$

The approximating q density for Δ_j, which is proportional to Δ_j but conditional on Υ_j is

$$\begin{aligned}
\log q(\Delta_j|\Upsilon_j) &\propto \mathbb{E}_{q(-\Delta_j,-\Upsilon_j)}\Big[\log p(\Omega_j|\Upsilon_j,\Delta_j) + \log p(\Delta_j|\Upsilon_j)\Big] \\
&\propto \mathbb{E}_{q(-\Delta_j,-\Upsilon_j)}\Big[\frac{1}{2}\log\Delta_j^{-1}\Upsilon_j - \frac{1}{2}\Omega_j^2\Upsilon_j\Delta_j^{-1} + \Upsilon_j(a_\Delta+1)\log\Delta_j^{-1} + \\
&\quad - b_\Delta\Upsilon_j\Delta_j^{-1} + (1-\Upsilon_j)\delta_0(\Delta_j)\Big] \\
&\propto \mathbb{E}_{q(-\Delta_j,-\Upsilon_j)}\left[(\log\Delta_j^{-1})\Upsilon_j\left(\frac{1}{2}+a_\Delta+1\right) - \Delta_j^{-1}\Upsilon_j\left(\frac{1}{2}\Omega_j^2 + b_\Delta\right) + (1-\Upsilon_j)\delta_0(\Delta_j)\right]
\end{aligned}$$

which gives us

$$q(\Delta_j|\Upsilon_j) = \Big[IG(\Delta_j|a_{\Delta_j}^*, b_{\Delta_j}^*)\Big]^{\Upsilon_j}\Big[\delta_0(\Delta_j)\Big]^{(1-\Upsilon_j)}. \tag{6.7.111}$$

Under q

$$q(\Delta_j|\Upsilon_j = 1, y) = IG(\Delta_j|a_{\Delta_j}^*, b_{\Delta_j}^*), \quad q(\Delta_j|\Upsilon_j = 0, y) = \delta_0(\Delta_j)$$

with updates

$$a^*_{\Delta,j} = \frac{1}{2} + a_\Delta \tag{6.7.112}$$

$$\begin{aligned} b^*_{\Delta,j} &= \frac{1}{2}\mathbb{E}[\Omega_j^2|\Upsilon_j = 1] + b_\Delta \\ &= \frac{1}{2}(\sigma^2_{\Omega,j} + \mu^2_{\Omega,j}) + b_\Delta. \end{aligned} \tag{6.7.113}$$

This gives

$$\mathbb{E}_q(\Delta_j^{-1}|\Upsilon_j = 1) = a^*_{\Delta,j}/b^*_{\Delta,j} \tag{6.7.114}$$

$$\mathbb{E}_q(\log \Delta_j|\Upsilon_j) = \log\left(b^*_{\Delta_j}\right) - \Psi(a^*_{\Delta_j})$$

The auxiliary parameters create an alternative DAG which is updated via a “separate branch" of pseudo updates which helps us to approximate the model in order to guide the MCMC step. These updates are refined at each iteration by the full VI updates which account for the constraint. The “sparsity" parameter κ and the hyperparameters a_Δ, b_Δ which are set to a_ψ, b_ψ provide a link back to the constrained model.

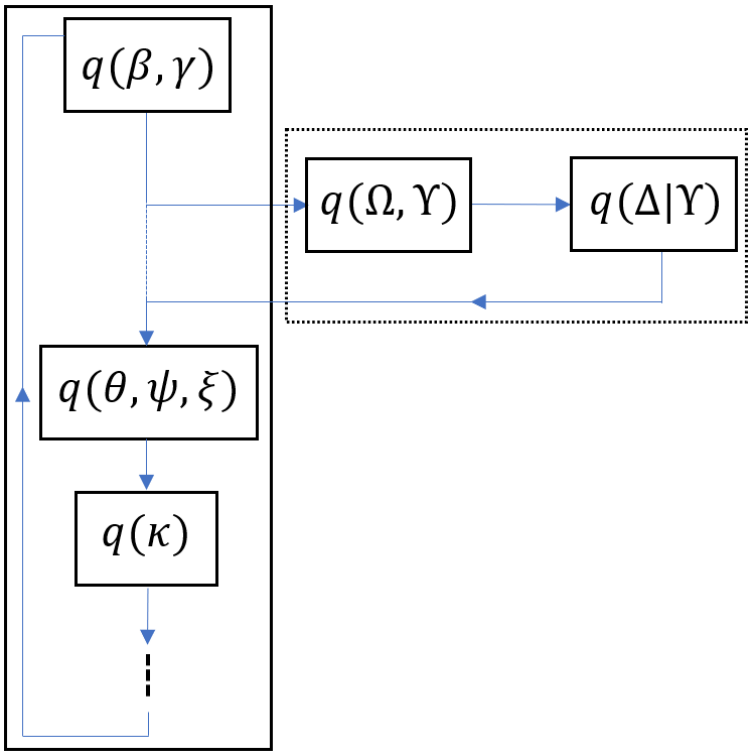

Figure 6.7.1: Diagram depicting the order and structure of the CAVI updates. Although the CAVI-MC permits any order, the pseudo updates for the auxiliary parameters help guide the MCMC and are performed directly before the $q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi})$ MC update. The pseudo updates for an unconstrained model are in the dashed box and branch off prior to the joint $q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi})$ update. The q approximating densities $q(\Delta_j|\Upsilon_j = 1)$ are then used to guide the MCMC step.

ELBO

The objective of VI is to find the candidate from a family of densities $\mathcal{D}$ which best approximates, the one closest in KL divergence, to exact conditional

$$q^*(\boldsymbol{\vartheta}) = \underset{q^*(\boldsymbol{\vartheta}) \in \mathcal{D}}{\arg\min}\, \mathrm{KL}(q(\boldsymbol{\vartheta})||p(\boldsymbol{\vartheta}|y))$$

This objective is not computable as it requires computing marginal likelihood. If we expand the expression

$$\mathrm{KL}(q(\boldsymbol{\vartheta})||p(\boldsymbol{\vartheta}|\mathbf{y})) = \mathbb{E}_{q(\boldsymbol{\vartheta})}[\log q(\boldsymbol{\vartheta})] - \mathbb{E}_{q(\boldsymbol{\vartheta})}[\log p(\boldsymbol{\vartheta}, \mathbf{y})] + \log p(\mathbf{y})$$

we can identify the elements which are a function of the parameters in the model. As the KL divergence cannot be computed, an alternative objective that is equivalent to the KL divergence up to an added constant is the evidence lower bound (ELBO).

$$\mathcal{L}(q) = \mathbb{E}_{q(\boldsymbol{\vartheta})}[\log p(\boldsymbol{\vartheta}, \mathbf{y})] - \log q(\boldsymbol{\vartheta}) \tag{6.7.115}$$

This function is the negative KL divergence plus the marginal likelihood, and is optimised at each iteration of the CAVI in order to monitor its convergence. The computational details are:

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_{q(\boldsymbol{\vartheta})}[\log p(\mathbf{y}, \boldsymbol{\vartheta})] - \mathbb{E}_{q(\boldsymbol{\vartheta})}[\log q(\boldsymbol{\vartheta})] \\
&= A(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\theta}, \sigma^2) + B^*(\alpha|w_\alpha) + \sum_s B(\beta_s, \gamma_s|w, \omega) + \tilde{B}(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi}|\kappa) + \sum_g \hat{B}(\boldsymbol{\zeta}_g, \chi_g|v, \varrho) + \\
&\quad + C(\omega) + \tilde{C}(\kappa) + \hat{C}(\varrho) + D(w) + D^*(w_\alpha) + \hat{D}(v) + \\
&\quad + F(\sigma^2|\tau, \nu) + G(\nu) + H(b_w) + H^*(b_\alpha) + \hat{H}(b_v).
\end{aligned}$$

The functions are

$$\begin{aligned}
A(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \sigma^2) &= \mathbb{E}_q[\log p(\mathbf{y}|\beta, \theta, \zeta, \sigma^2)] \\
&= \mathbb{E}_q\left[-\frac{n}{2}\log(2\pi) + \frac{n}{2}\log\left(\sigma^{-2}\right) - \frac{1}{2\sigma^2}\left\|\boldsymbol{u}\right\|^2\right] \\
&= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log\left(\sigma^2\right)^{(1)} - \frac{\left(\sigma^{-2}\right)^{(1)}\left\|\boldsymbol{u}\right\|^{(2)}}{2}
\end{aligned}$$

where $||\boldsymbol{u}||^2$ is defined in (6.7.86).

$$
\begin{aligned}
B^*(\alpha|w_\alpha) =& \mathbb{E}_q[\log p(\alpha|w_\alpha)] - \mathbb{E}_q[\log q(\alpha)] \\
=& -\frac{1}{2}\log(2\pi) + \frac{1}{2}(\log w_\alpha^{-1})^{(1)} - \frac{1}{2(w_\alpha)^{(1)}}(\alpha)^{(2)} - \\
& -\frac{1}{2}\log(2\pi) - \frac{1}{2}(\log\sigma_\alpha^2) - \frac{1}{2(\sigma_\alpha^2)}\mathbb{E}_q\left[(\alpha-\mu_\alpha)^2\right] \\
=& \frac{1}{2}\log\big(\sigma_\alpha^2\big) + \frac{1}{2}(\log w_\alpha^{-1})^{(1)} + \frac{1}{2} - \frac{1}{2}(w_\alpha^{-1})^{(1)}(\alpha)^{(2)}
\end{aligned} \tag{6.7.116}
$$

$$
\begin{aligned}
B(\beta_s,\gamma_s|w,\omega) =& \mathbb{E}_q[\log p(\beta_s|\gamma_s,w)] + \mathbb{E}_q[\log p(\gamma_s|\omega)] - \mathbb{E}_q[\log q(\beta_s,\gamma_s)] \\
=& (\gamma_s)^{(1)}\bigg(-\frac{1}{2}\log(2\pi) + \frac{1}{2}(\log w^{-1})^{(1)}\bigg) - \mathbb{E}_q\bigg[\frac{1}{2w}\gamma_s\beta_s^2\bigg] + \\
& + (1-(\gamma_s)^{(1)})\delta_0(\beta_s) + (\gamma_s)^{(1)}(\log\omega)^{(1)} + (1-(\gamma_s)^{(1)})(\log(1-\omega))^{(1)} + \\
& + \frac{1}{2}(\gamma_s)^{(1)}\bigg(\log(2\pi) + \log\sigma_{\beta,s}^2\bigg) + \mathbb{E}_q\bigg[\frac{1}{2\sigma_{\beta,s}^2}\gamma_s\Big(\beta_s^2 - 2\beta_s\mu_{\beta,s} + \mu_{\beta,s}^2\Big)\bigg] + \\
& - (1-(\gamma_s)^{(1)})\delta_0(\beta_s) - (\gamma_s)^{(1)}\log(\gamma_s)^{(1)} - (1-(\gamma_s)^{(1)})\log\big(1-(\gamma_s)^{(1)}\big)
\end{aligned} \tag{6.7.117}
$$

Simplifying using $\mathbb{E}_q\left[\frac{1}{2\sigma_{\beta,s}^2}\gamma_s\Big(\beta_s^2 - 2\beta_s\mu_{\beta,s} + \mu_{\beta,s}^2\Big)\right] = -\frac{(\gamma_s)^{(1)}}{2}$

$$
\begin{aligned}
B(\beta_s,\gamma_s|\cdot) =& \frac{(\gamma_s)^{(1)}}{2}\bigg((\log w^{-1})^{(1)} + 2(\log\omega)^{(1)} + 1 + \log\sigma_{\beta,s}^2 + 1 - 2\log(\gamma_s)^{(1)}\bigg) + \\
& - \frac{(\gamma_s)^{(1)}}{2}\bigg((\sigma_{\beta,s}^2 + \mu_{\beta,s}^2)(w)^{(-1)}\bigg) + (1-(\gamma_s)^{(1)})\bigg((\log(1-\omega))^{(1)} + \log\big(1-(\gamma_s)^{(1)}\big)\bigg).
\end{aligned}
$$

$$\begin{aligned}\hat{B}(\boldsymbol{\zeta}_g, \chi_g|v, \varrho) =&\mathbb{E}_q[\log p(\boldsymbol{\zeta}_g|\chi_g, v)] + \mathbb{E}_q[\log p(\chi_g|\varrho)] - \mathbb{E}_q[\log q(\boldsymbol{\zeta}_g, \chi_g)] \\ =&(\chi_g)^{(1)}\bigg(- \frac{m_g}{2}\log(2\pi) + \frac{m_g}{2}(\log v^{-1})^{(1)}\bigg) - \mathbb{E}_q\bigg[\frac{1}{2v}\chi_g\boldsymbol{\zeta}_g^T\boldsymbol{\zeta}_g\bigg] + \\ &+ (1-(\chi_g)^{(1)})\delta_0(\boldsymbol{\zeta}_g) + (\chi_g)^{(1)}(\log\varrho)^{(1)} + (1-(\chi_g)^{(1)})(\log(1-\varrho))^{(1)} + \\ &+ \frac{1}{2}(\chi_g)^{(1)}\bigg(m_g\log(2\pi) + \log\det(\Sigma_{\zeta_g})\bigg) + \mathbb{E}_q\bigg[\frac{1}{2}\chi_g(\boldsymbol{\zeta}_g - \boldsymbol{\mu}_{\zeta_g})^T\Sigma_{\zeta_g}^{-1}(\boldsymbol{\zeta}_g - \boldsymbol{\mu}_{\zeta_g})\bigg] + \\ &- (1-(\chi_g)^{(1)})\delta_0(\boldsymbol{\zeta}_g) - (\chi_g)^{(1)}\log(\chi_g)^{(1)} - (1-(\chi_g)^{(1)})\log\big(1-(\chi_g)^{(1)}\big)\end{aligned} \tag{6.7.118}$$

Simplifying using $\mathbb{E}_q\bigg[\chi_g\left(\boldsymbol{\zeta}_g^T\Sigma_{\zeta_g}^{-1}\boldsymbol{\zeta}_g\right)\bigg] = m_g(\chi_g)^{(1)}$

$$\begin{aligned}\hat{B}(\boldsymbol{\zeta}_g, \chi_g|v, \varrho) =&\frac{(\chi_g)^{(1)}}{2}\bigg(m_g(\log v^{-1})^{(1)} - \frac{1}{(v)^{(1)}}(tr(\Sigma_{\zeta_g}) + \boldsymbol{\mu}_{\zeta_g}^T\boldsymbol{\mu}_{\zeta_g}) + \log\det(\Sigma_{\zeta_g}) + m_g + \\ &+ 2(\log\varrho)^{(1)} - 2\log\big((\chi_g)^{(1)}\big)\bigg) + (1-(\chi_g)^{(1)})\bigg(\log\big(1-(\chi_g)^{(1)}\big) + (\log(1-\varrho))^{(1)}\bigg)\end{aligned}$$

$$\tilde{B}(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\psi}|\cdot) =\mathbb{E}_{q(\boldsymbol{\vartheta})}\Big[\log p(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{\xi}) + \log p(\boldsymbol{\psi}|\boldsymbol{\xi}) + \log p(\boldsymbol{\xi})\Big] - \mathbb{E}_{\log q(\boldsymbol{\vartheta})}\Big[\log q(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\xi})\Big] \tag{6.7.119}$$

The approximating density is only known up to a constant of proportionality but this is sufficient for the ELBO calculations.

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{\vartheta})}\Big[\log(p(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\psi}))\Big] =& -\frac{1}{2}((d_\xi)^{(1)} - 1)\log(2\pi) - \frac{1}{2}(\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi))^{(1)} + \\ &- \frac{1}{2}(\boldsymbol{\theta}_\xi^T(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)^+\theta_\xi)^{(1)} + \sum_j(\xi_j)^{(1)}(\log\kappa)^{(1)} + \\ &+ \sum_j(1-(\xi_j)^{(1)})(\log\kappa)^{(1)} - \sum_j(a_\psi+1)(\xi_j\log(\psi_j))^{(1)} + \\ &+ (a_\psi\log(b_\psi) - \log(\Gamma(a_\psi))\sum_j(\xi_j)^{(1)} - b_\psi\sum_j(\xi_j\psi_j^{-1})^{(1)}\end{aligned} \tag{6.7.120}$$

The q expectations $(\xi_j \log(\psi_j))^{(1)}$ and $(\xi_j \psi_j^{-1})^{(1)}$ can be found using the law of iterative expectations but these will cancel. The free parameters are a function of ξ so when we take an expectation we have

$$\begin{aligned}\mathbb{E}_{q(\vartheta)}\Big[\log q(\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\psi}|\mathbf{y})\Big] \propto &\mathbb{E}_{q(\boldsymbol{\vartheta})}\Big[\log(\text{SMVN}(\boldsymbol{\theta}))\Big] + \frac{1}{2}(\boldsymbol{\mu}_{\theta_\xi}^T \mathbf{T}_\xi (\mathbf{T}_\xi^T \Sigma_{\theta_\xi} \mathbf{T}_\xi)^+ \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi})^{(1)} + \\ &+ \frac{1}{2}(\log\big(\det{}^*(\mathbf{T}_\xi \Sigma_{\theta_\xi} \mathbf{T}_\xi)\big))^{(1)} - \frac{1}{2}(\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)))^{(1)} + \\ &+ \sum_j (1-\xi_j)(\log(1-\kappa))^{(1)} - \sum_j (a_\psi + 1)(\xi_j \log(\psi_j))^{(1)} + \\ &+ (a_\psi \log(b_\psi) - \log(\Gamma(a_\psi)) \sum_j (\xi_j)^{(1)} - b_\psi \sum_j (\xi_j \psi_j^{-1})^{(1)} + \sum_j \xi_j (\log\kappa)^{(1)}\end{aligned} \tag{6.7.121}$$

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{\vartheta})}\Big[\log(\text{SMVN}(\boldsymbol{\theta}))\Big] = &-\frac{1}{2}((d_\xi)^{(1)} - 1)\log(2\pi) - \frac{1}{2}(\log(\det{}^*(\mathbf{T}_\xi \Sigma_\xi \mathbf{T}_\xi)))^{(1)} + \\ &- \frac{1}{2}\Bigg((\boldsymbol{\theta}_\xi^T (\mathbf{T}_\xi \Sigma_\xi \mathbf{T}_\xi)^+ \boldsymbol{\theta}_\xi)^{(1)} - 2(\boldsymbol{\theta}_\xi^T (\mathbf{T}_\xi \Sigma_\xi \mathbf{T}_\xi)^+ \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi})^{(1)} + \\ &(\boldsymbol{\mu}_{\theta_\xi}^T \mathbf{T}_\xi (\mathbf{T}_\xi \Sigma_\xi \mathbf{T}_\xi)^+ \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi})^{(1)} \Bigg)\end{aligned} \tag{6.7.122}$$

Bringing together the expression for $\tilde{B}$

$$\begin{aligned}\tilde{B}(\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\xi}|\kappa,a_\psi,b_\psi) =&\mathbb{E}_{q(\boldsymbol{\vartheta})}\Big[\log p(\boldsymbol{\theta}|\boldsymbol{\xi},\boldsymbol{\psi}) + \log p(\boldsymbol{\psi}|\boldsymbol{\xi},a_\psi,b_\psi) + \log p(\boldsymbol{\xi}|\kappa)\Big] - \mathbb{E}_{q(\boldsymbol{\vartheta})}\Big[\log q(\boldsymbol{\theta},\boldsymbol{\xi})\Big] \\ =&-\frac{1}{2}(\log(\det{}^*(\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)))^{(1)} + \frac{1}{2}(\log(\det{}^*(\mathbf{T}_\xi \Sigma_\xi \mathbf{T}_\xi)))^{(1)} + \\ &-\frac{1}{2}\Big\{(\boldsymbol{\theta}_\xi^T (\mathbf{T}_\xi D(\boldsymbol{\psi}_\xi)\mathbf{T}_\xi)^+ \boldsymbol{\theta}_\xi)^{(1)} - (\boldsymbol{\theta}_\xi^T (\mathbf{T}_\xi \Sigma_\xi \mathbf{T}_\xi)^+ \boldsymbol{\theta}_\xi)^{(1)}\Big\} + \\ &+(\boldsymbol{\theta}_\xi^T (\mathbf{T}_\xi \Sigma_\xi \mathbf{T}_\xi)^+ \mathbf{T}_\xi \boldsymbol{\mu}_{\theta_\xi})^{(1)}\end{aligned} \tag{6.7.123}$$

$$\begin{aligned}\tilde{C}(\kappa) =&\mathbb{E}_q[\log p(\kappa)] - \mathbb{E}_q[\log q(\kappa)] \\ =&\log B(a_\kappa^*, b_\kappa^*) - \log B(a_\kappa, b_\kappa) + (a_\kappa^* - a_\kappa)(\log\kappa)^{(1)} + (b_\kappa^* - b_\kappa)(\log[1-\kappa])^{(1)}\end{aligned} \tag{6.7.124}$$

$$
\begin{aligned}
C(\omega) =& \mathbb{E}_q[\log p(\omega)] - \mathbb{E}_q[\log q(\omega)] \\
=& \log B(a^*_\omega, b^*_\omega) - \log B(a_\omega, b_\omega) + \\
&+ (a^*_\omega - a_\omega)(\log \omega)^{(1)} + (b^*_\omega - b_\omega)(\log(1-\omega))^{(1)}
\end{aligned} \tag{6.7.125}
$$

$$
\begin{aligned}
\hat{C}(\varrho) =& \mathbb{E}_q[\log p(\varrho)] - \mathbb{E}_q[\log q(\varrho)] \\
=& \log B(a^*_\varrho, b^*_\varrho) - \log B(a_\varrho, b_\varrho) + (a^*_\varrho - a_\varrho)(\log \varrho)^{(1)} + (b^*_\varrho - b_\varrho)(\log[1-\varrho])^{(1)}
\end{aligned} \tag{6.7.126}
$$

$$
\begin{aligned}
D(w) =& \mathbb{E}_q[\log p(w)] - \mathbb{E}_q[\log q(w)] \\
=& \mathbb{E}_q\Big[a_w \log b_w - \log\Gamma(a_w) + (a_w+1)\log w^{-1} - b_w w^{-1}\Big] + \\
&- \mathbb{E}_q\Big[a^*_w \log b^*_w - \log\Gamma(a^*_w) - (a^*_w+1)\log w^{-1} + b^*_w w^{-1}\Big] \\
=& a_w(\log b_w)^{(1)} - a^*_w \log b^*_w - \log\Gamma(a_w) + \log\Gamma(a^*_w) + \\
&+ (a_w - a^*_w)(\log w^{-1})^{(1)} + (b^*_w - (b_w)^{(1)})(w^{-1})^{(1)}
\end{aligned} \tag{6.7.127}
$$

$$
\begin{aligned}
D^*(w_\alpha) =& \mathbb{E}_q[\log p(w_\alpha)] - \mathbb{E}_q[\log q(w_\alpha)] \\
=& \mathbb{E}_q\Big[a_\alpha \log b_\alpha - \log\Gamma(a_\alpha) + (a_\alpha+1)\log w_\alpha^{-1} - b_\alpha w_\alpha^{-1}\Big] + \\
&- \mathbb{E}_q\Big[a^*_\alpha \log b^*_\alpha - \log\Gamma(a^*_\alpha) - (a^*_\alpha+1)\log w_\alpha^{-1} + b^*_\alpha w_\alpha^{-1}\Big] \\
=& a_\alpha(\log b_\alpha)^{(1)} - a^*_\alpha \log b^*_\alpha - \log\Gamma(a_\alpha) + \log\Gamma(a^*_\alpha) + \\
&+ (a_\alpha - a^*_\alpha)(\log w_\alpha^{-1})^{(1)} + (b^*_\alpha - (b_\alpha)^{(1)})(w_\alpha^{-1})^{(1)}
\end{aligned} \tag{6.7.128}
$$

$$
\begin{aligned}
\hat{D}(v) =& \mathbb{E}_q[\log p(v)] - \mathbb{E}_q[\log q(v)] \\
=& \mathbb{E}_q\Big[a_v \log b_v - \log\Gamma(a_v) + (a_v+1)\log v^{-1} - b_v v^{-1}\Big] + \\
& - \mathbb{E}_q\Big[a_v^* \log b_v^* - \log\Gamma(a_v^*) - (a_v^*+1)\log v^{-1} + b_v^* v^{-1}\Big] \\
=& a_v(\log b_v)^{(1)} - a_v^* \log b_v^* - \log\Gamma(a_v) + \log\Gamma(a_v^*) + \\
& + (a_v - a_v^*)(\log v^{-1})^{(1)} + (b_v^* - (b_v)^{(1)})(v^{-1})^{(1)} \qquad (6.7.129)
\end{aligned}
$$

$$
\begin{aligned}
F(\sigma^2|\tau,\nu) =& \mathbb{E}_q[\log p(\sigma^2|\tau,\nu)] - \mathbb{E}_q[\log q(\sigma^2)] \\
=& \tau(\log\nu)^{(1)} - \log\Gamma(\tau) + (\tau+1)(\log\sigma^{-2})^{(1)} - (\nu)^{(1)}(\sigma^{-2})^{(1)} + \\
& - \tau^*(\log\nu^*) - \log\Gamma(\tau^*) + (\tau^*+1)(\log\sigma^{-2})^{(1)} + \nu^*(\sigma^{-2})^{(1)} \\
=& \log\Gamma(\tau^*) - \log\Gamma(\tau) + (\tau-\tau^*)(\log\sigma^{-2})^{(1)} + \\
& + \tau(\log\nu)^{(1)} - \tau^*(\log\nu^*) + (\sigma^{-2})^{(1)}(\nu^* - (\nu)^{(1)}) \qquad (6.7.130)
\end{aligned}
$$

$$
\begin{aligned}
G(\nu) =& \mathbb{E}_q[\log p(\nu)] - \mathbb{E}_q[\log q(\nu)] \\
=& a_\nu \log b_\nu - a_\nu^* \log b_\nu^* + \log\Gamma(a_\nu^*) - \log\Gamma(a_\nu) + \\
& + (a_\nu - a_\nu^*)(\log\nu)^{(1)} + (b_\nu - b_\nu^*)(\nu)^{(1)}. \qquad (6.7.131)
\end{aligned}
$$

$$
\begin{aligned}
H(b_w) =& \mathbb{E}_q[\log p(b_w)] - \mathbb{E}_q[\log q(b_w)] \\
=& \mathbb{E}_q\Big[a_b \log b_b - \log\Gamma(a_b) + (a_b - 1)\log b_w - b_b b_w\Big] + \\
& \mathbb{E}_q\Big[a_b^* \log b_b^* - \log\Gamma(a_b^*) + (a_b^* - 1)\log b_w - b_b^* b_w\Big] \\
=& a_b \log b_b - a_b^* \log b_b^* - \log\Gamma(a_b) + \log\Gamma(a_b^*) + (\log b_w)^{(1)}(a_b - a_b^*) + \\
& + (b_w)^{(1)}(b_b^* - b_b)
\end{aligned} \tag{6.7.132}
$$

$$
\begin{aligned}
H^*(b_\alpha) =& \mathbb{E}_q[\log p(b_\alpha)] - \mathbb{E}_q[\log q(b_\alpha)] \\
=& \mathbb{E}_q\Big[a_{b,\alpha} \log b_{b,\alpha} - \log\Gamma(a_{b,\alpha}) + (a_{b,\alpha} - 1)\log b_\alpha - b_\alpha b_{b,\alpha}\Big] + \\
& \mathbb{E}_q\Big[a_{b,\alpha}^* \log b_\alpha^* - \log\Gamma(a_{b,\alpha}^*) + (a_{b,\alpha}^* - 1)\log b_\alpha - b_\alpha^* b_{b,\alpha}\Big] \\
=& a_{b,\alpha} \log b_{b,\alpha} - a_{b,\alpha}^* \log b_\alpha^* - \log\Gamma(a_{b,\alpha}) + \log\Gamma(a_{b,\alpha}^*) + (\log b_\alpha)^{(1)}(a_{b,\alpha} - a_{b,\alpha}^*) + \\
& + (b_\alpha)^{(1)}(b_{b,\alpha}^* - b_{b,\alpha})
\end{aligned} \tag{6.7.133}
$$

$$
\begin{aligned}
\hat{H}(b_v) =& \mathbb{E}_q[\log p(b_v)] - \mathbb{E}_q[\log q(b_v)] \\
=& \mathbb{E}_q\Big[a_{bv} \log b_{bv} - \log\Gamma(a_{bv}) + (a_{bv} - 1)\log b_v - b_{bv} b_v\Big] + \\
& \mathbb{E}_q\Big[a_{bv}^* \log b_{bv}^* - \log\Gamma(a_{bv}^*) + (a_{bv}^* - 1)\log b_v - b_{bv}^* b_v\Big] \\
=& a_{bv} \log b_{bv} - a_{bv}^* \log b_{bv}^* - \log\Gamma(a_{bv}) + \log\Gamma(a_{bv}^*) + (\log b_v)^{(1)}(a_{bv} - a_{bv}^*) + \\
& + (b_v)^{(1)}(b_{bv}^* - b_{bv})
\end{aligned} \tag{6.7.134}
$$

6.7.2 Proofs

Here are some simple proofs of the results used in the derivations.

Proof: Simplification of the constraint matrix

We can simplify the calculations. $\mathbf{T}\mathbf{T}^T = \mathbf{T}\mathbf{T} = \mathbf{T}$. If we define the matrix

$$\mathbf{T} = \begin{pmatrix} 1-1/d & -1/d & \dots & -1/d \\ -1/d & 1-1/d & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1/d \\ -1/d & \dots & -1/d & 1-1/d \end{pmatrix}$$

Then for the diagonal component of $\mathbf{T}\mathbf{T}$ we either have entries corresponding to the dot product of

$$\begin{bmatrix} 1-1/d \\ -1/d \\ \vdots \\ -1/d \end{bmatrix} \cdot \begin{bmatrix} 1-1/d \\ -1/d \\ \vdots \\ -1/d \end{bmatrix} = (1-1/d)^2 + \frac{d-1}{d^2} = 1-1/d \tag{6.7.135}$$

where $1-1/d$ is in the same position in the vector. The off-diagonal entries correspond to dot product of vectors where the position of the $1-1/d$ terms are not matched which always gives us

$$(1-1/d) \times (-2/d) + (d-2)/d^2 = -1/d \; \square \tag{6.7.136}$$

Using the matrix determinant lemma where A is an invertible square matrix and u, v are column vectors

$$\det\big(A + uv^T\big) = (1 + v^T A^{-1} u)\det(A) \tag{6.7.137}$$

we can prove that the determinant of this matrix is zero. Express $\mathbf{T}$ as

$$\begin{aligned}\mathbf{T} &= (\mathbf{I}_d - (1/d)\mathbf{1}_{d\times d}) \\ &= \mathbf{I}_d + \begin{bmatrix} -1/\sqrt{d} \\ \vdots \\ -1/\sqrt{d} \end{bmatrix} \begin{bmatrix} 1/\sqrt{d} & \dots & 1/\sqrt{d} \end{bmatrix}\end{aligned}$$

Thus

$$\det(\mathbf{T}) = 1 + \begin{bmatrix} 1/\sqrt{d} & \dots & -1/\sqrt{d} \end{bmatrix} \begin{bmatrix} -1/\sqrt{d} \\ \vdots \\ 1/\sqrt{d} \end{bmatrix} \tag{6.7.138}$$

$$= 1 - 1 = 0 \ \square \tag{6.7.139}$$

Proof: Eigenvalues of T comprise of $d-1$ 1's and one 0.

To find the eigenvalues of $\mathbf{T}$ need to solve

$$\det(\mathbf{T} - \lambda\mathbf{I}) = 0 \tag{6.7.140}$$

for λ. Using the lemma in Equation (6.7.137) and $\mathbf{T} - \lambda\mathbf{I} = A + uv^T$ where

$$A = \text{diag}(1-\lambda) \qquad u = \begin{bmatrix} -1/\sqrt{d} \\ \vdots \\ -1/\sqrt{d} \end{bmatrix} \qquad v = \begin{bmatrix} 1/\sqrt{d} \\ \vdots \\ 1/\sqrt{d} \end{bmatrix} \tag{6.7.141}$$

we have

$$\begin{aligned}\det(\mathbf{T}-\lambda\mathbf{I}) &= (1+v^T\mathrm{diag}((1-\lambda)^{-1}))u)(1-\lambda)^d \\ &= (1-(1-\lambda)^{-1})(1-\lambda)^d \\ &= \left(\frac{1-\lambda+1}{1-\lambda}\right)(1-\lambda)^d \\ &= -\lambda(1-\lambda)^{d-1}.\end{aligned} \tag{6.7.142}$$

Therefore the eigenvalues for $\mathbf{T}$ are

$$\lambda_1, \lambda_2, ..., \lambda_{d-1} = 1 \qquad \lambda_d = 0 \ \square \tag{6.7.143}$$

Proof: Pseudo Inverse of the constraint matrix

Using the SVD $\mathbf{T}$ can be expressed as $U\Lambda V$. As $\mathbf{T}$ is symmetric $U\Lambda V = U\Lambda U$. The pseudo inverse is

$$\mathbf{T}^+ = U\Lambda^+U^T = \begin{bmatrix} & & \\ u_1 & \cdots & u_d \\ & & \end{bmatrix} \begin{bmatrix} \lambda_1^{-1} & & & \\ & \ddots & & \\ & & \lambda_{d-1}^{-1} & \\ & & & 0 \end{bmatrix} \begin{bmatrix} & & \\ u_1 & \cdots & u_d \\ & & \end{bmatrix} \tag{6.7.144}$$

As the non zero eigenvalues all equal 1

$$\mathbf{T} = \begin{bmatrix} & & \\ u_1 & \cdots & u_d \\ & & \end{bmatrix} \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{bmatrix} \begin{bmatrix} & & \\ u_1 & \cdots & u_d \\ & & \end{bmatrix}. \tag{6.7.145}$$

This approach can also be used to solve the pseudo determinant $\det^*(\theta\mathbf{T})$ (where θ is a scalar) which is a product of the non-zero eigenvalues. The eigenvalues of the scaled matrix can be found

solving $\det(\theta\mathbf{T} - \lambda I) = 0$.

$$\det(\theta\mathbf{T} - \lambda\mathbf{I}) = (1 + v^T A^{-1} u)\det(A) \tag{6.7.146}$$

where

$$A = \operatorname{diag}(\theta - \lambda) \qquad u = \begin{bmatrix} -\sqrt{\theta/d} \\ \vdots \\ -\sqrt{\theta/d} \end{bmatrix} \qquad v = \begin{bmatrix} \sqrt{\theta/d} \\ \vdots \\ \sqrt{\theta/d} \end{bmatrix} \tag{6.7.147}$$

Simplifying gives

$$\det(\theta\mathbf{T} - \lambda\mathbf{I}) = (1 + v^T A^{-1} u)\det(A) \tag{6.7.148}$$

$$= -\lambda(\theta - \lambda)^{d-1} \tag{6.7.149}$$

The eigenvalues, found by setting this expression to zero are

$$\lambda_1, \lambda_2, ..., \lambda_{d-1} = \theta \qquad \lambda_d = 0 \ \square \tag{6.7.150}$$

Thus the expression

$$\det^*(2\pi w T) = (2\pi w)^{d-1}. \tag{6.7.151}$$

6.7.3 Simulation results

The full set of results from the simulation study are presented in Table (6.3) - Table (6.7).

Table 6.3: Table of true positive rate, false positive rate and the L2 loss of the estimated parameters for the true model with elements $\boldsymbol{\theta}$ as the only significant parameter for the variational Bayes (VB), OLS and group lasso (GL) approach and a SNR of 0.5. The total number of compositional, continuous and categorical covariates are represented by d, p and G respectively.

$(n, d, p+G)$	ρ	Method	TPR	FPR	L2 loss
(100, 45, 24)	0	VB	1.00	0.01	0.20
		OLS	0.53	0.02	3.56
		GL	1.00	0.40	4.02
(100, 45, 24)	0.2	VB	0.96	0.01	0.46
		OLS	0.67	0.06	5.53
		GL	1.00	0.48	8.42
(100, 45, 24)	0.5	VB	0.74	0.00	1.64
		OLS	0.51	0.04	4.67
		GL	0.98	0.50	5.08
(100, 100, 24)	0	VB	0.99	0.01	0.19
		GL	1.00	0.15	0.61
(100, 100, 24)	0.2	VB	0.99	0.00	0.25
		GL	1.00	0.19	1.10
(100, 100, 24)	0.5	VB	0.33	0.00	4.07
		GL	1.00	0.25	2.16
(200, 100, 24)	0	VB	1.00	0.01	0.09
		OLS	0.86	0.00	0.64
		GL	1.00	0.18	0.57
(200, 100, 24)	0.2	VB	1.00	0.00	0.09
		OLS	0.85	0.00	0.68
		GL	1.00	0.17	0.42
(200, 100, 24)	0.5	VB	1.00	0.04	0.04
		OLS	0.74	0.00	1.61
		GL	1.00	0.23	0.63

Table 6.4: Table of true positive rate, false positive rate and the L2 loss of the estimated parameters for the true model with elements $\boldsymbol{\theta}$ as the only significant parameter for the variational Bayes (VB), OLS and group lasso (GL) approach and a SNR of 1. The total number of compositional, continuous and categorical covariates are represented by d, p and G respectively.

$(n, d, p+G)$	ρ	Method	TPR	FPR	L2 loss
(100, 45, 24)	0	VB	1.00	0.00	0.08
		OLS	0.94	0.08	2.32
		GL	0.98	0.35	3.86
(100, 45, 24)	0.2	VB	1.00	0.01	0.04
		OLS	0.97	0.16	2.13
		GL	0.99	0.68	3.63
(100, 45, 24)	0.5	VB	0.94	0.00	0.39
		OLS	1.00	0.16	2.41
		GL	1.00	0.62	3.84
(100, 100, 24)	0	VB	1.00	0.00	0.06
		GL	1.00	0.18	0.26
(100, 100, 24)	0.2	VB	1.00	0.01	0.06
		GL	1.00	0.17	0.33
(100, 100, 24)	0.5	VB	1.00	0.00	0.05
		GL	1.00	0.22	0.75
(200, 100, 24)	0	VB	1.00	0.00	0.03
		OLS	0.99	0.00	0.23
		GL	1.00	0.22	0.16
(200, 100, 24)	0.2	VB	1.00	0.00	0.03
		OLS	1.00	0.00	0.13
		GL	1.00	0.15	0.13
(200, 100, 24)	0.5	VB	1.00	0.00	0.02
		OLS	1.00	0.00	0.88
		GL	1.00	0.23	0.25

Table 6.5: Table of true positive rate, false positive rate and the L2 loss of the estimated parameters for the true model with elements $\boldsymbol{\theta}$ as the only significant parameter for the variational Bayes (VB), OLS and group lasso (GL) approach and a SNR of 5. The total number of compositional, continuous and categorical covariates are represented by d, p and G respectively.

$(n, d, p+G)$	ρ	Method	TPR	FPR	L2 loss
(100, 45, 24)	0	VB	1.00	0.04	0.01
		OLS	0.99	0.10	2.06
		Lasso	1.00	0.59	0.74
(100, 45, 24)	0.2	VB	1.00	0.03	0.00
		OLS	1.00	0.06	1.64
		Lasso	1.00	0.66	2.91
(100, 45, 24)	0.5	VB	1.00	0.09	0.00
		OLS	0.84	0.07	2.37
		Lasso	1.00	0.54	7.41
(100, 100, 24)	0	VB	1.00	0.01	0.00
		Lasso	1.00	0.20	0.02
(100, 100, 24)	0.2	VB	1.00	0.04	0.00
		Lasso	1.00	0.22	0.02
(100, 100, 24)	0.5	VB	1.00	0.00	0.01
		Lasso	1.00	0.27	0.17
(200, 100, 24)	0	VB	1.00	0.00	0.02
		OLS	0.99	0.00	0.23
		Lasso	1.00	0.22	0.16
(200, 100, 24)	0.2	VB	1.00	0.00	0.00
		OLS	1.00	0.00	0.38
		Lasso	1.00	0.18	0.01
(200, 100, 24)	0.5	VB	1.00	0.04	0.00
		OLS	1.00	0.00	1.19
		Lasso	1.00	0.23	0.25

Table 6.6: Table of true positive rate, false positive rate and the L2 loss of the estimated parameters for the true model with elements of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ as significant parameters for the variational Bayes (VB), OLS and group lasso (GL) approach and a SNR of 1. The total number of compositional, continuous and categorical covariates are represented by d, p and G respectively and are described in the Cov column. The combinations denoted by A, B and C are
$(n = 100, d = 45, p + G = 24)$, $(100, 100, 24)$ and $(200, 100, 24)$ respectively.

Cov	ρ	Method	TPR	FPR	TPR($\boldsymbol{\theta}$)	FPR($\boldsymbol{\theta}$)	TPR($\boldsymbol{\beta}, \boldsymbol{\zeta}$)	FPR($\boldsymbol{\beta}, \boldsymbol{\zeta}$)	L2
A	0	VB	0.99	0.01	1.00	0.01	0.97	0.01	0.92
		OLS	0.49	0.11	0.84	0.15	0.23	0.05	12.99
		GL	0.70	0.40	0.98	0.55	0.50	0.18	8.09
A	0.2	VB	1.00	0.00	1.00	0.00	0.99	0.00	0.67
		OLS	0.46	0.09	0.72	0.14	0.26	0.03	10.40
		GL	0.79	0.61	1.00	0.72	0.63	0.45	19.39
A	0.5	VB	0.30	0.01	0.10	0.00	0.53	0.04	11.60
		OLS	0.44	0.08	0.61	0.12	0.31	0.03	9.86
		GL	0.74	0.65	0.96	0.71	0.57	0.58	2.79
B	0	VB	0.99	0.00	1.00	0.00	0.98	0.01	0.94
		GL	0.77	0.20	1.00	0.20	0.60	0.19	5.71
B	0.2	VB	0.99	0.00	1.00	0.00	0.98	0.01	0.99
		GL	0.74	0.65	0.96	0.71	0.57	0.58	2.79
B	0.5	VB	0.36	0.00	0.26	0.00	0.48	0.00	9.69
		GL	0.68	0.27	0.89	0.22	0.53	0.21	4.28
C	0	VB	1.00	0.00	1.00	0.00	1.00	0.00	0.37
		OLS	0.68	0.00	1.00	0.00	0.43	0.00	4.57
		GL	1.00	0.30	1.00	0.32	1.00	0.23	4.06
C	0.2	VB	1.00	0.00	1.00	0.00	1.00	0.01	0.40
		OLS	0.67	0.00	1.00	0.00	0.42	0.00	4.65
		GL	0.99	0.35	1.00	0.37	0.98	0.29	2.53
C	0.5	VB	1.00	0.00	1.00	0.00	1.00	0.00	0.02
		OLS	0.68	0.00	1.00	0.00	0.44	0.00	5.16
		GL	1.00	0.33	1.00	0.33	1.00	0.30	2.74

Table 6.7: Table of true positive rate, false positive rate and the L2 loss of the estimated parameters for the true model with elements of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ as significant parameters for the variational Bayes (VB), OLS and group lasso (GL) approach and a SNR of 5. The total number of compositional, continuous and categorical covariates are represented by d, p and G respectively and are described in the Cov column. The combinations denoted by A, B and C are $(n = 100, d = 45, p + G = 24)$, $(100, 100, 24)$ and $(200, 100, 24)$ respectively.

Cov	ρ	Method	TPR	FPR	TPR($\boldsymbol{\theta}$)	FPR($\boldsymbol{\theta}$)	TPR($\boldsymbol{\beta}, \boldsymbol{\zeta}$)	FPR($\boldsymbol{\beta}, \boldsymbol{\zeta}$)	L2
A	0	VB	1.00	0.05	1.00	0.07	1.00	0.01	0.04
		OLS	0.77	0.09	1.00	0.14	0.59	0.00	6.33
		GL	1.00	0.59	1.00	0.70	1.00	0.43	5.40
A	0.2	VB	1.00	0.07	1.00	0.10	1.00	0.00	0.06
		OLS	0.57	0.08	1.00	0.14	0.25	0.00	7.77
		GL	0.91	0.62	1.00	0.77	0.85	0.38	5.28
A	0.5	VB	1.00	0.03	1.00	0.04	1.00	0.00	0.07
		OLS	0.73	0.08	0.93	0.00	0.59	0.00	6.78
		GL	1.00	0.67	1.00	0.70	1.00	0.63	2.13
B	0	VB	1.00	0.00	1.00	0.00	1.00	0.00	0.10
		GL	1.00	0.16	1.00	0.13	1.00	0.26	5.84
B	0.2	VB	1.00	0.01	1.00	0.01	1.00	0.00	0.03
		GL	1.00	0.10	1.00	0.12	1.00	0.05	4.1
B	0.5	VB	1.00	0.00	1.00	0.00	1.00	0.00	0.05
		GL	0.68	0.27	0.89	0.22	0.53	0.21	4.28
C	0	VB	1.00	0.02	1.00	0.03	1.00	0.00	0.04
		OLS	0.92	0.00	1.00	0.00	0.88	0.00	0.86
		GL	1.00	0.35	1.00	0.36	1.00	0.30	1.21
C	0.2	VB	1.00	0.03	1.00	0.04	1.00	0.00	0.03
		OLS	0.89	0.00	1.00	0.00	0.81	0.00	1.60
		GL	1.00	0.31	1.00	0.33	1.00	0.21	0.81
C	0.5	VB	1.00	0.00	1.00	0.00	1.00	0.00	0.02
		OLS	0.91	0.00	1.00	0.00	0.84	0.00	1.90
		GL	1.00	0.41	1.00	0.41	1.00	0.43	0.54

CHAPTER 7

Bayesian Multiple Response Compositional Regression with Microbiome Features via Variational Inference

7.1 Abstract

The microbiome has an important role within the human body. As we seek to reveal the pathways that underlie common human disease, interest lies in finding microbiome features which are correlated with the hosts physiology. An important challenge in microbiome research is that current sequencing protocols can only provide information about the relative abundance of its constituting components. This compositionality cannot be accounted for by standard statistical frameworks. Almost all of the approaches which have been developed to account for the compositional nature only allow a single response. Multivariate approaches which capture the latent structure of the responses to increase statistical power and data understanding and improve model estimation, provide a considerable improvement to a univariate approach. We develop a Bayesian hierarchical multiple response linear log-contrast model which is estimated by mean field Monte-Carlo

co-ordinate ascent variational inference (CAVI-MC) to address these challenges. By a reparameterisation of the seemingly unrelated regression framework, correlation between the responses is captured and different regressors are free to be associated with different responses. We use priors which account for the large difference in scale and constrained parameter space associated with the compositional covariates. Intractable marginal expectations are estimated by a reversible jump Monte Carlo Markov Chain guided by the data through univariate approximations of the variational posterior probability of inclusion, with proposal parameters informed by approximating variational densities via auxiliary parameters. Software has been developed in python which is freely available. We apply our CAVI-MC model to the "Know Your Heart" study, exploring the relationship between gut microbiome, health covariates and a set of biomarkers.

7.2 Introduction

One of the most widely used approaches for enumerating the microbiome is amplicon sequencing with the 16S ribosomal DNA marker gene. After preprocessing the raw sequences from the samples, the 16S sequence reads are clustered into operational taxonomic units (OTUs) (Bharti and Grimm, 2021). The abundances of microbial OTUs are compositional. They are not independent and only provide information about the relative magnitudes of the components because they have an arbitrary total imposed by the sequencing instruments (Gloor et al., 2017). This means that the standard methods of analysis such as linear regression are not applicable to microbiome data (Li, 2015), unless an appropriate transformation is performed.

The high dimensionality of these datasets, where the space of possible combinations of significant variables is large, imposes a high computational burden. Typically, sparsity is expected where just a few species are associated with the response, but these associations will vary across the responses. Bayesian variable selection approaches have the advantage of being able to include prior knowledge and simultaneously incorporate many sources of variation. Explicit variable selection (George and McCulloch (1993), Kuo and Mallick (1998), Dellaportas et al. (2002)) produces posterior distributions of model inclusion and parameter values which enable model choice and a

probabilistic understanding of the strength and nature of the association.

To model compositional data, a transformation must be performed to transfer the compositional vectors into the Euclidean space. A variety of log-ratio transformations have been proposed including additive log-ratio (alr), centred log-ratio (Aitchison, 1982) and more recently isometric log-ratio (Egozcue et al., 2003). The alr transformation allows a direct inference in frequentist regression problems between selected covariates and the compositional data set (Aitchison and Bacon-Shone, 1984). Lin et al. (2014) propose an adaptive l_1 regularisation regression for sparsity with the constraint imposed by the log contrasts. Zhang et al. (2020) introduce a generalised transformation matrix on the parameters in the Bayesian framework, similar to the generalized lasso, which does not require constraining the parameters to the affine hyperplane. By treating the constraint as a tuning parameter within the generalised matrix which is never strictly imposed, a conjugate prior parametrisation allows that the marginal posterior of the selection parameter to be derived within a Gibbs sampler.

Often interest falls in understanding the relationship between the microbiome and a complex set of phenotypes such as lipids (Matey-Hernandez et al., 2018) or metabolites (Bharti and Grimm, 2021). A multivariate approach which is able to capture the latent structure of the responses thereby increasing statistical power (Inouye et al., 2012) and improving model estimation and data understanding, offers a considerable improvement to the univariate approach. Extending linear models to multivariate outcomes creates a large and complex posterior space, presenting computational and statistical problems which have been addressed in a variety of applications.

The seemingly unrelated regression (SUR) framework is applied in the Bayesian framework by Holmes et al. (2002), allowing the residuals across the regression model to be correlated. Partially conjugate priors are defined on parameters in the original parametrisation of Zellner (1962) to obtain conditional posteriors (as the marginal posteriors are intractable). The MCMC approach, which is adapted for a random design matrix, is only compatible with very small datasets given the size of both the design matrix and precision matrix which the conditional posteriors are a function of. Motivated by the SUR model, Bhadra and Mallick (2013) combine a matrix variate normal

likelihood with explicit variable selection and Gaussian graphical modelling. With a focus on a sparse covariance matrix, Gaussian graphical modelling with decomposable graphs is used to model the precision matrix where the edges of the graph between nodes correspond to non-zero entries in the precision matrix (Wermuth, 1976). Although this achieves computational improvements, the approach is still restricted to a small number of responses which are all associated with the same set of regressors. Banterle and Lewin (2018) use a reparametrisation of the SUR to make it computationally feasible to capture the correlation across hundreds of responses whilst allowing different covariates to be associated with different responses.

Despite adaptations to Bayesian multiple response algorithms such as MT-HESS (Lewin et al., 2016) with adaptive parallel tempering or factorisation of the likelihood into conditionally independent products (Banterle and Lewin, 2018), large datasets can still prohibit the MCMC from fully searching the large model space. Variational inference is an alternative approach which uses optimisation to achieve large computational savings by approximating the marginal posterior densities. Carbonetto and Stephens (2012) use variational inference for linear regression with a univariate response Carbonetto and Stephens (2012) for large omics datasets. This is extended to multiple responses by Ruffieux et al. (2017) who use a similar hierarchy framework as Bottolo et al. (2011). By choosing conditionally conjugate prior distributions and specifying a mean field variational family, closed form iterative updates which minimise the Kullback-Leibler divergence between the approximating densities and the exact posterior densities are obtained. However many models of interest, such as logistic regression and non conjugate topic models, do not enjoy the properties required to exploit this algorithm.

We extend the Bayesian hierarchical linear log-contrast model for compositional data in Scott and Lewin (2021) to multi-dimensional responses, linking high-dimensional multivariate regressions in a computationally efficient way. The latent response structure is captured by a covariance matrix within a SUR framework, before the properties of a bivariate normal are exploited to iteratively factorise the matrix. Feature selection priors on the reparameterised model introduces convenient covariance selection, bypassing the computational challenges encountered with Gaussian graphical models. The flexible model framework enables us to avoid the restrictive assumption of either

independent conditional residuals or association of the same set of regressors with all the responses. By capturing the information across the responses, the ability to detect covariates associated with the response improves. This is particularly important in the context of high dimensional microbiome data where $p >> n$ and the response of interest often comprises a complex biological phenotype.

The model is estimated by mean field Monte Carlo co-ordinate ascent variational inference (CAVI-MC). A reparameterised *alr* transformation on the compositional data avoids the need for any reference category, but imposes a sum to zero constraint on the respective parameters. We account for this, as well as the large differences in the abundances of features in the microbiome data, with priors within a hierarchical prior framework. Monte Carlo expectations are used to approximate intractable integrals because the priors associated with the compositional data are not conditionally conjugate. These expectations are estimated by a reversible jump Monte Carlo Markov chain (RJMCMC) (Green, 1995), guided by the data through a univariate approximation of the intractable variational probability of inclusion. Auxiliary parameters are introduced, with their corresponding variational densities used as proposal distributions. Model averaging over all the explored models can be performed and shrunk estimates of the regression coefficient (by the model uncertainty) are available.

The multiple response CAVI-MC model is applied to a subset of the "Know Your Heart" cross-sectional study of cardiovascular disease (Cook et al., 2018), examining the association between a set of biochemistries analysed using blood and urine samples and a set of covariates containing both unconstrained and compositional data. The set of biochemistries comprises seven biomarkers which includes lipids, renal function, liver function and metabolites. These are measured by nineteen biological quantities such as creatine and albumin and exhibit large correlation, particularly between the quantities within the target biomarker. The study was conducted in two Russian cities Novosibirsk and Arkhangelsk, containing 45,252 men and women aged between 35-69 years recruited from the general population. A health check questionnaire was completed, providing information on age, sex, alcohol, diet, smoking status and education level. We analyse the microbiome of 685 subjects from the Arkhangelsk region at the phylum level, as the 16S rRNA

sequencing of faecal samples was only performed for these participants. We find creatinine from urine samples to be associated with Actinobacteria and Verrucomicrobia after controlling for age, body mass index (BMI) and smoking status.

7.3 Model

7.3.1 Microbiome data

We start from a model with a multivariate response $\boldsymbol{Y} = (\mathbf{y}_1, ..., \mathbf{y}_T)$, $\mathbf{y}_t = (y_{t1}, ..., y_{tn})$ for $t = 1, ..., T$ and an $n \times d$ design matrices $\boldsymbol{Q}_{n\times d}$ which contain the standardised rows of the microbiome OTU raw counts (each row sums to 1), where zeros have been replaced by a small pseudo-cont (typically 0.5). The set of compositional explanatory variables can be transformed onto the unconstrained sample space $\mathbb{R}^{d-1}$ using the alr transformation

$$alr(\boldsymbol{q}_i) = \Big[\log\left(\frac{q_{i1}}{q_{id}}\right), \log\left(\frac{q_{i2}}{q_{id}}\right), ..., \log\left(\frac{q_{id-1}}{q_{id}}\right)\Big], \tag{7.3.1}$$

where $\boldsymbol{q}_i$ is the ith row of $\boldsymbol{Q}$ and the ratios have been arbitrarily chosen to involve the division of each of the first $d-1$ components by the final component. The log linear model, with the alr transformed variables as proposed by Aitchison and Bacon-Shone (1984), can be expressed as a set of linked regressions

$$y_{it} = alr(\boldsymbol{q}_i)\tilde{\boldsymbol{\theta}}_t + u_{it} \qquad t = 1, ..., T. \tag{7.3.2}$$

with $\tilde{\boldsymbol{\theta}}_t = (\theta_{t1}, ..., \theta_{t,d-1})^T$ as the corresponding $(d-1)$ vector of regression coefficients. Importantly the residuals will be correlated $\mathbf{u}_i = (u_{i1}, ..., u_{iT}) \sim N_T(\mathbf{0}, \boldsymbol{C})$, where $\boldsymbol{C}$ is a $T \times T$ non-diagonal positive definite matrix. Although convenient, the interpretation of the model depends on the arbitrary choice of the reference category. If we expand the dot product $alr(\boldsymbol{q}_i)\tilde{\boldsymbol{\theta}}_t$ and set

$$\theta_{td} = -\sum_{j}^{d-1} \tilde{\theta}_{tj} \tag{7.3.3}$$

the linked linear model for a response can be expressed in matrix form (Lin et al., 2014) as

$$\mathbf{y}_t = \mathbf{Z}\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t, \quad t = 1, ..., T \quad \text{subject to } \sum_{j=1}^{d} \theta_{tj} = 0 \tag{7.3.4}$$

where $\boldsymbol{Z} = (\log \boldsymbol{q}_1, ..., \log \boldsymbol{q}_d)$ is the log of the $n \times d$ compositional design matrix $\boldsymbol{Q}$ and $\boldsymbol{\theta}_t = (\theta_{t1}, ..., \theta_{td})^T$ is a d-vector of regression coefficients constrained to sum to zero.

7.3.2 Factorisation of the likelihood

The linked linear model in (7.3.4) can be expressed as a SUR model (Zellner, 1962) with the T vector equations stacked on top of each other in the form

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z} & \mathbf{0} & \dots & \mathbf{0} \\ 0 & \boldsymbol{Z} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ \vdots \\ \boldsymbol{\theta}_T \end{pmatrix} + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_T \end{pmatrix} = \tilde{\boldsymbol{Z}}\boldsymbol{\theta} + \mathbf{u}$$
$$\mathbf{u} \sim N_{n \times T}(\mathbf{0}, \boldsymbol{C} \otimes \mathbb{I}_n). \tag{7.3.5}$$

The error terms $\mathbf{u}_t$ from the same regression are assumed to be independent given the model covariates, and the residual variance is free to change across the models. Importantly, correlation between the error terms of different models is captured in $\boldsymbol{C}$, allowing the responses to be correlated between themselves.

In the standard regression setting (where $\boldsymbol{\theta}$ is unconstrained), assuming the same $\boldsymbol{\gamma}_t$ for all t or a diagonal $\boldsymbol{C}$ and conjugate priors for $\boldsymbol{\theta}$ and $\boldsymbol{C}$, $\boldsymbol{C}$ and $\boldsymbol{\theta}$ can be integrated out analytically (Petretto et al. (2010), Bhadra and Mallick (2013)). In the more general case, the usual priors on the parameters are no longer conjugate and can not be integrated out. A Gibbs sampler for the posterior distribution is straightforward to write as the full conditionals retain their simple forms (Holmes et al., 2002), however the computational time is prohibitive for most dimensional-settings. Variational inference, can be a considerably cheaper alternative to MCMC techniques

in high-dimensional settings. Although direct comparison can be difficult, Ruffieux et al. (2017) achieve a favourable result with a varitional algorithm that converges within tens of iterations, as opposed to MCMC sampling which may require thousands of iterations to converge.

To overcome the computational challenges we begin by factorising the likelihood to

$$p(\boldsymbol{Y}|\tilde{\boldsymbol{Z}}, \boldsymbol{\theta}, \boldsymbol{C}) = \prod_{t=1}^{T} \psi(\mathbf{y}_t|\boldsymbol{Z}\boldsymbol{\theta}_t + \boldsymbol{U}_{(t-1)}\boldsymbol{\rho}_t, \sigma_t^2\mathbb{I}_n) \tag{7.3.6}$$

where the matrix $\boldsymbol{U}_{(t-1)} = \boldsymbol{Y}_{(t-1)} - (\boldsymbol{Z}\boldsymbol{\theta}_1 \ ... \ \boldsymbol{Z}\boldsymbol{\theta}_{t-1})$ consists of the first $t-1$ residuals from the linked regression and $\psi(\mathbf{y}|\boldsymbol{\mu}, \Sigma)$ is the probability density function for the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. The ordering of the decomposition does not affect the joint distribution $p(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{C})$ as the factoring is by chain-conditioning. The parameter σ_t^2 is the residual variance of the response t conditioned on the $\boldsymbol{U}_{t-1}$ residuals, $\boldsymbol{\rho}_t$ is a real valued vector of regression coefficients.

We include other covariates of interest within the reparameterised likelihood (7.3.6). This takes the form of continuous covariates $\boldsymbol{X}$ and a categorical design matrix $\boldsymbol{W}$ which contains dummy variables for the $g = 1, ..., G$ categorical variables coded to indicate the m_g levels with respect to an intercept

$$p(\boldsymbol{Y}|\cdot) = \prod_{t=1}^{T} \psi(\mathbf{y}_t|\alpha_t\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}_t + \boldsymbol{W}\boldsymbol{\zeta}_t + \boldsymbol{Z}\boldsymbol{\theta}_t + \boldsymbol{U}_{(t-1)}\boldsymbol{\rho}_t, \sigma_t^2\mathbb{I}_n) \quad \text{subject to } \sum_{j=1}^{d} \theta_{tj} = 0 \tag{7.3.7}$$

where the matrix of residuals in the mean function are defined as

$$\boldsymbol{U}_{(t-1)} = \boldsymbol{Y}_{(t-1)} - (\mathbf{1}_n\alpha_1 + \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{W}\boldsymbol{\zeta}_1 + \boldsymbol{Z}\boldsymbol{\theta}_1 \ ... \ \mathbf{1}_n\alpha_{t-1} + \boldsymbol{X}\boldsymbol{\beta}_{t-1} + \boldsymbol{W}\boldsymbol{\zeta}_{t-1} + \boldsymbol{Z}\boldsymbol{\theta}_{t-1}). \tag{7.3.8}$$

The parametrisation of the likelihood breaks up the stacked design matrix in Equation (7.3.5) and produces a diagonal covariance matrix which ensures a quicker computational time and a more manageable algebraic form when deriving the complete conditional in variational inference. The product of conditionally independent Gaussian densities allows us to exploit the statistical

framework in Scott and Lewin (2021) and greatly improves the computational feasibility of the model.

If we define the residual from a draw across the T responses as $\boldsymbol{\epsilon}_i \sim N_T(\mathbf{0}, \boldsymbol{D})$ with $\boldsymbol{D}$ as a diagonal matrix $D_{tt} = \sigma_t^2$, the likelihood for this single observation vector is

$$\boldsymbol{u}_{i,T}^T(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{\epsilon}_i$$

with the vector $\boldsymbol{u}_{i,T}$ as the ith row from the $\boldsymbol{U}_T$ matrix and $\boldsymbol{P}$ as a lower triangle nilpotent matrix with $\boldsymbol{P}_{ts} = \rho_{ts}$ $(t > s)$. Taking the variance of this expression gives us

$$\boldsymbol{D} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{C}(\boldsymbol{I} - \boldsymbol{P})^T. \tag{7.3.9}$$

This factorisation is popular in autoregressive modelling and graphical models. Banterle and Lewin (2018) use a Cholesky factorisation of the precision matrix and perfect elimination ordering so that the zeros in the ρ_t correspond to zeros in the precision matrix represented by a decomposable graph structure. Pourahmadi (1999) use expression (7.3.5) within linear regression to add significance testing to the now unconstrained transformed off-diagonal elements for covariance selection whilst maintaining the positive definite property of the covariance matrix. Smith and Kohn (2002) extend this interpretation of the Cholesky decomposition to the Bayesian framework.

7.3.3 Unconstrained Priors

The parameters in the model are estimated completely in the reparameterised space, where the priors on the new parameters $\{\sigma_t^2, \boldsymbol{\rho}_t\}$ are determined by starting with an Inverse Wishart prior on the positive definite matrix $\boldsymbol{C} \sim IW(\nu, M)$, in the original parametrisation of the model (7.3.5). As $\boldsymbol{C}_{(t)}$ is a submatrix of $\boldsymbol{C}$ it also has an Inverse Wishart distribution. The new parameters are related to the inverse of this matrix, σ_t^2 is the Schur complement of c_t in $\boldsymbol{C}_{(t)}$ and $\boldsymbol{\rho}_t = \boldsymbol{C}_{(t-1)}^{-1}\mathbf{c}_t$ (proofs are in Dawid (1981)).The priors are determined by decomposing $M = \tau\mathbf{I}_T$ conformally with $\boldsymbol{C}$ and are independent of the order of the factorisation. The prior parameterisation for σ_t^2

is thus

$$\sigma_t^2|\tau,\nu \sim IG\left(\frac{\nu-T+t}{2},\frac{\tau}{2}\right), \tag{7.3.10}$$

where the parameters in the bracket refer to the shape and scale respectively, with a gamma hyperprior on τ. The prior for $\boldsymbol{\rho}_t$ given σ_t^2 (Schur complement) is a multivariate normal

$$\boldsymbol{\rho}_t|\sigma_t^2 \sim \mathcal{N}_{T-1}\Big(\mathbf{0},\frac{\sigma_t^2}{\tau}\mathbf{I}_{T-1}\Big) \tag{7.3.11}$$

Each covariate response pair for the unconstrained continuous data has its own independent regression parameter β_{ts}, where the prior is augmented with a latent indicator variable in the form of a "spike-and-slab" (George and McCulloch, 1997) to perform explicit variable selection. The spike is a point mass at 0 (Dirac distribution) with probability $1-p(\gamma_{ts})=1-\omega_s$ and the slab is a zero centred Gaussian with variance w_t. The binary latent indicator variable γ_{ts} represents the inclusion of the sth covariate in the model. We take advantage of the multiple responses by allowing the sparsity parameter ω to vary over the covariate space, an option which is rarely available with a univariate response.

In the case of the categorical data matrix, we are interested in selecting the group of variables associated with the response into the model, rather than a particular level. Each factor variable (or group) $g=1,..,G$ has $j=1,...,m_g,m_g+1$ levels which are coded as dummy variables in $\boldsymbol{W}$ with reference to the intercept. The spike is a point mass at 0 with probability $1-p(\chi_{tg})=1-\varrho_g$ and the slab is a zero centred Gaussian with variance v_t.

As $\boldsymbol{\rho}_t$ can be interpreted as an additional set of regression parameters alongside a design matrix of residuals $\boldsymbol{U}_{t-1}$, a latent variable η_{tk} is augmented to the normal prior for $\boldsymbol{\rho}_t$. η_{tk} reduces the noise in the model by performing a reparameterised form of covariance selection, conveniently bypassing the difficulties which can be encountered when selecting parameters within a positive definite matrix. This approach is an alternative to Gaussian graphical models (Wang, 2015) which allows us to scale up the model to high dimensions whilst imposing sparsity over the reparameterised space and maintaining computational feasibility.

The spike-and-slab priors on the unconstrained data mean parameters are

$$\beta_{ts}|\gamma_{ts}, w_t \sim (\gamma_{ts})\mathcal{N}(0, w_t) + (1-\gamma_{ts})\delta_0, \tag{7.3.12}$$

$$\boldsymbol{\zeta}_{tg}|\chi_{tg}, v_t \sim (\chi_{tg})\mathcal{N}_{m_g}(\mathbf{0}, v_t\mathbf{I}_{m_g}) + (1-\chi_{tg})\delta_0, \tag{7.3.13}$$

$$\rho_{tk}|\sigma^2, \tau, \eta_{tk} \sim (\eta_{tk})\mathcal{N}(0, \sigma_t^2/\tau) + (1-\eta_{tk})\delta_0, \tag{7.3.14}$$

where δ_0 is the Dirac distribution. Each latent indicator variable is assigned an independent Bernoulli prior. The probability that a given covariate in the design matrices $\boldsymbol{X}$, $\boldsymbol{W}$ and $\boldsymbol{U}_{(t-1)}$ affect any response is modelled through parameters ω_s, ϱ_g and κ_j respectively, which are shared across responses. Beta priors are placed on these parameters. The prior variance parameters, which are free to vary across the responses w_t, v_t and σ_t^2 are given inverse gamma hyperpriors with a gamma hyperprior on the respective scales.

7.3.4 Priors on constrained parameters

The convenient form of the likelihood in (7.3.7) allows us to easily extend the prior structure of Scott and Lewin (2021) for a univariate model containing a compositional design matrix, to a multivariate response model with a latent structure $\boldsymbol{C}$.

The linear constraint on the vector of parameters for each response $\boldsymbol{\theta}_t$ is captured by positing the degenerate singular multivariate normal prior

$$\boldsymbol{\theta}_t|\boldsymbol{\mu}_t, \boldsymbol{\psi}_t \sim SMVN_d(\mathbf{T}\boldsymbol{\mu}_t, \mathbf{T}\text{diag}(\boldsymbol{\psi}_t)\mathbf{T}) \tag{7.3.15}$$

where $\mathbf{T} = (\mathbf{I}_d - (1/d)\mathbb{J}_d)$ is an idempotent matrix of rank $d-1$ and $\mathbb{J}$ is a matrix of ones. The addition of a separate variance parameter for each θ_{tj} parameter adds additional flexibility to the model to account for the large differences in the order of magnitude of each row in the compositional design matrix. We augment the prior on $\boldsymbol{\theta}_t$ with dependent latent indicator variables from the

truncated distribution

$$p(\boldsymbol{\xi}_t|\kappa_j) \propto \prod_{j=1} \kappa_j^{\xi_{tj}}(1-\kappa_j)^{1-\xi_{tj}} \ \mathrm{I}\Big[\sum_j \xi_{tj}. \neq 1\Big] \tag{7.3.16}$$

which accounts for the alr transformation by preventing the selection of a single microbe into the model through the indicator function (I[·]).

The full singular multivariate normal spike-and-slab prior for $p(\boldsymbol{\theta}_t|\Sigma_t, \boldsymbol{\xi}_t) = p(\boldsymbol{\theta}_{\xi_t}|\Sigma_t, \boldsymbol{\xi}_t)p(\boldsymbol{\theta}_{\bar{\xi}_t}|\boldsymbol{\xi}_t)$, where $\boldsymbol{\theta}_{\xi_t}$ and $\boldsymbol{\theta}_{\bar{\xi}_t}$ are subvectors of $\boldsymbol{\theta}_t$, is

$$p(\boldsymbol{\theta}_{\xi_t}|\Sigma_t, \boldsymbol{\xi}_t) = \frac{1}{(\det{}^*(2\pi\Sigma^+_{\xi_t}))^{(-1/2)}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}_{\xi_t}\Sigma^+_{\xi_t}\boldsymbol{\theta}_{\xi_t}\right) \quad \text{and} \quad p(\boldsymbol{\theta}_{\bar{\xi}_t} = \mathbf{0}|\boldsymbol{\xi}_t) = 1. \tag{7.3.17}$$

$\Sigma^+_{\xi_t}$ denotes the Moore-Penrose pseudo-inverse (Penrose, 1955) of the matrix $\mathbf{T}_{\xi_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\xi_t}$, a function of the $\mathbf{T}$ matrix and a diagonal matrix of the parameters $\boldsymbol{\psi}_{\xi_t}$, defined by $A^+ = VS^+U^T$ if $A = USV^T$ is the singular value decomposition of A and S^+ is diagonal matrix where $S^+_{ii} = 1/S_{ii}$ for the non-zero diagonal entries of S. $\boldsymbol{\theta}_t$ is the vector of parameters $1 \times d_{\xi_t}$, $\det^*$ denotes the pseudo-determinant defined as the product of the non-zero eigenvalues of the matrix and $\boldsymbol{\xi}_t$ is a vector of zeros and ones. This prior also implies a univariate spike-and-slab on the diagonal elements of the covariance matrix in (7.3.17).

$$p(\boldsymbol{\psi}_t|\boldsymbol{\xi}_t) = \prod_{j=1}^{d}\Big[\frac{b^{a_{\psi_t}}_{\psi_t}}{\Gamma(a_{\psi_t})}(\psi_{tj})^{-a_\psi-1}\exp\{-b_{\psi_t}\psi^{-1}_{tj}\}\Big]^{\xi_{tj}}\delta_0(\psi_{tj})^{1-\xi_{tj}} \quad \psi_{tj} > 0, \ \forall\ j. \tag{7.3.18}$$

Here we place the hierarchical prior directly on each independent scale parameter ψ_{tj}. The specification of the spike-and-slab priors on all the of parameters in the mean of the likelihood in (7.3.6), modifies the conditional normal so that its final form is

$$p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\beta}, \mathbf{C}) = \prod_{t=1}^{T}\psi(\mathbf{y}_t|\boldsymbol{X}_{\gamma_t}\boldsymbol{\beta}_{\gamma_t} + \boldsymbol{W}_{\chi_t}\boldsymbol{\zeta}_{\chi_t} + \boldsymbol{Z}_{\xi_t}\boldsymbol{\theta}_{\xi_t} + \boldsymbol{U}_{(t-1)\eta_t}\boldsymbol{\rho}_{\eta_t}, \sigma^2_t\mathbb{I}_n) \tag{7.3.19}$$

where $\boldsymbol{\gamma}_t$, $\boldsymbol{\chi}_t$, $\boldsymbol{\xi}_t$ and $\boldsymbol{\eta}_t$ are vectors of 0's and 1's which partition the respective design matrices and vectors of parameters.

7.4 Variational Inference Updates

We employ variational inference (e.g. Blei et al. (2017)) with a *mean field variational family* where the latent variables are mutually independent and each governed by a distinct factor in the variational density is used, but dependencies are allowed within each member (block). We define the blocks to ensure the dependency between the latent indicator variable(s) and their associated parameter(s) is captured. The full mean-field approximation distribution is defined as

$$\begin{aligned} q(\boldsymbol{\vartheta}) = &\left\{\prod_t q(\alpha_t)\right\} \times \left\{\prod_t \prod_s q(\beta_{ts}, \gamma_{ts})\right\} \times \left\{\prod_t q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t)\right\} \times \left\{\prod_t \prod_g q(\boldsymbol{\zeta}_{tg}, \chi_{tg})\right\} \times \\ &\left\{\prod_s q(\omega_s)\right\} \times \left\{\prod_j q(\kappa_j)\right\} \times \left\{\prod_g q(\varrho_g)\right\} \times \left\{\prod_t q(\sigma_t^2) \prod_{k<t} q(\rho_{tk}, \eta_{tk}|\sigma_t^2)\right\} \times \\ &\left\{\prod_t q(w_t)\right\} \times \left\{\prod_t q(w_{\alpha_t})\right\} \times \left\{\prod_t q(v_t)\right\} \times q(\lambda) \times q(b_w) \times q(b_v) \times q(\tau), \end{aligned} \tag{7.4.1}$$

with $q(.)$ as the analytical approximation restricted to belong to a class of tractable distributions by the factorization of (7.4.1). The inference is transformed into an optimisation problem where $q(\boldsymbol{\theta})$ is obtained by minimizing its Kullback-Liebler divergence from the target distribution $p(\boldsymbol{\theta}|\boldsymbol{Y})$.

The variational inference updates are available analytically for all parameters and hyperparameters in the model except for the joint update $q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t)$ and are derived in the Supplementary Section. The linked likelihood factorisation for the multiple responses in (7.3.19) alters the free variational parameter updates, directly associated with the multivariate regression. Unlike independent updates, information is borrowed across the responses as q expectations from parameters in the other $T-1$ regressions are now included in the analytical update.

The likelihood factorisation and prior parameterisation of the multivariate response model allows us to conveniently exploit the CAVI-MC approach in Scott and Lewin (2021) for the joint update $q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t)$. The conditional vector update $q(\boldsymbol{\theta}_t|\boldsymbol{\psi}_t, \boldsymbol{\xi}_t)$ is available analytically and takes the form

$$q(\boldsymbol{\theta}_{\xi_t}|\boldsymbol{\xi}_t, \boldsymbol{\psi}_t) = SMVN_{d_{\xi_t}}(\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\theta_{\xi_t}}, \mathbf{T}_{\xi_t}\Sigma_{\theta_{\xi_t}}\mathbf{T}_{\xi_t}), \quad q(\boldsymbol{\theta}_{\bar{\xi}_t} = 0|\boldsymbol{\xi}_t) = \delta_0(\boldsymbol{\theta}_{\bar{\xi}_t}), \tag{7.4.2}$$

where q denotes the probability with respect to the approximating distribution. The updates for the vector $\boldsymbol{\mu}_{\theta_{\xi_t}}$ and matrix $\Sigma_{\theta_{\xi_t}}$

$$\boldsymbol{\mu}_{\theta_{\xi_t}} = \Sigma_{\theta_{\xi_t}}\Big(\boldsymbol{Z}_{\xi_t}^T\Big((\sigma_t^{-2})^{(1)}\Big((\boldsymbol{u}_{t,\not j})^{(1)} - \sum_{k<t}(\boldsymbol{u}_k)^{(1)}(\rho_{tk})^{(1)}\Big) - \sum_{k>t}(\sigma_k^{-2})^{(1)}(\rho_{kt})^{(1)}(\boldsymbol{u}_k)^{(1)} + \\ + \sum_{k>t}\sum_{h<k,h\neq t}(\sigma_k^{-2})^{(1)}(\boldsymbol{u}_h)^{(1)}(\rho_{kh})^{(1)}(\rho_{kt})^{(1)} + \sum_{k>t}(\sigma_k^{-2})^{(1)}(\boldsymbol{u}_{t,\not j})^{(1)}(\rho_{kt})^{(2)}\Big)\Big) \tag{7.4.3}$$

$$\Sigma_{\theta_{\xi_t}} = \left((\mathbf{T}_{\xi_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\xi_t})^+ + (\sigma_t^{-2})^{(1)}\boldsymbol{Z}_{\xi_t}^T\boldsymbol{Z}_{\xi_t} + \sum_{k>t}(\sigma_k^{-2})^{(1)}(\rho_{kt})^{(2)}\boldsymbol{Z}_{\xi_t}^T\boldsymbol{Z}_{\xi_t}\right)^{-1} \tag{7.4.4}$$

where $f(\boldsymbol{z})^{(j)}$ as the j-th moment of $f(\boldsymbol{z})$ with respect to $q(\boldsymbol{z})$, $\mathbb{E}_q\left[f(\boldsymbol{z}^j)\right]$. However, the truncated Bernoulli prior distributions for $\boldsymbol{\xi}_t$ and unique scale parameter ψ_{tj}, for each element of $\boldsymbol{\theta_t}$ and each response, prevent a conjugate posterior update for the joint block $q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t)$. This is proportional to

$$\begin{aligned} q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t) &\propto q(\boldsymbol{\theta}_t|\boldsymbol{\psi}_t, \boldsymbol{\xi}_t)q(\boldsymbol{\psi}_t, \boldsymbol{\xi}_t) \\ &\propto \ \text{SMVN}(\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\theta_{\xi_t}}, \mathbf{T}_{\xi_t}\Sigma_{\theta_{\xi_t}}\mathbf{T}_{\xi_t})\delta_0(\boldsymbol{\theta}_{\bar{\xi}_t})\exp\left(\frac{1}{2}\boldsymbol{\mu}_{\theta_{\xi_t}}^T\mathbf{T}_{\xi_t}(\mathbf{T}_{\xi_t}^T\Sigma_{\theta_{\xi_t}}\mathbf{T}_{\xi_t})^+\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\theta_{\xi_t}} + \right. \\ &\quad + \frac{1}{2}\log\big(\det{}^*(\mathbf{T}_{\xi_t}\Sigma_{\theta_{\xi_t}}\mathbf{T}_{\xi_t})\big) - \frac{1}{2}\log(\det{}^*(\mathbf{T}_{\xi_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\xi_t})) + \sum_j \xi_{tj}(\log\kappa_j)^{(1)} + \\ &\quad + \sum_j(1-\xi_{tj})(\log(1-\kappa_j))^{(1)} + (a_{\psi_t}\log(b_{\psi_t}) - \log(\Gamma(a_{\psi_t}))\sum_j \xi_{tj} + \\ &\quad \left. - \sum_j(a_{\psi_t}+1)\xi_{tj}\log(\psi_{tj}) - b_{\psi_t}\sum_j \xi_{tj}\psi_{tj}^{-1}\right). \end{aligned} \tag{7.4.5}$$

We use the CAVI-MC approach of Scott and Lewin (2021) where a RJMCMC is incorporated into the variational inference updates to sample from $q(\boldsymbol{\psi}_t, \boldsymbol{\xi}_t)$ in (7.4.5) and calculate the intractable marginal expectations over $q(\cdot)$. The birth-death and swap scheme is guided by a univariate approximation of $q(\boldsymbol{\xi_t}, \boldsymbol{\psi}_t)$ relative to the jth element for each response. The proposal distributions for $\boldsymbol{\psi}_t$ are obtained by introducing auxiliary parameters (upper case Greek letters), which are unconstrained versions of the constrained parameters, with a simpler prior parameterisation. The auxiliary parameters create an alternative directed acyclic graph (DAG) which is updated via a

“separate branch" of pseudo updates which helps us to approximate the model in order to guide the MCMC step. These updates are refined at each iteration by the full variational inference updates which account for the constraint.

7.4.1 Algorithm

Co-ordinate ascent variational inference is performed by iterating through the analytical variational updates, maximising the evidence lower bound (ELBO) with respect to each coordinate direction whilst fixing the other coordinate values. For the $q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t)$ updates an MCMC approach samples from the intractable q density to obtain Monte Carlo estimates of the expectations where the proposal probabilities for the sampling scheme are a function of the data and the variational parameters, and are updated at each iteration of the co-ordinate ascent variational inference.

For each run we compute the evidence lower bound , (derived in the Supplementary Section) with the updated free parameters, until this converges to the local optimum. The ELBO is no longer smooth because of the Monte Carlo variability, but we are able to declare convergence when the random fluctuations are small around a fixed point. The implementation of the overall approach is described in Algorithm 7, with the MCMC move detailed in Algorithm 8.

Algorithm 7: CAVI-MC for variable selection.

Input : A model $p(\boldsymbol{Y}, \boldsymbol{\vartheta})$, a data set $\{\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Z}\}$. Number of Monte Carlo samples m.
Output : Variational densities $q(\boldsymbol{\vartheta}_{-(\theta,\psi,\xi)}) = \prod_j q_j(\vartheta_j)$ and Monte Carlo expectations.
Intialize: First and second order raw moments of the variational factors, prior hyperparameters.

for k = *1,...,K* **do**

  for j = *1,...,J* **do**

    Set $q_j(\vartheta_j) \propto \exp\{\mathbb{E}_{-j}[\log p(\vartheta_j|\boldsymbol{\vartheta}_{-j}, \mathbf{Y})]\}$

  end

  for t = *1,...,T* **do**

    Calculate the arguments for proposal distribution for $\boldsymbol{\psi}_t$ from the psuedo variational updates.

$$a^*_{\Delta_{tj}} = \frac{1}{2}(\Upsilon_{tj})^{(1)} + a_{\Delta_t} \qquad b^*_{\Delta_{tj}} = \frac{1}{2}(\Omega_{tj})^{(1)} + b_{\Delta_t}$$
$$\psi_{tj} \sim IG_q(a^*_{\Delta_{tj}}, b^*_{\Delta_{tj}})$$

    Calculate the probabilites $\tilde{p}(\boldsymbol{\xi}_t|\boldsymbol{\vartheta})$ for the $\boldsymbol{\xi}_t$ proposal (by approximating $q(\boldsymbol{\xi}_t|\boldsymbol{Y})$ and normalising) in the RJMCMC.

$$\tilde{p}(\xi_{tj} = |\boldsymbol{\vartheta}) \equiv \Bigg[\exp\Bigg\{ (\log(1-\kappa_j))^{(1)} - \frac{1}{2}\log\big(\bar{\sigma}^2_{\theta,tj}\big) + \frac{1}{2}(\log\psi_{tj})^{\{1\}}_{\emptyset} - (\log\kappa_j)^{(1)} +$$
$$+ (\log\Gamma(a_{\psi_t}) - a_{\psi_t}\log b_{\psi_t}) + (a_{\psi_t}+1)(\log\psi_{tj})^{\{1\}}_{\emptyset} + b_{\psi_t}(\psi^{-1}_{tj})^{\{1\}}_{\emptyset} \Bigg\} +$$
$$- \frac{1}{2\bar{\sigma}^2_{\theta,tj}} \left((1 - 1/\{d_{\xi_t}\}^{\{1\}})(\{\mu_{\theta_{tj}}\}^{\{1\}}_{\emptyset})^2 - \frac{2}{\{d_{\xi_t}\}^{\{1\}}}\{\mu_{\theta_{\xi_{tj}}}\}^{\{1\}}_{\emptyset} \sum_{j' \neq j} \{\mu_{\theta_{\xi_{tj'}}}\}^{\{1\}} \right) + 1 \Bigg]^{-1}$$

  end

  Perform MCMC step: Algorithm 8.

  return $\mathbb{E}_q(\boldsymbol{\xi}_t|\boldsymbol{Y})^{[k]}$, $\mathbb{E}_q(\boldsymbol{\psi}_t|\boldsymbol{Y})^{[k]}$, $\mathbb{E}_q(\boldsymbol{\theta}_t|\boldsymbol{Y})^{[k]}$, $\mathbb{E}_q(\boldsymbol{\theta}^T_{\xi_t}\boldsymbol{Z}^T_{\xi_t}\boldsymbol{Z}_{\xi_t}\boldsymbol{\theta}_{\xi_t}|\boldsymbol{Y})^{[k]}$ *and cross product terms in the ELBO calculation.*

  Compute ELBO.

end

return $\Big(q(\boldsymbol{\vartheta}_{-(\theta,\psi,\xi)})$, $\mathbb{E}_q(\boldsymbol{\xi}_t|\boldsymbol{Y})$, $\mathbb{E}_q(\boldsymbol{\psi}_t|\boldsymbol{Y})$, $\mathbb{E}_q(\boldsymbol{\theta}_t|\boldsymbol{Y})\Big)$ $\forall$ t.

Algorithm 8: MCMC step for CAVI-MC.

Input: k current loop of CAVI-MC, q expectations, pseudo VB updates, normalised approximate marginal probability $p(\boldsymbol{\xi}_t|\boldsymbol{\vartheta})$.

for *t = 1, ..., T* **do**

  for *i = 1,...,m* **do**

    if *Between model move proposed* **then**

      Given the current position of the variational samples $\boldsymbol{\xi}_t$, $\boldsymbol{\psi}_{\xi_t}$ and $\boldsymbol{\theta}_{(\psi_t,\xi_t)}$, propose either a birth-death move or swap move. Propose a new model with probability

$$j_m(\boldsymbol{\xi}_t, \boldsymbol{\xi}_t') \propto p(\boldsymbol{\xi}_t|(\log(1-\kappa_j))^{(1)}, (\log\kappa_j)^{(1)}, (\boldsymbol{\mu}_{\theta_{\xi_t}})_{\emptyset}^{\{1\}[k-1]}, (\boldsymbol{\psi}_t^{-1})^{\{1\}[k-1]}, (\log\boldsymbol{\psi}_t)^{\{1\}[k-1]}, (\bar{\sigma}_{\theta_t}^2), (d_{\xi_t})^{\{1\}[k-1]}).$$

$$\pi(\boldsymbol{\psi}_t'|\boldsymbol{\xi}_t', a_{\Delta_{tj}}^*, b_{\Delta_{tj}}^*) = \prod_j \left[IG_q\Big(\psi_{tj}|\frac{1}{2}(\Upsilon_{tj})^{(1)} + a_{\Delta_t}, \frac{1}{2}(\Omega_{tj})^{(1)} + b_{\Delta_t}\Big)\right]^{\xi_{tj}'} \delta_0(\psi_{tj}')^{1-\xi_{tj}'}.$$

$$\boldsymbol{\theta}'_{(\psi_t,\xi_t)} \sim \text{SMVN}_{d'_{\xi_t}}((\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\theta_{\xi_t}})', (\mathbf{T}_{\xi_t}\Sigma_{\theta_{\xi_t}}\mathbf{T}_{\xi_t})'|\boldsymbol{\psi}_t', \boldsymbol{\xi}_t', \mathbf{Z}, (u_{tj})^{(1)}, (\sigma_t^{-2})^{(1)}).$$

      The acceptance probability is

$$\alpha_{tb} = \min\left\{\frac{q(\boldsymbol{\psi}_t', \boldsymbol{\xi}_t'|\boldsymbol{Y})j_m(\boldsymbol{\xi}_t', \boldsymbol{\xi}_t)\pi(\boldsymbol{\psi}_t|\boldsymbol{\xi}_t, a_{\Delta_t}^*, b_{\Delta_t}^*)}{q(\boldsymbol{\psi}_t, \boldsymbol{\xi}_t|\boldsymbol{Y})j_m(\boldsymbol{\xi}_t, \boldsymbol{\xi}_t')\pi(\boldsymbol{\psi}_t'|\boldsymbol{\xi}_t', a_{\Delta_t}^*, b_{\Delta_t}^*)}, 1\right\}$$

      with the target density simplifed to:

$$q(\boldsymbol{\xi}_t, \boldsymbol{\psi}_t|\boldsymbol{Y}) \propto \exp\left(\frac{1}{2}(\boldsymbol{\mu}_{\theta_{(\xi_t,\psi_t)}}^T\mathbf{T}_{\xi_t}(\mathbf{T}_{\xi_t}^T\Sigma_{\theta_{(\xi_t,\psi_t)}}\mathbf{T}_{\xi_t})^+\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\theta_{(\xi_t,\psi_t)}}) + \frac{1}{2}\log\Big(\det{}^*(\mathbf{T}_{\xi_t}\Sigma_{\theta_{(\xi_t,\psi_t)}}\mathbf{T}_{\xi_t})\Big)\right.$$
$$\sum_j \xi_{tj}(\log\kappa_j)^{(1)} - \frac{1}{2}\log(\det{}^*(\mathbf{T}_{\xi_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\xi_t})) + \sum_j(1-\xi_{tj})(\log(1-\kappa_j))^{(1)} +$$
$$\left. - (a_{\psi_t}+1)\sum_j \xi_{tj}\log(\psi_{tj}) - b_{\psi_t}\sum_j \xi_{tj}\psi_{tj}^{-1} + (a_{\psi_t}\log(b_{\psi_t})) - \log(\Gamma(a_{\psi_t}))\sum_j \xi_{tj}\right).$$

      for *l=1,...,L* **do**

        Perform within-model moves: Given the current position of the variational samples $\boldsymbol{\xi}_t$, $\boldsymbol{\psi}_t$ and $\boldsymbol{\theta}_t$, draw proposals $\boldsymbol{\psi}_t'|\boldsymbol{\xi}_t$ and $\boldsymbol{\theta}_t'|\boldsymbol{\psi}_t', \boldsymbol{\xi}_t$ using the same distributions as the between model move.

        Proposed move accepted with probability

$$\alpha_{tw} = \min\left\{\frac{q(\boldsymbol{\psi}_t', \boldsymbol{\xi}_t|\boldsymbol{Y})\pi(\boldsymbol{\psi}_t|\boldsymbol{\xi}_t, a_{\Delta_t}^*, b_{\Delta_t}^*)}{q(\boldsymbol{\psi}_t, \boldsymbol{\xi}_t|\boldsymbol{Y})\pi(\boldsymbol{\psi}_t'|\boldsymbol{\xi}_t, a_{\Delta_t}^*, b_{\Delta_t}^*)}, 1\right\}.$$

      end

    else

      for *l=1,...,L* **do**

        Perform within-model moves with probability α_{tw}.

      end

    end

  end

end

7.5 Data application

We apply our proposed method to a subset of the "Know your Heart" cross-sectional study of cardiovascular disease (Cook et al., 2018). Information on age, sex, alcohol consumption, diet quality, education level and smoking status was obtained from 685 men and women of the Arkhangelsk branch, aged between 35 and 69 years and recruited from the general population, by a baseline questionnaire. A CAGE score (Demmie et al., 2015) for detecting problem drinking (labelled as *alcohol* in the Figures, where as total alcohol consumption is *totalvol*) was derived from the answers. BMI was calculated from information collected at a physical examination.

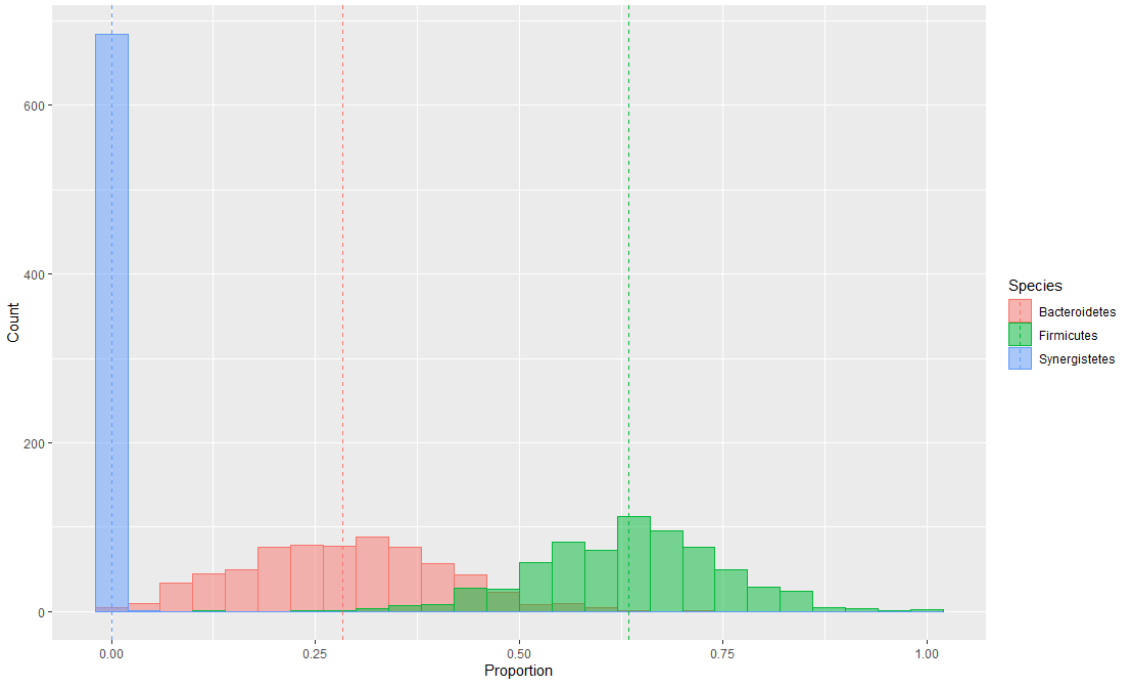

Figure 7.5.1: Histogram of the standardised OTU counts, for the gut microbiome species of Bacteroidetes, Firmicutes and Synergistetes. Their respective means of 0.2840, 0.6337 and 0.0004, are represented by the dashed lines. There are large differences in the size of the proportions for Synergistetes, compared with Bacteroidetes.

Participants of the sudy were asked to volunteer faecal samples for analysis of the gut microbiome as part of the study. 16S rRNA sequencing (using the variable regions V3-V4) followed by taxonomic classification using a Naive Bayes Classifier (Bokulich et al., 2018) was used to determine the relative abundances of the microbes. There are very large differences in the orders of

magnitude of the standardised OTU counts, illustrated by the histogram for the bacterial phyla of Bacteroidetes, Firmicutes and Synergistetes in Figure 7.5.1. The large blue column for Synergistetes highlights that many of the proportions are close to 0, in stark contrast to the distribution of values for Bacteroidetes. The mean Synergistetes value is just 0.0004, compared with a mean of 0.6337 for Firmicutes.

Our response matrix $\boldsymbol{Y}$ is the core set of biochemistries analysed using the blood and urine samples, listed in Table 7.1. Two data points were removed due to missing values, under the assumption of "missing at random". Each response is logged because of the positive skew of many of the responses. The correlation plot of the empirical residuals after independent univariate regressions of the log responses (left plot in Figure 7.5.2), highlights the dependency between many of the biomarkers and the importance of our likelihood specification which is able to capture this latent structure. This is particularly obvious between the first six of the lipid biomarkers and the liver function tests in Table 7.1.

Table 7.1: Core set of biological analyses on blood and urine samples with labels used in Figures. Unit is mmol/L unless specified.

Biomarker Target	Label	Specific Measure	Biological Sample
Lipid Metabolism	apoa1	Apolipoprotein A1 g/L	Serum
	apob	Apolipoprotein B g/L	Serum
	hdl	High Density Lipoprotein Cholesterol (HDL)	Serum
	ldl	Low Density Lipoprotein Cholesterol (LDL)	Serum
	trig	Triglycerides	Serum
	lpa	Lp(a) mg/dl	Serum
Renal Function	crea_s	Creatinine	Serum
	crea_u	Creatinine	Urine
	cyc	Cystatin C mg/L	Serum
	malb	Albumin mg/L	Urine
Inflammatory Markers	crphs	High sensitivity C reactive protein mg/L	Serum
Metabolites	thb	Haemoglobin A1c	Whole Blood
Iron Pathways	trf	Transferrin g/L	Serum
Liver function tests	alt	Alanine transaminase (ALT) U/L	Serum
	ast	Aspartate transaminase (AST) U/L	Serum
	ggt	Gamma-glutamyl transferase (GGT) U/L	Serum
Caridiac Micronecrosis	bnp	NT-Pro-B-type Natriuretic Peptide pg/ml	Serum
	tropt	High sensitivity Troponin T pg/L	Serum

The microbiome taxa at either the phylum or genus level are included in the model alongside the unconstrained covariates. As is common in abundance data, taxon which has more than 93% of zeros is removed to protect against taxa with a small mean and a trivially large coefficient of variation. The counts are transformed into relative abundances after adding a small constant of 0.5 to replace the zero counts (Aitchison, 2003) and then log-transformed. Each column is centred and divided by the standard deviation across all of the mean centred log compositional data. This scales the data whilst respecting the sum to zero constraint of each $\boldsymbol{\theta}_t$ vector (7.3.4). All continuous unconstrained covariates are standardised and the dummy variables for the categorical covariates in the design matrix are coded relative to a reference level. In the case of smoking, a three level categorical smoking covariate is determined with non-smoker as the reference level and ex-smoker (smoke1), regular smoker (smoke2) and less than 1 cigarette a day (smoke3) as the respective levels. Vague priors are placed on the hyperparameters and q expectations are initialised by randomly sampling from the prior distributions, which ensures different starting points for each run of the algorithm. Four separate runs are performed with 30 VI iterations each to check for multi-modality of the posterior space, as the CAVI-MC converges to a local optima. The initial number of between-model MCMC iterations is set to 5000, before 10000 iterations are performed after the 5th iteration of the variational updates and the ELBO is monitored to confirm convergence.

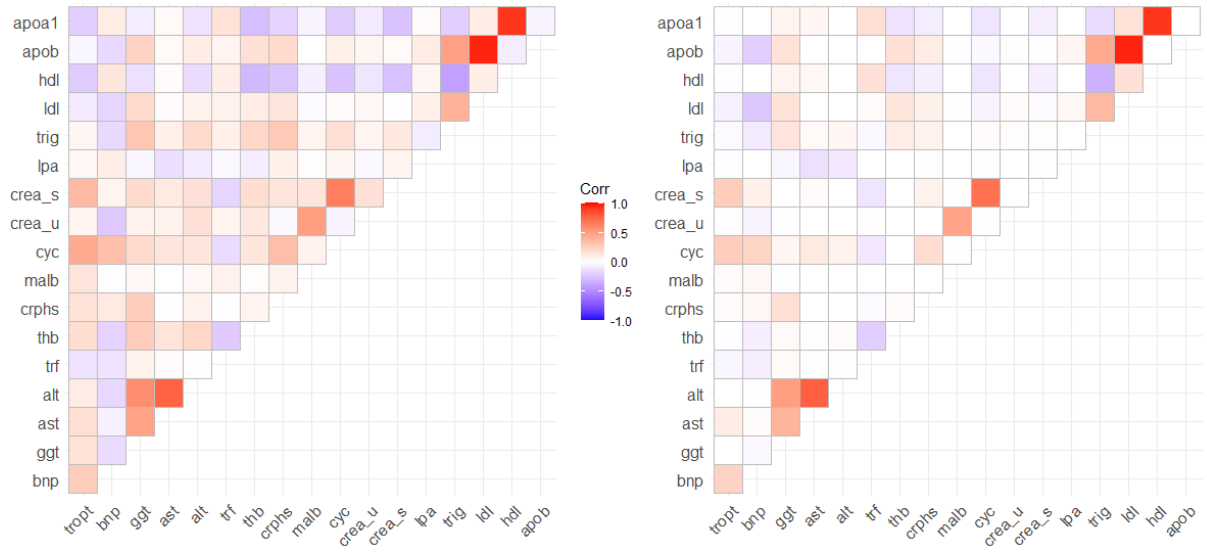

Figure 7.5.2: Left plot is the empirical residual correlation obtained from independent univariate regressions. Right plot is the upper triangular correlation matrix from the empirical residual correlation matrix $\boldsymbol{C}$ in the original model parameterisation after shrinkage. Most low correlations have been shrunk to 0. Red blocks represent a strong positive correlation, blue blocks represent a strong negative correlation. The labels in the x and y axes are defined in Table 7.1.

For each run, despite different starting points, the CAVI-MC converges on to the same optimum. Although 30 variational iterations are performed the algorithm converges after approximately 10 iterations. This can be observed from the plot of the ELBO (Figure 7.5.3) from the first run, with only very small increases after the 10th iteration. Despite the MCMC component, the large number of MCMC sampler iterations which are averaged over in the CAVI calculation, the ELBO remains monotonically increasing (Figure 7.5.3).

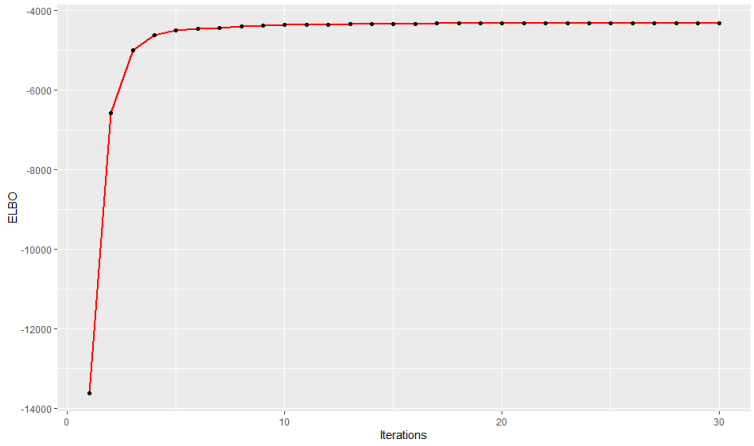

Figure 7.5.3: Plot of the ELBO against iterations for the CAVI-MC applied to the "Know Your Heart" data set with the microbiome grouped at the phylum level. 30 iterations are performed, with 30,000 between state space moves by the RJMCMC after 4 iterations. The approximately flat line after 10 iterations implies that the model has reached convergence. Despite the MCMC component removing the monotonic properties of the ELBO, the fluctuations are relatively small.

As the model is unable to identify any taxa of interest with the matrix of log responses at the genus level, once thresholding is performed, only the results at the phylum level are discussed. The marginal expectations of the approximate posterior distribution of inclusion $((\gamma_{ts})^{(1)}$, $(\chi_{tg})^{(1)}$ and $(\xi_{tj})^{(1)})$ and effect size $((\beta_{ts})^{(1)}$, $(\boldsymbol{\zeta}_{tg})^{(1)}$ and $(\theta_{tj})^{(1)})$ for each covariate against the respective response are plotted as heat maps in Figure 7.5.4 and Figure 7.5.5 respectively. Any effect size is accompanied by the respective standard deviation variational parameter estimate from the approximating density, (for $(\beta_{ts})^{(1)}$ this is $\sigma^2_{\beta,ts}$). All the continuous covariates have been standardised, so their respective shrunken parameter estimates (marginal expectation) represent a change in the log response from an increase of one standard deviation. For each response, the parameter with the largest absolute value identifies the covariate with the largest effect size (e.g. sex for albumin).

The estimated covariance matrix $\boldsymbol{C}$ in the SUR model (7.3.5), can be obtained from the q approximating densities of $\boldsymbol{\rho}_t$ and σ^2_t. The marginal expectation $(\rho_{ts})^{(1)}$ is available directly from the CAVI-MC updates and is an average weighted by the probability of model inclusion $\mathbb{E}_q[\eta_{tk}]$,

$$\begin{aligned}\mathbb{E}_q[\rho_{tk}] = (\rho_{tk})^{(1)} &= \mathbb{E}_q[\mathbb{E}_q[\rho_{tk}|\eta_{tk}]] \\ &= \mu_{\rho_{tk}}(\eta_{tk})^{(1)} + 0(1-(\eta_{tk})^{(1)}). \end{aligned} \tag{7.5.1}$$

The marginal expectation of σ_t^2 can be calculated from the variational free parameters $a^*_{\sigma^2,t}$ and $b^*_{\sigma^2,t}$

$$\mathbb{E}_q[\sigma_t^2] = (\sigma_t^2)^{(1)} = \frac{b^*_{\sigma^2,t}}{a^*_{\sigma^2,t} - 1} \qquad a^*_{\sigma^2,t} > 1. \tag{7.5.2}$$

To see how the covariance feature selection priors in (7.3.14) shrink the off-diagonal elements, the matrix can be recovered by using the variational expectations $(\sigma_t^2)^{(1)}$ and $(\rho_{tk})^{(1)}$ and iteratively solving for the elements in $\boldsymbol{C}$. Figure 7.5.2 displays the correlations of the residuals after mean conditioning before and after shrinkage. The overall effect of the latent indicator variable η_{tk} is to shrink many of the smaller correlations, whilst retaining the stronger correlations in the model. In terms of the mean squared error of a future value (where the expectation is with respect to the data), the shrinkage from the latent indicator variables adds bias to the model estimation, in return for a large reduction in model estimation variance, to ensure the model is generalisable.

By thresholding the marginal probability of inclusion at 0.5 to declare a significant association, the covariates of age, BMI and sex have the largest overall "effect" on the set of responses. We subsequently compare our findings with the literature where analysis has been performed on a univariate response, without accounting for the correlation across the biomarkers.

We find gamma-glutamyl transferase, a liver function test, is associated with alcohol consumption alongside sex and BMI. However, for aspartate transaminase and alanine transaminase, the two other biomarkers for liver function, an association with alcohol is absent (either from alcohol consumption or the 4 level categorical alcohol variable derived from the CAGE score to identify problem drinking). The positive association detected between smoking and cystatin C, with regular smoking (smoke2) having the largest effect (0.0986 ± 0.0115), has been documented by Funamoto et al. (2019) and Drummond et al. (2017). As in our analysis, Drummond et al. (2017) control for age and sex, however we also account for BMI which has a positive effect on the log response. A finding replicated in Muntner et al. (2008).

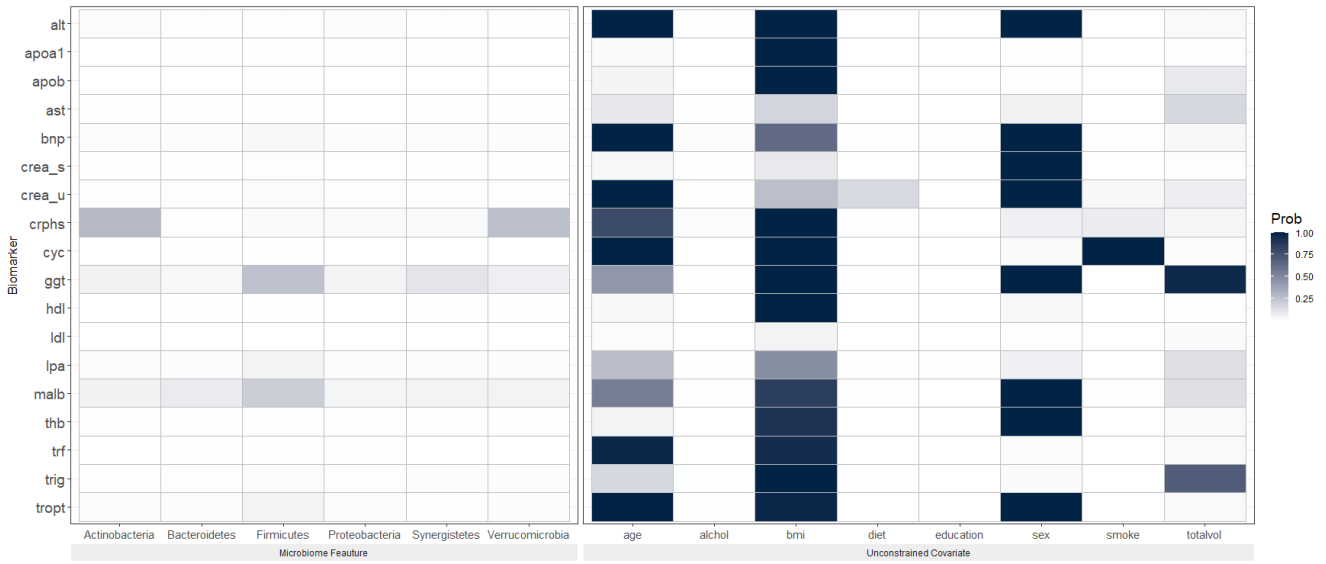

Figure 7.5.4: Heatmap of the marginal posterior probability of inclusion for the compositional and unconstrained covariates. Thresholding at 0.5 prevents a declaration of association between any of the microbiome features and the responses. Darker shades represent a higher probability of inclusion.

We find BMI to be associated with all the measures that characterise the lipid profile except lipoprotein (a), which has a 0.4740 marginal probability of association. This correlation between BMI and lipoprotein levels, especially low density lipoprotein, has been proposed to be a strong contributing risk factor for cardiovascular diseases in obese individuals. Our findings of a positive association with log low density lipoprotein and log triglycerides, and a negative association with log high density lipoprotein are common in the literature (Sandhu et al., 2008). Despite expecting all three to be associated with BMI, often studies which treat the lipids as independent only find significance for a subset (Shamai et al., 2011).

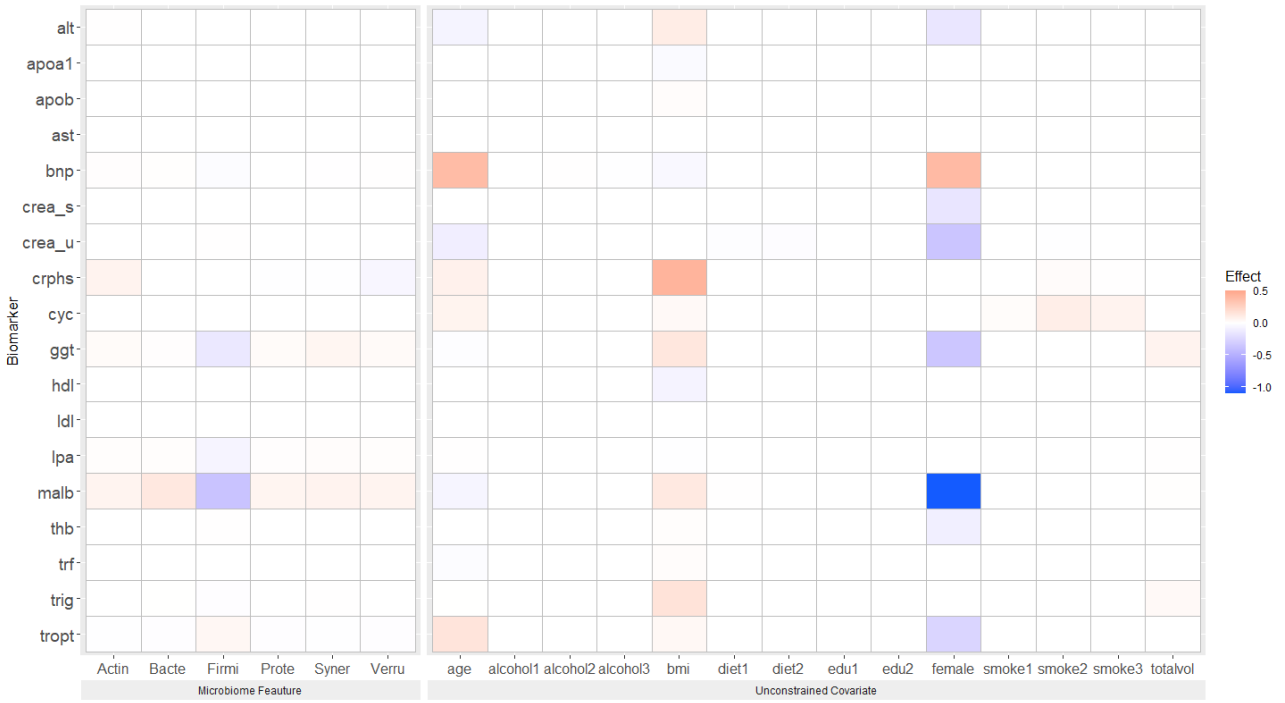

Figure 7.5.5: Heatmap of the marginal effect size for the compositional and unconstrained covariates on the log responses. The microbiome labels for the species at the Phylum level have been reduced to their first 5 letters. The female category has the largest negative effect, a (-1.090 ± 0.1292) difference on log albumin. Standardised BMI has the largest positive effect, a (0.4210 ± 0.0368) increase on log high sensitivity C reactive protein.

We are unable to declare any significant associations between the gut microbiota and the responses. An increase in *Actinobacteria* and a decrease in *Verrucomicrobia* leads to an increase in the inflammatory biomarker high sensitivity C reactive protein. However, there remains a large amount of uncertainty in this relationship, as the probability of inclusion is 0.2860 and 0.2580 respectively. The selection of two compositional covariates should not be confused with the constraint imposed on the latent indicator variable (7.3.16) by the log transformation in (7.3.4) which prevents the selection of a single microbe in the model. Although the constraint does apply for any move made in the RJMCMC, the constraint does not apply to the marginal q posterior distribution. However, if the compositional space is small, the effect of the constraint is much more noticeable, since there are much fewer two variable combinations.

The model detects a possible negative association between Firmicutes and both albumin and gamma-glutamyl transferase but the marginal probability of inclusion of 0.2000 and 0.2481 re-

spectively, prevents any declaration of a significant association. Similar associations have been found to be significant in other studies. High gamma-glutamyl transferase is an indicator of liver disease, which is well known to be accompanied by reductions in Firmicutes (Chen et al., 2011). Low albumin levels are indicative of a decline in kidney function, which are associated with increases in Firmicutes (Hobby et al., 2019).

7.6 Discussion

Our model extends the Bayesian hierarchical linear log-contrast model for compositional data framework in Scott and Lewin (2021) to multi-dimensional phenotypes which are related through a latent structure. Variable selection priors exploit the expected sparsity and allow the associated variables to vary across the responses. The reparameterisation of matrix normal likelihood alongside feature selection allows the model to accommodate either sparse or dense residual covariance structures. A hierarchical prior framework enables the leveraging of information across responses within the model, aiding identification of important covariates. The approach should facilitate research in the relationship between compositional data and multivariate phenotypes.

Current literature suggests a possible sex difference in the gut microbiome at the phylum level (Haro et al. (2016), Dominianni et al. (2015)). Koliada et al. (2021) identify the relative abundances of Firmicutes and Actinobacteria to be increased, while Bacteroidetes was decreased in females compared to males. The model is easily adapted to account for this type of interaction between a categorical covariate and the compositional data, by including an additional compositional design matrix for each level of the covariate. However as the model grows in complexity, the computational burden increases, particularly as one moves down the taxonomic rank for the classification of species in the microbiome.

The model offers an opportunity to further investigate the relationship between gut microbiome an short chain fatty acids. Short chain fatty acids play a critical role in the interplay between diet, the gut microbiota and downstream activation or inhibition of inflammatory cascades such as gas-

trointestinal tract inflammation and inflammatory bowl diseases (ulcerative colitis and Chrohn's disease) (Bander et al., 2020). Our model may provide additional insight as it would use their latent structure to help identify the compositional covariates associated with each response.

The model has not been adapted to account for a strong correlation across the microbiome design matrix $\boldsymbol{Z}$. This was shown to effect the performance of the univariate response model for large datasets with a low signal-to-noise ratio and high correlation in Scott and Lewin (2021). A Markov Random Field prior (Chen and Welling, 2012) can impose structure on the latent indicator variable $\boldsymbol{\xi}_t$ which could potentially improve identification of the constrained covariates. This prior was used to incorporate the phylogenetic relationship among the bacterial taxa by Zhang et al. (2020) in a model that partially accounted for the constraint imposed on the parameters by a compositional transformation. Alternatively, a Dirichlet Process prior (Curtis and Ghosh, 2011) may be used to account for the correlation by using the information within the design matrix. This avoids having to pre-define the structure of the taxa. In both cases, the CAVI-MC element which is incorporated into the variational inference approach, permits very flexible priors for the compositional selection priors.

Acknowledgments

This work was supported by the UK Medical Research Council grant MR/N013638/1 and, MR/M013138/1 "Methods and tools for structural models integrating multiple high-throughput omics data sets in genetic epidemiology". The approach is applied to data from the the Know Your Heart study, a component of International Project on Cardiovascular Disease in Russia (IPCDR) and funded by Wellcome Trust Strategic Award [100217], UiT The Arctic University of Norway (UiT), Norwegian Institute of Public Health, and Norwegian Ministry of Health and Social Affairs. The funding bodies had no role in the design of the study, data collection, analysis, interpretation of data, or in writing the manuscript. *Conflict of Interest*: None declared.

7.7 Supplementary Material

7.7.1 CAVI-MC closed-form updates

This section contains all of the variational inference updates for the CAVI-MC.

Parameterisation

The prior parameterisation is defined below, where the indexes (g, h, j, k, l, s, t) assign unique variables per index. The full prior parameterisation with covariate and covariance variable selection is:

$$p(\boldsymbol{y}_t|\cdot) = \frac{1}{(2\pi\sigma_t^2)^{-n/2}} \exp\left\{-\frac{1}{2\sigma_t^2}\left\|\boldsymbol{y}_t - \alpha_t \mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}_t - \boldsymbol{W}\boldsymbol{\zeta}_t - \boldsymbol{Z}\boldsymbol{\theta}_t - \sum_{k<t}\boldsymbol{u}_k\rho_{tk}\right\|^2\right\} \tag{7.7.1}$$

$$p(\alpha_t|w_{\alpha_t}) = (2\pi w_{\alpha_t})^{-1/2} \exp\left\{-\frac{1}{2w_{\alpha_t}}\alpha_t^2\right\} \tag{7.7.2}$$

$$p(\beta_{ts}|\gamma_{ts}, w_t) = \left[(2\pi)^{-1/2}(w_t)^{-1/2}\exp\left\{-\frac{1}{2w_t}\|\beta_{ts}\|^2\right\}\right]^{\gamma_{ts}} \delta_0(\beta_{ts})^{1-\gamma_{ts}} \quad \beta_{ts} \in \mathbb{R}^1 \tag{7.7.3}$$

$$p(\gamma_{ts}|\omega_s) = \omega_s^{\gamma_{ts}}(1-\omega_s)^{1-\gamma_{ts}} \quad \gamma_{ts} \in \{0, 1\} \tag{7.7.4}$$

$$p(\boldsymbol{\zeta}_{tg}|\chi_{tg}, v_t) = \left(\frac{1}{(2\pi v_t)^{m_g/2}}\exp\left(-\frac{1}{2v_t}\boldsymbol{\zeta}_{tg}^T\boldsymbol{\zeta}_{tg}\right)\right)^{\chi_{tg}} \delta_0(\boldsymbol{\zeta}_{tg})^{1-\chi_{tg}} \tag{7.7.5}$$

$$p(\chi_{tg}|\varrho_g) = \varrho_g^{\chi_{tg}}(1-\varrho_g)^{1-\chi_{tg}} \tag{7.7.6}$$

$$p(\boldsymbol{\theta}_t|\boldsymbol{\xi}_t, \Sigma_t(\boldsymbol{\psi}_{\xi_t}, \mathbf{T})) = \frac{1}{\det^*(2\pi\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t}^T)^{(-1/2)}}\exp\left(-\frac{1}{2}(\theta_{\boldsymbol{\xi}_t})^T(\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\xi_t}^{\mathbf{T}})^+(\boldsymbol{\theta}_{\xi_t})\right)\delta(\theta_{\bar{\xi}_t}) \tag{7.7.7}$$

$$p(\boldsymbol{\xi}_t) \propto \prod_{j=1}\kappa_j^{\xi_{tj}}(1-\kappa_j)^{1-\xi_{tj}}\mathrm{I}\Big[\sum_j \xi_{tj} \neq 1\Big] \quad \boldsymbol{\xi}_t \in \mathbb{N}_{\leq d, \neq 1} \tag{7.7.8}$$

$$p(\boldsymbol{\psi}_t|\boldsymbol{\xi}_t) = \prod_{j=1}^{d}\left[\frac{b_{\boldsymbol{\psi}_t}^{a_{\boldsymbol{\psi}_t}}}{\Gamma(a_{\boldsymbol{\psi}_t})}(\psi_{tj})^{-a_{\boldsymbol{\psi}_t}}\exp\{-b_{\boldsymbol{\psi}_t}\psi_{tj}^{-1}\}\right]^{\xi_{tj}}\delta_0(\psi_{tj})^{1-\xi_{tj}} \tag{7.7.9}$$

$$p(\rho_{tk}|\sigma_t^2,\tau,\eta_{tk}) = \left[\frac{1}{\sqrt{2\pi}}\left(\frac{\tau}{\sigma_t^2}\right)^{\frac{1}{2}}\exp\left\{-\frac{\tau}{2\sigma_t^2}\rho_{tk}^2\right\}\right]^{\eta_{tk}}\delta_0(\rho_{tk})^{1-\eta_{tk}} \quad \rho_{tk}\in\mathbb{R}^1 \tag{7.7.10}$$

$$p(\eta_{tk}|\lambda) = \lambda^{\eta_{tk}}(1-\lambda)^{1-\eta_{tk}} \quad \eta_{tk}\in\{0,1\} \tag{7.7.11}$$

$$p(\sigma_t^2|\tau,\nu) = \frac{1}{\Gamma\left(\frac{\nu-T+t}{2}\right)}\left(\frac{\tau}{2\sigma_t^2}\right)^{\frac{\nu-T+t}{2}}\frac{1}{\sigma_t^2}\exp\left\{-\frac{\tau(\sigma_t^2)^{-1}}{2}\right\} \quad \sigma_t^2>0 \tag{7.7.12}$$

The prior distribution on the hyperparameters is

$$p(w_{\alpha_t}|a_\alpha,b_\alpha) = \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)}(w_{\alpha_t})^{-a_\alpha-1}\exp\left(-b_\alpha w_{\alpha_t}^{-1}\right) \tag{7.7.13}$$

$$p(v_t|a_v,b_v) = \frac{b_v^{a_v}}{\Gamma(a_v)}(v_t)^{-a_v-1}\exp\{-b_v v_t^{-1}\} \quad v>0 \tag{7.7.14}$$

$$p(w_t|a_w,b_w) = \frac{b_w^{a_w}}{\Gamma(a_w)}(w_t)^{-a_w-1}\exp\{-b_w w_t^{-1}\} \quad w_t>0 \tag{7.7.15}$$

$$p(\omega_s|a_\omega,b_\omega) = \frac{1}{B(a_\omega,b_\omega)}\omega_s^{a_\omega-1}(1-\omega_s)^{b_\omega-1} \quad 0\le\omega_s\le 1 \tag{7.7.16}$$

$$p(\kappa_j) = \frac{1}{B(a_\kappa,b_\kappa)}\kappa_j^{a_\kappa-1}(1-\kappa_j)^{b_\kappa-1} \quad 0\le\kappa_j\le 1 \tag{7.7.17}$$

$$p(\varrho_g) = \frac{1}{B(a_\varrho,b_\varrho)}\varrho_g^{a_\varrho-1}(1-\varrho_g)^{b_\varrho-1} \quad 0\le\varrho_g\le 1 \tag{7.7.18}$$

$$p(b_\alpha) = \frac{b_{b_\alpha}^{a_{b_\alpha}}}{\Gamma(a_{b_\alpha})}(b_\alpha)^{a_{b_\alpha}-1}\exp(-b_{b_\alpha}b_\alpha) \tag{7.7.19}$$

$$p(b_v) = \frac{b_{bv}^{a_{bv}}}{\Gamma(a_{bv})}(b_v^{a_{bv}-1})\exp\{-b_{bv}b_v\} \quad b_v>0 \tag{7.7.20}$$

$$p(\lambda) = \frac{1}{B(a_\lambda,b_\lambda)}\lambda^{a_\lambda-1}(1-\lambda)^{b_\lambda-1} \quad \lambda>0 \tag{7.7.21}$$

$$p(\tau) = \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)}(\tau)^{a_\tau-1}\exp\{-b_\tau\tau\} \quad \tau>0 \tag{7.7.22}$$

The joint posterior is

$$
\begin{aligned}
p(y,\theta) = & \left\{\prod_t p(\boldsymbol{y}_t|\mathbf{X},\mathbf{W},\mathbf{Z},\mathbf{U}_t,\alpha_t,\boldsymbol{\beta}_t,\boldsymbol{\zeta}_t,\boldsymbol{\theta}_t,\sigma_t^2,\boldsymbol{\rho}_t)\right\} \times \left\{\prod_t p(\sigma_t^2|\tau,\nu)\prod_{k<t} p(\rho_{tk}|\sigma_t^2,\tau,\eta_{tk})\right\} \times \\
& \left\{\prod_t p(\boldsymbol{\theta}_t|\Sigma_t(\boldsymbol{\psi}_{\xi_t},\mathbf{T}),\boldsymbol{\xi}_t) \times \prod_t p(\boldsymbol{\psi}_t|\boldsymbol{\xi}_t) \times \prod_t p(\boldsymbol{\xi}_t|\kappa_j)\right\} \times \\
& \left\{\prod_t\prod_g p(\boldsymbol{\zeta}_{tg}|v_t,\chi_{tg}) \times \prod_t\prod_s p(\chi_{tg}|\varrho_g)\right\} \times \left\{\prod_s p(\omega_s)\right\} \times \left\{\prod_j p(\kappa_j)\right\} \times \\
& \left\{\prod_t\prod_s p(\beta_{ts}|w_t,\gamma_{ts}) \times \prod_t\prod_s p(\gamma_{ts}|\omega_s)\right\} \times \left\{\prod_g p(\varrho_g)\right\} \times \\
& \left\{\prod_t\prod_{k<t} p(\eta_{tk}|\lambda)\right\} \times \left\{\prod_t p(w_t|a_w,b_w)\right\} \times \left\{\prod_t p(w_{\alpha_t}|a_\alpha,b_\alpha)\right\} \times \left\{\prod_t p(v_t|a_v,b_v)\right\} \times \\
& \left\{\prod_t p(\alpha_t|w_{\alpha_t})\right\} \times p(\lambda) \times p(b_w) \times p(b_v) \times p(\tau)
\end{aligned}
$$

The mean-field approximation distribution is defined as

$$
\begin{aligned}
q(\boldsymbol{\vartheta}) = & \left\{\prod_t q(\alpha_t)\right\} \times \left\{\prod_t\prod_s q(\beta_{ts},\gamma_{ts})\right\} \times \left\{\prod_t q(\boldsymbol{\theta}_t,\boldsymbol{\psi}_t,\boldsymbol{\xi}_t)\right\} \times \left\{\prod_t\prod_g q(\boldsymbol{\zeta}_{tg},\chi_{tg})\right\} \times \\
& \left\{\prod_s q(\omega_s)\right\} \times \left\{\prod_j q(\kappa_j)\right\} \times \left\{\prod_g q(\varrho_g)\right\} \times \left\{\prod_t q(\sigma_t^2)\prod_{k<t} q(\rho_{tk},\eta_{tk}|\sigma_t^2)\right\} \times \\
& \left\{\prod_t q(w_t)\right\} \times \left\{\prod_t q(w_{\alpha_t})\right\} \times \left\{\prod_t q(v_t)\right\} \times q(\lambda) \times q(b_w) \times q(b_v) \times q(\tau)
\end{aligned}
$$

with $f(\mathbf{z})^{(j)}$ as the j-th moment of $f(\mathbf{z})$ with respect to $q(\mathbf{z})$, $\mathbb{E}_q\left[f(\boldsymbol{z})^j\right]$.

Variational inference updates

To simplify the updates the following first order expectations are defined

$$(\boldsymbol{u}_t)^{(1)} = \boldsymbol{y}_t - \alpha_t \mathbf{1}_n - \sum_s X_s(\beta_{ts})^{(1)} - \sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_{tg})^{(1)} - \boldsymbol{Z}(\boldsymbol{\theta}_t)^{(1)} \tag{7.7.23}$$

$$(\boldsymbol{u}_{t,-\alpha})^{(1)} = \boldsymbol{y}_t - \sum_s X_s(\beta_{ts})^{(1)} - \sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_{tg})^{(1)} - \boldsymbol{Z}(\boldsymbol{\theta}_t)^{(1)} \tag{7.7.24}$$

$$(\boldsymbol{u}_{t,-s})^{(1)} = \boldsymbol{y}_t - \alpha_t \mathbf{1}_n - \sum_{l \neq s} X_l(\beta_{tl})^{(1)} - \boldsymbol{Z}(\boldsymbol{\theta}_t)^{(1)} \tag{7.7.25}$$

$$(\boldsymbol{u}_{t,-g})^{(1)} = \boldsymbol{y}_t - \alpha_t \mathbf{1}_n - \sum_s X_s(\beta_{ts})^{(1)} - \sum_{l \neq g} \boldsymbol{W}_l(\boldsymbol{\zeta}_{tl})^{(1)} - \boldsymbol{Z}(\boldsymbol{\theta}_t)^{(1)} \tag{7.7.26}$$

$$(\boldsymbol{u}_{t\not\beta})^{(1)} = \boldsymbol{y}_t - \alpha_t \mathbf{1}_n - \sum_s X_s(\beta_{ts})^{(1)} - \sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_{tg})^{(1)} \tag{7.7.27}$$

$$(\boldsymbol{u}_t)^{(1)} = (\boldsymbol{u}_{t\not\beta})^{(1)} - \boldsymbol{Z}(\boldsymbol{\theta}_t)^{(1)} \tag{7.7.28}$$

as $E_q[\boldsymbol{Z}_{\xi_t}\boldsymbol{\theta}_{\xi_t}] = \boldsymbol{Z}(\boldsymbol{\theta}_t)^{(1)}$. We also define the second order expectation as

$$\begin{aligned}
||\boldsymbol{u}_t||^{(2)} =& ||\boldsymbol{y}_t||^2 + n(\alpha_t)^{(2)} + \sum_s ||X_s||^2(\beta_{ts})^{(2)} + \sum_g (\boldsymbol{\zeta}_{tg}^T \boldsymbol{W}_g^T \boldsymbol{W}_g \boldsymbol{\zeta}_{tg})^{(1)} + \\
&+ \mathbb{E}_q[\boldsymbol{\theta}_{\xi_t}^T \boldsymbol{Z}_{\boldsymbol{\xi}_t}^T \boldsymbol{Z}_{\boldsymbol{\xi}_t} \boldsymbol{\theta}_{\xi_t}] - 2\sum_s \boldsymbol{y}_t^T X_s(\beta_{ts})^{(1)} - 2\boldsymbol{y}_t^T \boldsymbol{Z}(\boldsymbol{\theta}_t)^{(1)} + \\
&- 2\sum_g \boldsymbol{y}_t^T \boldsymbol{W}_g(\boldsymbol{\zeta}_{tg})^{(1)} - 2(\alpha_t)^{(1)} \boldsymbol{y}_t^T \mathbf{1}_n + 2\sum_{s \neq s', s < s'} X_s^T X_{s'}(\beta_{ts})^{(1)}(\beta_{ts'})^{(1)} + \\
&+ 2(\boldsymbol{Z}(\boldsymbol{\theta}_t)^{(1)})^T (\sum_s X_s(\beta_{ts})^{(1)}) + 2\sum_s \sum_g (\beta_{ts})^{(1)} X_s^T \boldsymbol{W}_g(\boldsymbol{\zeta}_{tg})^{(1)} + \\
&+ 2(Z(\theta)^{(1)})^T (\sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_{tg})^{(1)}) + 2\sum_{g \neq g', g < g'} (\boldsymbol{\zeta}_{tg})^{(1)T} \boldsymbol{W}_g^T \boldsymbol{W}_{g'}(\boldsymbol{\zeta}_{tg'})^{(1)} + \\
&+ 2(\alpha_t)^{(1)} \mathbf{1}_n^T \sum_s X_s(\beta_{ts})^{(1)} + 2(\alpha_t)^{(1)} \mathbf{1}_n^T \boldsymbol{Z}(\boldsymbol{\theta}_t)^{(1)} + 2(\alpha_t)^{(1)} \mathbf{1}_n^T \sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_{tg})^{(1)}.
\end{aligned} \tag{7.7.29}$$

The parameter updates are as follows:

$$\log q(\alpha_t) \propto \mathbb{E}_{-(\alpha_t)}\left[\log p(\boldsymbol{y}_t|\mathbf{X},\mathbf{W},\mathbf{Z},\mathbf{U}_t,\alpha_t,\boldsymbol{\beta}_t,\boldsymbol{\zeta}_t,\boldsymbol{\theta}_t,\sigma_t^2,\rho_t)+\sum_{k>t}\log p(\boldsymbol{y}_k|\cdot)+\log p(\alpha_t)\right] \tag{7.7.30}$$

$$\begin{aligned} p(\boldsymbol{y}_t|\cdot) &\propto \mathbb{E}_{-\beta_{ts}}\left[-\frac{1}{2\sigma_t^2}\left\|\left(\boldsymbol{y}_t-\alpha_t\mathbf{1}_n-\sum_s X_s\gamma_{ts}\beta_{ts}-\sum_g \boldsymbol{W}_g\boldsymbol{\zeta}_{tg}\chi_{tg}-\boldsymbol{Z}_{\xi_t}\boldsymbol{\theta}_{\xi_t}-\sum_{k<t}\boldsymbol{u}_k\rho_{tk}\right)\right\|^2\right] \\ &\propto -\frac{1}{2}(\sigma_t^{-2})^{(1)}\left(\alpha_t^2 n-2\alpha_t\left(\mathbf{1}_n^T\left((\boldsymbol{u}_{t,-\alpha})^{(1)}-\sum_{k<t}(\boldsymbol{u}_k)^{(1)}(\rho_{tk})^{(1)}\right)\right)\right) \end{aligned}$$

$$\begin{aligned} \log p(\boldsymbol{y}_k|\cdot) &\propto \mathbb{E}_{(-\alpha_t)}\Bigg[-\frac{1}{2\sigma_k^2}\Bigg\|\boldsymbol{y}_k-\alpha_k\mathbf{1}_n-\sum_s X_s\gamma_{ks}\beta_{ks}-\boldsymbol{Z}_{\xi_k}\boldsymbol{\theta}_{\xi_k}-\sum_g \boldsymbol{W}_g\boldsymbol{\zeta}_{kg}\chi_{kg}+ \\ &\quad -\sum_{h<k,h\neq t}\boldsymbol{u}_h\rho_{kh}-\left(\boldsymbol{y}_t-\sum_s X_s\gamma_{ts}\beta_{ts}-\sum_g \boldsymbol{W}_g\boldsymbol{\zeta}_{tg}\chi_{tg}-\boldsymbol{Z}_{\xi_t}\boldsymbol{\theta}_{\xi_t}\right)\rho_{kt}\Bigg)+\alpha_t\rho_{kt}\mathbf{1}_n\Bigg\|^2\Bigg] \\ &\propto \mathbb{E}_{(-\alpha_t)}\left[-\frac{\sigma_k^{-2}}{2}\left\|\boldsymbol{u}_k-\sum_{h<k,h\neq t}\boldsymbol{u}_h\rho_{kh}-\boldsymbol{u}_{(t,-\alpha)}\rho_{kt}+\alpha_t\rho_{kt}\mathbf{1}_n\right\|^2\right] \\ &\propto -\frac{(\sigma_k^{-2})^{(1)}}{2}\Bigg(n\alpha_t^2(\rho_{kt})^{(2)}-2\alpha_t\mathbf{1}_n^T\Big(\sum_{h<k,h\neq t}(\boldsymbol{u}_h)^{(1)}(\rho_{kh})^{(1)}(\rho_{tk})^{(1)}-(\boldsymbol{u}_k)^{(1)}(\rho_{kt})^{(1)}+ \\ &\quad +(\boldsymbol{u}_{t,-\alpha})^{(1)}(\rho_{kt})^{(2)}\Bigg)\Bigg] \end{aligned}$$

Bringing together

$$\begin{aligned} \log(\alpha_t) \propto -\frac{1}{2}\Bigg(&\Big((\sigma_t^{-2})^{(1)}+\sum_{k>t}(\sigma_k^{-2})^{(1)}(\rho_{kt})^{(2)}\Big)\alpha_t^2 n+ \\ &-2\alpha_t\mathbf{1}_n^T\Big((\sigma_t^{-2})^{(1)}\big(\boldsymbol{u}_{(t,-\alpha)}+\sum_{k<t}(\boldsymbol{u}_k)^{(1)}(\rho_{tk})^{(1)}\big)-\sum_{k>t}(\sigma_k^{-2})^{(1)}(\boldsymbol{u}_k)^{(1)}(\rho_{kt})^{(1)}+ \\ &+\sum_{k>t}(\sigma_k^{-2})^{(1)}\Big(\sum_{h<k,h\neq t}(\boldsymbol{u}_h)^{(1)}(\rho_{kh})^{(1)}(\rho_{kt})^{(1)}+(\boldsymbol{u}_{t,-\alpha})^{(1)}(\rho_{kt})^{(2)}\Big)+\alpha_t^2(w_{\alpha t})^{(-1)}\Bigg) \end{aligned}$$

defining the q free variational parameters

$$\sigma^2_{\alpha,t} = \left[n \left((\sigma_t^{-2})^{(1)} + \sum_{k>t} (\rho_{kt})^{(2)} (\sigma_k^{-2})^{(1)} \right) + (w_{\alpha t}^{-1})^{(1)} \right]^{-1} \tag{7.7.31}$$

$$\begin{aligned} \mu_{\alpha,t} =& \sigma^2_{\alpha,t} \mathbf{1}_n^T \left[(\sigma_t^{-2})^{(1)} \left((\boldsymbol{u}_{t,-\alpha})^{(1)} - \sum_{k<t} (\boldsymbol{u}_k)^{(1)} (\rho_{tk})^{(1)} \right) + \right. \\ &\left. + \sum_{k>t} (\sigma_k^{-2})^{(1)} (\rho_{kt})^{(2)} (\boldsymbol{u}_{t,-\alpha})^{(1)} - \sum_{k>t} (\sigma_k^{-2})^{(1)} \rho_{kt}^{(1)} \left((\boldsymbol{u}_k)^{(1)} - \sum_{h<k \;\; h \neq t} (\boldsymbol{u}_h)^{(1)} (\rho_{kh})^{(1)} \right) \right] \end{aligned} \tag{7.7.32}$$

we have $q(\alpha_t) = \mathcal{N}(\mu_{\alpha,t}, \sigma^2_{\alpha,t})$, where $(\alpha_t)^{(1)} = \mu_{\alpha,t}$.

$$\log q(\beta_{ts}, \gamma_{ts}) \propto \mathbb{E}_{-(\beta_{ts},\gamma_{ts})} \left[\log p(\boldsymbol{y}_t | \mathbf{X}, \mathbf{W}, \mathbf{Z}, \mathbf{U}_t, \alpha_t, \boldsymbol{\beta}_t, \boldsymbol{\zeta}_t, \boldsymbol{\theta}_t, \sigma_t^2, \rho_t) + \sum_{k>t} \log p(\boldsymbol{y}_k | \cdot) \right] + \tag{A}$$

$$+ \mathbb{E}_{-(\beta_{ts},\gamma_{ts})} \left[\log p(\beta_{ts} | \gamma_{ts}, w_t) \right] + \tag{B}$$

$$+ \mathbb{E}_{-(\beta_{ts},\gamma_{ts})} \left[\log p(\gamma_{ts} | \omega_s) \right] \tag{C}$$

(B) and (C) can be easily computed as and are proportional to

$$\text{(B)}: \quad -\gamma_{ts} \left(\frac{1}{2} (w_t^{-1})^{(1)} \left\| \beta_{ts} \right\|^2 \right) + (1 - \gamma_{ts}) \, \delta_0(\beta_{ts}) + \frac{\gamma_{ts}}{2} \left[(\log w_t^{-1})^{(1)} - \log 2\pi \right]$$

$$\text{(C)}: \quad \gamma_{ts} \left(\log \omega_s \right)^{(1)} + (1 - \gamma_{ts}) \left(\log(1 - \omega_s) \right)^{(1)}$$

and we can write A, inserting the latent variable which augments the likelihood because of the spike-and-slab priors as

$$\text{(A)}: \quad \mathbb{E}_{-(\beta_{ts},\gamma_{ts})} \left[\log p(\boldsymbol{y}_t | \cdot) \right] + \sum_{k>t} \mathbb{E}_{-(\beta_{ts},\gamma_{ts})} \left[\log p(\boldsymbol{y}_k | \cdot) \right] = A_t^1 + \sum_{k>t} A_{tk}^2$$

$$A_t^1 :\mathbb{E}_{-\beta_{ts}}\Big[-\frac{1}{2\sigma_t^2}\Big\|\Big(\boldsymbol{y}_t-\alpha_t\mathbf{1}_n-\sum_{l\neq s}X_l\gamma_{tl}\beta_{tl}-\sum_g \boldsymbol{W}_g\boldsymbol{\zeta}_{tg}\chi_{tg}-\boldsymbol{Z}_{\xi_t}\boldsymbol{\theta}_{\xi_t}-\sum_{k<t}\boldsymbol{u}_k\rho_{tk}\Big)+$$
$$-X_s\gamma_{ts}\beta_{ts}\Big\|^2\Big]=\mathbb{E}_{-\beta_{ts}}\left[-\frac{1}{2\sigma_t^2}\left\|\boldsymbol{u}_{t,-s}-\sum_{k<t}\boldsymbol{u}_k\rho_{tk}-X_s\gamma_{ts}\beta_{ts}\right\|^2\right]$$
$$\propto-\frac{1}{2}(\sigma_t^{-2})^{(1)}\gamma_{ts}\Big(\|X_s\|^2\beta_{ts}^2-2\beta_{ts}\Big(X_s^T\Big((\boldsymbol{u}_{t,-s})^{(1)}-\sum_{k<t}(\boldsymbol{u}_k)^{(1)}(\rho_{tk})^{(1)}\Big)\Big)\Big)$$

$$A_{tk}^2 :\mathbb{E}_{-\beta_{ts}}\Bigg[-\frac{1}{2\sigma_k^2}\Bigg\|\Bigg(\boldsymbol{y}_k-\alpha_k\mathbf{1}_n-\sum_j X_j\gamma_{kj}\beta_{kj}-\sum_g\boldsymbol{W}_g\boldsymbol{\zeta}_{kg}-\boldsymbol{Z}_{\xi_k}\boldsymbol{\theta}_{\xi_k}+$$
$$-\sum_{h<k,h\neq t}\boldsymbol{u}_h\rho_{kh}-\Big(\boldsymbol{y}_t-\sum_{j\neq s}X_j\gamma_{tj}\beta_{tj}-\boldsymbol{Z}_{\xi_t}\boldsymbol{\theta}_{\xi_t}-\sum_g\boldsymbol{W}_g\boldsymbol{\zeta}_{tg}\Big)\rho_{kt}\Big)+X_s\gamma_{ts}\beta_{ts}\rho_{kt}\Bigg\|^2\Bigg]$$
$$=-\frac{(\sigma_k^{-2})^{(1)}}{2}\left\|(\boldsymbol{u}_k)^{(1)}-\sum_{h<k,h\neq t}(\boldsymbol{u}_h)^{(1)}(\rho_{kh})^{(1)}-(\boldsymbol{u}_{t,-s})^{(1)}\rho_{kt}^{(1)}+X_s\gamma_{ts}\beta_{ts}(\rho_{kt})^{(1)}\right\|^2$$
$$\propto-\frac{(\sigma_k^{-2})^{(1)}}{2}\gamma_{ts}\Bigg(\|X_s\|^2(\beta_{ts})^{(2)}(\rho_{kt})^{(2)}+2\gamma_{ts}\beta_{ts}X_s^T((\boldsymbol{u}_k)^{(1)}(\rho_{kt})^{(1)}+$$
$$-\sum_{h<k,h\neq t}(\boldsymbol{u}_h)^{(1)}(\rho_{kh})^{(1)}(\rho_{tk})^{(1)}-(\boldsymbol{u}_{t,-s})^{(1)}(\rho_{kt})^{(2)}\Bigg)$$

For (A)

$$A_t^1+\sum_{k>t}A_k^2\propto-\frac{1}{2}\Bigg(\Big((\sigma_t^{-2})^{(1)}+\sum_{k>t}(\sigma_k^{-2})^{(1)}(\rho_{kt})^{(2)}\Big)\|X_s\|^2\gamma_{ts}\beta_{ts}^2+$$
$$-2\gamma_{ts}\beta_{ts}\Bigg(X_s^T\Big((\sigma_t^{-2})^{(1)}\big(\boldsymbol{u}_{t,-s}+\sum_{k<t}(\boldsymbol{u}_k)^{(1)}(\rho_{tk})^{(1)}\big)-\sum_{k>t}(\sigma_k^{-2})^{(1)}(\boldsymbol{u}_k)^{(1)}(\rho_{kt})^{(1)}+$$
$$+\sum_{k>t}(\sigma_k^{-2})^{(1)}\Big(\sum_{h<k,h\neq t}(\boldsymbol{u}_h)^{(1)}(\rho_{kh})^{(1)}(\rho_{kt})^{(1)}+(\boldsymbol{u}_{t,-s})^{(1)}(\rho_{kt})^{(2)}\Big)\Bigg)\Bigg)$$

Bringing together we have

$$\begin{aligned}\log q(\beta_{ts},\gamma_{ts}) \propto &-\frac{\gamma_{ts}\beta_{ts}^2}{2}\Bigg(\Big((\sigma_t^{-2})^{(1)}+\sum_{k>t}(\sigma_k^{-2})^{(1)}(\rho_{kt})^{(2)}\Big)\|X_s\|^2+(w_t^{-1})^{(1)}\Bigg)+\\ &-2\gamma_{ts}\beta_{ts}\Bigg(X_s^T\Big((\sigma_t^{-2})^{(1)}\big(\boldsymbol{u}_{t,-s}+\sum_{k<t}(\boldsymbol{u}_k)^{(1)}(\rho_{tk})^{(1)}\big)+\sum_{k>t}(\sigma_k^{-2})^{(1)}(\boldsymbol{u}_{t,-s})^{(1)}(\rho_{kt})^{(2)}+\\ &-\sum_{k>t}(\sigma_k^{-2})^{(1)}(\rho_{kt})^{(1)}\Big((\boldsymbol{u}_k)^{(1)}-\sum_{h<k,h\neq t}(\boldsymbol{u}_h)^{(1)}(\rho_{kh})^{(1)}\Big)\Bigg)\end{aligned}$$

defining the q free variational parameters

$$\sigma^2_{\beta,ts}=\left[\|X_s\|^2\left((\sigma_t^{-2})^{(1)}+\sum_{k>t}(\rho_{kt})^{(2)}(\sigma_k^{-2})^{(1)}\right)+(w_t^{-1})^{(1)}\right]^{-1} \tag{7.7.33}$$

$$\begin{aligned}\mu_{\beta,ts}=&\sigma^2_{\beta,ts}X_s^T\Bigg[(\sigma_t^{-2})^{(1)}\left((\boldsymbol{u}_{t,-s})^{(1)}-\sum_{k<t}(\boldsymbol{u}_k)^{(1)}(\rho_{tk})^{(1)}\right)+ \\ &+\sum_{k>t}(\sigma_k^{-2})^{(1)}(\rho_{kt})^{(2)}(\boldsymbol{u}_{t,-s})^{(1)}-\sum_{k>t}(\sigma_k^{-2})^{(1)}\rho_{kt}^{(1)}\left((\boldsymbol{u}_k)^{(1)}-\sum_{h<k\ h\neq t}(\boldsymbol{u}_h)^{(1)}(\rho_{kh})^{(1)}\right)\Bigg]\end{aligned} \tag{7.7.34}$$

(7.7.35)

The law of iterative expectations is used to obtain the expectation $(\beta_{ts})^{(1)}=\mathbb{E}_{q(\mathbf{z})}[\beta_{ts}]$, given that β_{ts} is parametrised by a mixture distribution

$$\begin{aligned}\mathbb{E}_{q(\mathbf{z})}[\beta_{ts}]&=\mathbb{E}_{q(\gamma_{ts})}[\mathbb{E}_q[\beta_{ts}|\gamma_{ts}]]\\ &=\mu_{\beta,ts}(\gamma_{ts})^{(1)}+0(1-(\gamma_{ts})^{(1)})=\mu_{\beta,ts}(\gamma_{ts})^{(1)}\end{aligned}$$

By exponentiating and completing the square we arrive at

$$\begin{aligned}q(\beta_{ts},\gamma_{ts})\propto&\left[(2\pi\sigma^2_{\beta,ts})^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2_{\beta,ts}}(\beta_{ts}-\mu_{\beta,ts})^2\right\}\right]^{\gamma_{ts}}\times \\ &\times\left[\left\{\exp\left((\log w_t^{-1})^{(1)}+(\log\sigma_t^{-2})^{(1)}\right)\sigma^2_{\beta,ts}\right\}^{\frac{1}{2}}\exp\left\{\frac{1}{2}\mu_{\beta,ts}\sigma^{-2}_{\beta,ts}\right\}\exp\left\{(\log\omega_s)^{(1)}\right\}\right]^{\gamma_{ts}}\times \\ &\times\delta_0(\beta_{ts})^{1-\gamma_{ts}}\exp\left\{(\log 1-\omega_s)^{(1)}\right\}^{1-\gamma_{ts}}\end{aligned} \tag{7.7.36}$$

and thus by calling

$$(\gamma_{ts})^{(1)} = \left[1 + \sqrt{\sigma^{-2}_{\beta,ts}} \exp\left\{(\log 1 - \omega_s)^{(1)} - (\log \omega_s)^{(1)} - \frac{1}{2}(\log w_t^{-1})^{(1)} - \frac{1}{2}\mu^2_{\beta,ts}\sigma^{-2}_{\beta,ts}\right\}\right]^{-1} \tag{7.7.37}$$

we have that under q

$$q(\beta_{ts}|\gamma_{ts} = 1) = \mathcal{N}(\mu_{\beta,ts}, \sigma^2_{\beta,ts}), \quad q(\beta_{ts}|\gamma_{ts} = 0) = \delta_0(\beta_{ts})$$
$$q(\gamma_{ts}) = Bern((\gamma_{ts})^{(1)}).$$

Note that now

$$(\beta_{ts})^{(1)} = \mu_{\beta,ts}(\gamma_{ts})^{(1)} \tag{7.7.38}$$
$$(\beta_{ts})^{(2)} = (\sigma^2_{\beta,ts} + \mu^2_{\beta,ts})(\gamma_{ts})^{(1)}. \tag{7.7.39}$$

$$\log q(\boldsymbol{\zeta}_{tg}, \chi_{tg}) \propto \mathbb{E}_{(\boldsymbol{\zeta}_{tg},\chi_{tg})}\Big[\log p(\boldsymbol{y}_t|\cdot) + \sum_{k>t} \log p(\boldsymbol{y}_k|\cdot) + \log p(\boldsymbol{\zeta}_{tg}|\chi_{tg}, v_t) + \log p(\chi_{tg}|\varrho_g)\Big]$$

The index g denotes the categorical factor groupings $g = 1, ..., G$ and m_g is the dimension of the vector $\boldsymbol{\zeta}_g$. As the categorical factors are coded with reference to the intercept, m_g is always 1 less than the levels in the categorical factor. The first likelihood component is proportional to

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{\zeta}_{tg},\chi_{tg})}[\log p(\boldsymbol{y}_t|\cdot)] =& \mathbb{E}_{(\boldsymbol{\zeta}_{tg},\chi_{tg})}\Big[-\frac{1}{2\sigma_t^2}\Big\|\boldsymbol{y}_t - \alpha_t\mathbf{1}_n - \sum_s X_s\gamma_{ts}\beta_{ts} - \sum_{l\neq g}\boldsymbol{W}_l\boldsymbol{\zeta}_{tl} + \\
&- \boldsymbol{W}_g\boldsymbol{\zeta}_{tg} - \mathbf{Z}_{\boldsymbol{\xi}_t}\boldsymbol{\theta}_{\xi_t} - \sum_{k<t}\boldsymbol{u}_k\rho_{tk}\Big\|^2\Big] \\
=& \mathbb{E}_{(\boldsymbol{\zeta}_{tg},\chi_{tg})}\Big[-\frac{1}{2\sigma_t^2}\Big\|\boldsymbol{u}_{t,-g} - \boldsymbol{W}_g\boldsymbol{\zeta}_{tg} - \sum_{k<t}\boldsymbol{u}_k\rho_{tk}\Big\|^2\Big] \\
\propto& -\frac{\chi_{tg}}{2\sigma_t^2}\left(\boldsymbol{\zeta}_{tg}^T\boldsymbol{W}_g^T\boldsymbol{W}_g\boldsymbol{\zeta}_{tg} - 2\boldsymbol{\zeta}_{tg}^T\boldsymbol{W}_g^T\Big((\boldsymbol{u}_{t,-g})^{(1)} - \sum_{k<t}\boldsymbol{u}_k(\rho_{tk})^{(1)}\Big)\right)
\end{aligned}$$

The spike-and-slab prior forces the latent selection variables into the likelihood component. The second likelihood component is proportional to

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{\zeta}_{tg},\chi_{tg})}[\log p(\boldsymbol{y}_k|\cdot)] =& \mathbb{E}_{(\boldsymbol{\zeta}_{tg},\chi_{tg})}\Big[-\frac{1}{2\sigma_k^2}\Big\| \boldsymbol{y}_k - \alpha_k \mathbf{1}_n - \sum_s X_s\gamma_{ks}\beta_{ks} + \\
&- \sum_j \boldsymbol{W}_j\chi_{kj}\boldsymbol{\zeta}_{kj} - \mathbf{Z}_{\xi_k}\boldsymbol{\theta}_{\xi_k} - \sum_{h<k,h\neq t} \boldsymbol{u}_k\rho_{kh} + \\
&- \Big(\boldsymbol{y}_t - \alpha_t\mathbf{1}_n - \sum_s X_s\gamma_{ts}\beta_{ts} - \sum_{j\neq g}\boldsymbol{W}_j\chi_{tj}\boldsymbol{\zeta}_{tj} - \boldsymbol{W}_g\chi_{tg}\boldsymbol{\zeta}_{tg} - \mathbf{Z}_{\boldsymbol{\xi}_t}\boldsymbol{\theta}_{\xi_t}\Big)\rho_{kt}\Big\|^2\Big] \\
=& \mathbb{E}_{(\boldsymbol{\zeta}_{tg},\chi_{tg})}\Big[-\frac{1}{2\sigma_k^2}\Big\| \boldsymbol{u}_k - \sum_{h<k,h\neq t}\boldsymbol{u}_h\rho_{kh} - \boldsymbol{u}_{t,-g}\rho_{kt} + \boldsymbol{W}_g\chi_{tg}\boldsymbol{\zeta}_{tg}\rho_{kt}\Big\|^2\Big] \\
\propto& -\frac{1}{2(\sigma_k^2)^{(1)}}\chi_{tg}\Bigg(\boldsymbol{\zeta}_{tg}^T\boldsymbol{W}_g^T\boldsymbol{W}_g\boldsymbol{\zeta}_{tg}(\rho_{kt})^{(2)} + 2\boldsymbol{\zeta}_{tg}^T\boldsymbol{W}_g^T\Bigg[(\boldsymbol{u}_k)^{(1)}(\rho_{kt})^{(1)} - \\
&\sum_{h<k,h\neq t}(\boldsymbol{u}_h)^{(1)}(\rho_{kh})^{(1)}(\rho_{kt})^{(1)} - (\boldsymbol{u}_{t,-g})^{(1)}(\rho_{kt})^{(2)}\Bigg]\Bigg)
\end{aligned}$$

Bringing the likelihood components together $\log p(\boldsymbol{Y}|\cdot) = \log p(\boldsymbol{y}_t|\cdot) + \sum_{k>t}\log p(\boldsymbol{y}_k|\cdot)$ gives

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{\zeta}_{tg},\chi_{tg})}[\log p(\boldsymbol{Y}|\cdot)] \propto& -\frac{1}{2}\Bigg(\Bigg((\sigma_t^2)^{(1)} + \sum_{k>t}(\sigma_k^{-2})^{(1)}(\rho_{kt})^{(2)}\Bigg)\chi_{tg}\boldsymbol{\zeta}_{tg}^T\boldsymbol{W}_g^T\boldsymbol{W}_g\boldsymbol{\zeta}_{tg} + \\
&- 2\chi_{tg}\boldsymbol{\zeta}_{tg}^T\boldsymbol{W}_g^T\Bigg[(\sigma_t^{-2})^{(1)}\Big((\boldsymbol{u}_{t,-g})^{(1)} - \sum_{k<t}(\boldsymbol{u}_k)^{(1)}(\rho_{tk})^{(1)}\Big) + \\
&+ (\sigma_k^{-2})^{(1)}\Big(-\sum_{k>t}(\boldsymbol{u}_k)^{(1)}(\rho_{kt})^{(1)} + \sum_{k>t}\sum_{h<k,h\neq t}(\boldsymbol{u}_h)^{(1)}(\rho_{kh})^{(1)}(\rho_{kt})^{(1)} + \\
&+ \sum_{k>t}(\boldsymbol{u}_{t,-g})^{(1)}(\rho_{kt})^{(2)}\Bigg]\Bigg)
\end{aligned}$$

Bringing together we have

$$\begin{aligned}
\log q(\boldsymbol{\zeta}_g,\chi_g) \propto \mathbb{E}_{(\boldsymbol{\zeta}_{tg},\chi_{tg})}[\log p(Y|\cdot)] - \frac{(v_t^{-1})^{(1)}}{2}\chi_{tg}\boldsymbol{\zeta}_{tg}^T\boldsymbol{\zeta}_{tg} + \chi_{tg}\frac{m_g}{2}\Big((\log v_t^{-1})^{(1)} - \log 2\pi\Big) + \\
\chi_{tg}(\log(\varrho_g))^{(1)} + (1-\chi_{tg})(\log(1-\varrho_g))^{(1)} + (1-\chi_{tg})\delta_0(\boldsymbol{\zeta}_{tg})
\end{aligned}$$

defining

$$\Sigma_{\boldsymbol{\zeta}_{tg}} = \left[(v_t^{-1})^{(1)}\mathbf{I}_{m_g} + \Big((\sigma_t^{-2})^{(1)} + \sum_{k>t}(\sigma_k^{-2})(\rho_{kt})^{(2)}\Big)\boldsymbol{W}_g^T\boldsymbol{W}_g\right]^{-1} \tag{7.7.40}$$

$$\begin{aligned}\boldsymbol{\mu}_{\zeta_g} = \Sigma_{\boldsymbol{\zeta}_{tg}}\boldsymbol{W}_g^T\Bigg[&(\sigma_t^{-2})^{(1)}\Big((\boldsymbol{u}_{t,-g})^{(1)} - \sum_{k<t}(\boldsymbol{u}_k)^{(1)}(\rho_{tk})^{(1)}\Big) + \sum_{k>t}(\boldsymbol{u}_{t,-g})^{(1)}(\rho_{kt})^{(2)}(\sigma_k^{-2})^{(1)} + \\ &- \sum_{k>t}(\boldsymbol{u}_k)^{(1)}(\rho_{kt})^{(1)}(\sigma_k^{-2})^{(1)} + \sum_{k>t}\sum_{h<k,h\neq t}(\boldsymbol{u}_h)^{(1)}(\rho_{kh})^{(1)}(\rho_{kt})^{(1)}(\sigma_k^{-2})^{(1)}\Bigg]\end{aligned} \tag{7.7.41}$$

by exponentiating, completing the square we have

$$\begin{aligned}q(\boldsymbol{\zeta}_{tg},\chi_{tg}|y) = &\left[\frac{1}{(2\pi)^{m_g/2}}\det(\Sigma_{\boldsymbol{\zeta}_{tg}})^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\zeta}_{tg}-\boldsymbol{\mu}_{\zeta_{tg}})^T\Sigma_{\boldsymbol{\zeta}_{tg}}^{-1}(\boldsymbol{\zeta}_{tg}-\boldsymbol{\mu}_{\zeta_{tg}})\right\}\right]^{\chi_{tg}} \\ &\delta_0(\boldsymbol{\zeta}_{tg})^{1-\chi_{tg}}\left[\exp\big((\log(1-\varrho_g))^{(1)}\big)\right]^{1-\chi_{tg}} \\ &\left[\exp\left(\frac{1}{2}\boldsymbol{\mu}_{\zeta_{tg}}^T\Sigma_{\boldsymbol{\zeta}_{tg}}^{-1}\boldsymbol{\mu}_{\zeta_{tg}} + \frac{1}{2}\log\det(\Sigma_{\boldsymbol{\zeta}_{tg}}) + \frac{m_g}{2}(\log v_t^{-1})^{(1)} + (\log\varrho_g)^{(1)}\right)\right]^{\chi_{tg}}\end{aligned} \tag{7.7.42}$$

and thus by calling

$$\begin{aligned}(\chi_{tg})^{(1)} = &\Big[1+\exp\Big((\log 1-\varrho_g)^{(1)} - (\log\varrho_g)^{(1)} - \frac{m_g}{2}(\log v_t^{-1})^{(1)} - \frac{1}{2}\boldsymbol{\mu}_{\zeta_{tg}}^T\Sigma_{\boldsymbol{\zeta}_{tg}}^{-1}\boldsymbol{\mu}_{\zeta_{tg}} + \\ &- \frac{1}{2}\log\big(\det(\Sigma_{\boldsymbol{\zeta}_{tg}})\big)\Big)\Big]^{-1}\end{aligned} \tag{7.7.43}$$

we have under q

$$\begin{aligned}q(\boldsymbol{\zeta}_{tg}|\chi_{tg}=1,y) = \mathcal{N}_{m_g}(\boldsymbol{\mu}_{\zeta_{tg}},\Sigma_{\boldsymbol{\zeta}_{tg}}), \quad q(\boldsymbol{\zeta}_{tg}|\chi_{tg}=0,y) = \delta_0(\boldsymbol{\zeta}_{tg}) \\ q(\chi_{tg}|y) = Bern((\chi_{tg})^{(1)}).\end{aligned}$$

Note that now

$$(\boldsymbol{\zeta}_{tg})^{(1)} = \boldsymbol{\mu}_{\zeta_{tg}}(\chi_{tg})^{(1)} \tag{7.7.44}$$

$$(\boldsymbol{\zeta}_{tg}^T\boldsymbol{\zeta}_{tg})^{(1)} = (\text{tr}(\Sigma_{\boldsymbol{\zeta}_{tg}}) + \boldsymbol{\mu}_{\zeta_{tg}}^T\boldsymbol{\mu}_{\zeta_{tg}})(\chi_{tg})^{(1)} \tag{7.7.45}$$

$$(\boldsymbol{\zeta}_{tg}^T\boldsymbol{W}_g^T\boldsymbol{W}_g\boldsymbol{\zeta}_{tg})^{(1)} = (\text{tr}(\boldsymbol{W}_g\Sigma_{\boldsymbol{\zeta}_{tg}}\boldsymbol{W}_g^T) + \boldsymbol{\mu}_{\zeta_g}^T\boldsymbol{W}_g^T\boldsymbol{W}_g\boldsymbol{\mu}_{\zeta_{tg}})(\chi_{tg})^{(1)} \tag{7.7.46}$$

$$\log\, q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t) \propto \mathbb{E}_{-(\boldsymbol{\theta}_t,\boldsymbol{\psi}_t,\boldsymbol{\xi}_t)}\left[\log p(\boldsymbol{\psi}_t|\boldsymbol{\xi}_t) + \log p(\boldsymbol{\xi}_t|\kappa)\right] +$$
$$+ \left(\mathbb{E}_{-(\boldsymbol{\theta}_t,\boldsymbol{\psi}_t,\boldsymbol{\xi}_t)}\left[\log p(\boldsymbol{y}_t|\cdot) + \sum_{k>t}\log p(\boldsymbol{y}_k|\cdot)\right] + \mathbb{E}_{-(\boldsymbol{\theta}_t,\boldsymbol{\psi}_t,\boldsymbol{\xi}_t)}\left[\log p(\boldsymbol{\theta}_t|\boldsymbol{\xi}_t, \boldsymbol{\psi}_t)\right]\right)_{[I(\sum_j \theta_{tj}=0)=1]}$$

$$\mathbb{E}_{-(.)}\left[\log p(\boldsymbol{y}_t|\cdot) + \sum_{k>t}\log p(\boldsymbol{y}_k|\cdot)\right] \propto -\frac{1}{2\sigma_t^2}\Big\|\boldsymbol{y}_t - \alpha_t\mathbf{1}_n - \sum_s X_s\gamma_{ts}\beta_{ts} - \sum_g \boldsymbol{W}_g\chi_{tg}\boldsymbol{\zeta}_{tg} - \boldsymbol{Z}_{\xi_t}\boldsymbol{\theta}_{\xi_t} +$$
$$- \sum_{k<t}\boldsymbol{u}_k\rho_{tk}\Big\|^2\Bigg) + \sum_{k>t} -\frac{1}{2\sigma_k^2}\Big\|\boldsymbol{y}_k - \alpha_k\mathbf{1}_n - \sum_s X_s\gamma_{ks}\beta_{ks} +$$
$$- \boldsymbol{Z}_{\xi_k}\boldsymbol{\theta}_{\xi_k} - \sum_g \boldsymbol{W}_g\boldsymbol{\zeta}_{kg}\chi_{kg} - \sum_{h<k}\boldsymbol{u}_h\rho_{kh}\Big\|^2$$

$$\mathbb{E}_{-(.)}\left[\log p(\boldsymbol{\theta}_t|\boldsymbol{\xi}_t, \boldsymbol{\psi}_t)\right] \propto -\frac{1}{2}(d_{\boldsymbol{\xi}_t} - 1)\log(2\pi) - \frac{1}{2}\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})) +$$
$$- \frac{1}{2}\boldsymbol{\theta}_{\xi_t}^T(\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})^+\boldsymbol{\theta}_{\xi_t} + \log\delta(\theta_{\bar{\xi}_t})$$

$$\mathbb{E}_{-(.)}\left[\log(p(\boldsymbol{\psi}_t|\boldsymbol{\xi}_t)p(\boldsymbol{\xi}_t|\kappa))\right] \propto + \sum_j \left(\xi_{tj}\log(\kappa_j) + (1-\xi_{tj})\log(1-\kappa_j)\right) + \sum_j \xi_{tj}(a_{\boldsymbol{\psi}_t}\log(b_{\boldsymbol{\psi}_t}) +$$
$$- \sum_j \xi_{tj}\log(\Gamma(a_{\boldsymbol{\psi}_t})) - \sum_j (a_{\boldsymbol{\psi}_t}+1)\xi_{tj}\log(\psi_{tj}) - b_{\boldsymbol{\psi}_t}\sum_j \xi_{tj}\psi_{tj}^{-1}$$

Thus we can express

$$\log q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t|\boldsymbol{y}_t, \cdot) \propto \mathbb{E}_{-(\boldsymbol{\theta}_t,\boldsymbol{\psi}_t,\boldsymbol{\xi}_t)}\left[A + B + C\right] \tag{7.7.47}$$

where A is proportional to

$$A \propto -\frac{1}{2}(d_{\boldsymbol{\xi}_t} - 1)\log(2\pi) - \frac{1}{2}\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t} D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})) + \tag{7.7.48}$$
$$- \frac{1}{2}\left(\boldsymbol{\theta}_{\xi_t}^T\left((\mathbf{T}_{\boldsymbol{\xi}_t} D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})^+ + \sigma_t^{-2}\boldsymbol{Z}_{\xi_t}^T\boldsymbol{Z}_{\xi_t}\right)\boldsymbol{\theta}_{\xi_t} - 2\sigma_t^{-2}\boldsymbol{\theta}_{\xi_t}^T\boldsymbol{Z}_{\xi_t}^T\left(\boldsymbol{y}_t - (\boldsymbol{u}_{t\not j})^{(1)} - \sum_{k<t}\boldsymbol{u}_k\rho_{tk}\right)\right)$$

and B is proportional to

$$B \propto -\frac{1}{2}\sigma_k^{-2}\Big\|\boldsymbol{y}_k - \alpha_k\mathbf{1}_n - \sum_s X_s\gamma_{ks}\beta_{ks} - \boldsymbol{Z}_{\xi_k}\boldsymbol{\theta}_{\xi_k} - \sum_g \boldsymbol{W}_g\boldsymbol{\zeta}_{kg}\chi_{kg} +$$
$$- \sum_{h<k, h\neq t}\boldsymbol{u}_h\rho_{kh} - u_{t\not j}\rho_{kt} + \boldsymbol{Z}_{\xi_t}\boldsymbol{\theta}_{\xi_t}\rho_{kt}\Big\|^2$$
$$\propto -\frac{1}{2}\sigma_k^{-2}\left(\boldsymbol{\theta}_{\xi_t}^T\boldsymbol{Z}_{\xi_t}^T\boldsymbol{Z}_{\xi_t}\boldsymbol{\theta}_{\xi_t}\rho_{kt}^2 + 2\boldsymbol{\theta}_{\xi_t}^T\boldsymbol{Z}_{\xi_t}^T\boldsymbol{u}_k\rho_{kt} - 2\boldsymbol{\theta}_{\xi_t}^T\boldsymbol{Z}_{\xi_t}^T\sum_{h<k, h\neq t}\boldsymbol{u}_h\rho_{kh}\rho_{kt} +\right.$$
$$\left. - 2\boldsymbol{\theta}_{\xi_t}^T\boldsymbol{Z}_{\xi_t}^T\boldsymbol{u}_{t\not j}\rho_{kt}^2\right) \tag{7.7.49}$$

Bringing together

$$\log p(\boldsymbol{\theta}_{\xi_t}, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t | y, .) \propto \mathbb{E}_{(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t)} \Bigg[-\frac{1}{2}(d_{\boldsymbol{\xi}_t} - 1)\log(2\pi) - \frac{1}{2}\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t} D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})) +$$
$$- \frac{1}{2}\boldsymbol{\theta}_{\xi_t}^T \Bigg((\mathbf{T}_{\boldsymbol{\xi}_t} D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})^+ + \sigma_t^{-2} \boldsymbol{Z}_{\xi_t}^T \boldsymbol{Z}_{\xi_t} + \sum_{k>t} \sigma_k^{-2} \rho_{kt}^2 \boldsymbol{Z}_{\xi_t}^T \boldsymbol{Z}_{\xi_t} \Bigg) \boldsymbol{\theta}_{\xi_t} +$$
$$- 2\boldsymbol{\theta}_{\xi_t}^T \boldsymbol{Z}_{\xi_t}^T \Bigg(\sigma_t^{-2} (\boldsymbol{u}_{t,\not{J}} - \sum_{k<t} \boldsymbol{u}_k \rho_{tk}) - \sum_{k>t} \sigma_k^{-2} \rho_{kt} \boldsymbol{u}_k +$$
$$+ \sum_{k>t} \sum_{h<k, h\neq t} \sigma_k^{-2} \boldsymbol{u}_h \rho_{kh} \rho_{kt} + \sum_{k>t} \sigma_k^{-2} \boldsymbol{u}_{t,\not{J}} \rho_{kt}^2 \Bigg) \Bigg] \tag{7.7.50}$$

Defining the vector $\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}$ and matrix $\Sigma_{\boldsymbol{\theta}_{\xi_t}}$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}} = \Sigma_{\boldsymbol{\theta}_{\xi_t}} \Bigg(\boldsymbol{Z}_{\xi_t}^T \Bigg((\sigma_t^{-2})^{(1)} \Bigg((\boldsymbol{u}_{t,\not{J}})^{(1)} - \sum_{k<t} (\boldsymbol{u}_k)^{(1)} (\rho_{tk})^{(1)} \Bigg) - \sum_{k>t} (\sigma_k^{-2})^{(1)} (\rho_{kt})^{(1)} (\boldsymbol{u}_k)^{(1)} +$$
$$+ \sum_{k>t} \sum_{h<k, h\neq t} (\sigma_k^{-2})^{(1)} (\boldsymbol{u}_h)^{(1)} (\rho_{kh})^{(1)} (\rho_{kt})^{(1)} + \sum_{k>t} (\sigma_k^{-2})^{(1)} (\boldsymbol{u}_{t,\not{J}})^{(1)} (\rho_{kt})^{(2)} \Bigg) \tag{7.7.51}$$

$$\Sigma_{\boldsymbol{\theta}_{\xi_t}} = \Bigg((\mathbf{T}_{\boldsymbol{\xi}_t} D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})^+ + (\sigma_t^{-2})^{(1)} \boldsymbol{Z}_{\xi_t}^T \boldsymbol{Z}_{\xi_t} + \sum_{k>t} (\sigma_k^{-2})^{(1)} (\rho_{kt})^{(2)} \boldsymbol{Z}_{\xi_t}^T \boldsymbol{Z}_{\xi_t} \Bigg)^{-1} \tag{7.7.52}$$

which are still function of the vector $\boldsymbol{\xi}_t$.

$$\log q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t | \boldsymbol{y}_t, .) \propto \Bigg[-\frac{1}{2}(d_{\boldsymbol{\xi}_t} - 1)(\log 2\pi) - \frac{1}{2}\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t} D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})) +$$
$$- \frac{1}{2} \Bigg([\boldsymbol{\theta}_{\xi_t} - \boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}]^T \Sigma_{\boldsymbol{\theta}_{\xi_t}}^{-1} [\boldsymbol{\theta}_{\xi_t} - \boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}] \Bigg) - \boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}^T \Sigma_{\boldsymbol{\theta}_{\xi_t}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}} \Bigg]_{[I(\sum_j \theta_{\xi_{tj}} = 0) = 1]} +$$
$$+ \sum_j \xi_{tj} (\log \kappa_j)^{(1)} + \sum_j (1 - \xi_{tj}) (\log(1 - \kappa_j))^{(1)} - b_{\boldsymbol{\psi}_t} \sum_j \xi_{tj} \psi_{tj}^{-1} +$$
$$+ (a_{\boldsymbol{\psi}_t} \log(b_{\boldsymbol{\psi}_t}) - \log(\Gamma(a_{\boldsymbol{\psi}_t})) \sum_j \xi_{tj} - \sum_j (a_{\boldsymbol{\psi}_t} + 1) \xi_{tj} \log(\psi_{tj}) \tag{7.7.53}$$

We want to identify the parts related to $\boldsymbol{\xi}_t$ and $\boldsymbol{\theta}_t$. We can remove the index by adding the

constraint on $\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}$ and $\Sigma_{\boldsymbol{\theta}_{\xi_t}}$ with the matrix $\mathbf{T}_{\xi_t}$.

$$\begin{aligned}\log q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t|\boldsymbol{y}_t, .) \propto & -\frac{1}{2}(d_{\boldsymbol{\xi}_t} - 1)\log(2\pi) - \frac{1}{2}\Bigg([\boldsymbol{\theta}_{\xi_t} - \mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}]^T(\mathbf{T}_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})^+[\boldsymbol{\theta}_{\xi_t} - \mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}]\Bigg) + \\ & - \frac{1}{2}\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})) + \sum_j \xi_{tj}(\log\kappa_j)^{(1)} + \\ & + \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}^T\mathbf{T}_{\xi_t}^T(\mathbf{T}_{\xi_t}^T\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})^+\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}} + (a_{\boldsymbol{\psi}_t}\log(b_{\boldsymbol{\psi}_t}) - \log(\Gamma(a_{\boldsymbol{\psi}_t}))\sum_j \xi_{tj} + \\ & - \sum_j (a_{\boldsymbol{\psi}_t} + 1)\xi_{tj}\log(\psi_{tj}) - b_{\boldsymbol{\psi}_t}\sum_j \xi_{tj}\psi_{tj}^{-1} + \sum_j (1 - \xi_{tj})(\log(1 - \kappa_j))^{(1)}\end{aligned}$$

We can then identify the singular multivariate density

$$\begin{aligned}\log q(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t|\boldsymbol{y}_t, .) \propto & -\frac{1}{2}(d_{\boldsymbol{\xi}_t} - 1)\log(2\pi) - \frac{1}{2}\log\big(\det{}^*(\mathbf{T}_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})\big) + \frac{1}{2}\log\big(\det{}^*(\mathbf{T}_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})\big) + \\ & - \frac{1}{2}\Bigg([\boldsymbol{\theta}_{\xi_t} - \mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}]^T(\mathbf{T}_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})^+[\boldsymbol{\theta}_{\xi_t} - \mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}]\Bigg) + \\ & + \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}^T\mathbf{T}_{\xi_t}^T(\mathbf{T}_{\xi_t}^T\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})^+\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}} + \\ & - \frac{1}{2}\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})) + \sum_j \xi_{tj}(\log\kappa_j)^{(1)} + \sum_j (1 - \xi_{tj})(\log(1 - \kappa_j))^{(1)} + \\ & + (a_{\boldsymbol{\psi}_t}\log(b_{\boldsymbol{\psi}_t}) - \log(\Gamma(a_{\boldsymbol{\psi}_t}))\sum_j \xi_{tj} - \sum_j (a_{\boldsymbol{\psi}_t} + 1)\xi_{tj}\log(\psi_{tj}) - b_{\boldsymbol{\psi}_t}\sum_j \xi_{tj}\psi_{tj}^{-1}\end{aligned} \tag{7.7.54}$$

which can be expressed as

$$\begin{aligned}q(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t|.) \propto & \ \mathrm{SMVN}_{d_{\boldsymbol{\xi}_t}}(\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}, \mathbf{T}_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})\delta_0(\bar{\boldsymbol{\xi}}_t)\times \\ & \exp\Bigg(\frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}^T\mathbf{T}_{\xi_t}(\mathbf{T}_{\xi_t}^T\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})^+\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}} + \frac{1}{2}\log\big(\det{}^*(\mathbf{T}_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})\big) + \\ & - \frac{1}{2}\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})) + \sum_j \xi_{tj}(\log\kappa_j)^{(1)} + \sum_j (1 - \xi_{tj})(\log(1 - \kappa_j))^{(1)} + \\ & + (a_{\boldsymbol{\psi}_t}\log(b_{\boldsymbol{\psi}_t}) - \log(\Gamma(a_{\boldsymbol{\psi}_t}))\sum_j \xi_{tj} - \sum_j (a_{\boldsymbol{\psi}_t} + 1)\xi_{tj}\log(\psi_{tj}) - b_{\boldsymbol{\psi}_t}\sum_j \xi_{tj}\psi_{tj}^{-1}\Bigg)\end{aligned} \tag{7.7.55}$$

We can see that the ξ_{tj} does not follow an independent Bernoulli density because the update

with respect to ξ_{tj} is dependent on the other $\xi_{tj'}$ values. There is also an issue with separating the elements of $\boldsymbol{\xi}_t$ from the pseudo determinant $\log\big(\det^*(\mathbf{T}_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})\big)$ and the first term.

Thus

$$q(\boldsymbol{\theta}_t|\boldsymbol{\xi}_t,\boldsymbol{\psi}_t) = \text{ SMVN}(\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}},\mathbf{T}_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})\delta_0(\theta_{\bar{\xi}_t}) \tag{7.7.56}$$

and

$$\begin{aligned}\log q(\boldsymbol{\theta}_t,\boldsymbol{\psi}_t,\boldsymbol{\xi}_t|.) \propto \log\big(\text{ SMVN}(\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}},\mathbf{T}_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})\delta_0(\theta_{\bar{\xi}_t})\big) + \frac{1}{2}\boldsymbol{\mu}^T_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t}(\mathbf{T}^T_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})^+\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}+\\ + \frac{1}{2}\log\big(\det^*(\mathbf{T}_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})\big) - \frac{1}{2}\log(\det^*(\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})) + \sum_j \xi_{tj}(\log\kappa_j)^{(1)}+\\ + \sum_j(1-\xi_{tj})(\log(1-\kappa_j))^{(1)} + (a_{\boldsymbol{\psi}_t}\log(b_{\boldsymbol{\psi}_t}) - \log(\Gamma(a_{\boldsymbol{\psi}_t}))\sum_j \xi_{tj}+\\ - \sum_j(a_{\boldsymbol{\psi}_t}+1)\xi_{tj}\log(\psi_{tj}) - b_{\boldsymbol{\psi}_t}\sum_j \xi_{tj}\psi_{tj}^{-1}\end{aligned} \tag{7.7.57}$$

Only part of the update is available in closed form. The full update is performed by an MCMC move, which is described in Section 7.7.2.

$$\begin{aligned}\log q(\rho_{tk},\eta_{tk}) = \ &\mathbb{E}_{-(\rho_{tk},\eta_{tk})}[\log p(\boldsymbol{y}_t|\alpha_t,\boldsymbol{\beta}_t,\sigma_t^2,\rho_t)] + \mathbb{E}_{-(\rho_{tk},\eta_{tk})}[\log p(\rho_{tk}|\sigma_t^2,\tau,\eta_{tk})]+\\ &+\mathbb{E}_{-(\rho_{tk},\eta_{tk})}[\log p(\eta_{tk}|\lambda)] + cst\end{aligned}$$

$$\begin{aligned}\log q(\rho_{tk},\eta_{tk}) =&\mathbb{E}_{-(\rho_{tk},\eta_{tk})}\Bigg[-\eta_{tk}\left(\frac{\tau\sigma_t^{-2}}{2}\rho_{tk}^2\right) + \eta_{tk}\left(\frac{1}{2}(\log\tau) + \frac{1}{2}\log\big(\sigma_t^{-2}\big) - \frac{1}{2}\log 2\pi + \log\lambda\right)+\\ &-\frac{\sigma_t^{-2}}{2}\left\|\boldsymbol{y}_t - \alpha_t\mathbf{1}_n - \sum_j X_j\beta_{tj} - \boldsymbol{Z}\boldsymbol{\theta}_t - \boldsymbol{W}\boldsymbol{\zeta}_{tg} - \sum_{h<t,h\neq k}\boldsymbol{u}_h\rho_{th} - \boldsymbol{u}_k\rho_{tk}\right\|^2+\\ &+(1-\eta_{tk})\delta_0(\rho_{tk}) + (1-\eta_{tk})(\log(1-\lambda)\Bigg] + cst\end{aligned}$$

Expanding the expectation gives

$$\begin{aligned}
\log q(\rho_{tk}, \eta_{tk}) \propto & -\frac{1}{2}(\sigma_t^{-2})^{(1)}\mathbb{E}_{-(\rho_{tk},\eta_{tk})}\left[\|\boldsymbol{u}_k\|^2\rho_{tk}^2 + 2\sum_{h<t,h\neq k}\boldsymbol{u}_h^T\boldsymbol{u}_k\rho_{tk} - 2\boldsymbol{u}_t^T\boldsymbol{u}_k\rho_{tk}\right]+ \\
& - \eta_{tk}\left(\frac{(\tau)^{(1)}(\sigma_t^{-2})^{(1)}}{2}\rho_{tk}^2\right) + \eta_{tk}\left(\frac{1}{2}(\log\tau)^{(1)} + \frac{1}{2}(\log\sigma_t^{-2})^{(1)} - \frac{1}{2}\log 2\pi + \right. \\
& \left. + (\log\lambda)^{(1)}\right) + (1-\eta_{tk})\delta_0(\rho_{tk}) + (1-\eta_{tk})(\log(1-\lambda))^{(1)} \\
\propto & \eta_{tk}\left(- \frac{1}{2}(\sigma_t^{-2})^{(1)}\left(\left(\|\boldsymbol{u}_k\|^{(2)} + (\tau)^{(1)}\right)\rho_{tk}^2 + \right.\right. \\
& \left.\left. - 2\left((\boldsymbol{u}_t)^{T(1)}(\boldsymbol{u}_k)^{(1)} - \sum_{h<t,h\neq k}(\boldsymbol{u}_h)^{T(1)}(\boldsymbol{u}_k)^{(1)}(\rho_{th})^{(1)}\right)\rho_{tk}\right)\right)+ \\
& - \eta_{tk}\left[\frac{(\tau)^{(1)}(\sigma_t^{-2})^{(1)}}{2}\rho_{tk}^2\right] + \eta_{tk}\left(\frac{1}{2}(\log\tau)^{(1)} + \frac{1}{2}(\log\sigma_t^{-2})^{(1)} - \frac{1}{2}\log 2\pi + \right. \\
& \left. + (\log\lambda)^{(1)}\right) + (1-\eta_{tk})\delta_0(\rho_{tk}) + (1-\eta_{tk})(\log(1-\lambda))^{(1)}.
\end{aligned}$$

The parameter updates for the q density are

$$\begin{aligned}
\mu_{\rho_{tk}} &= \frac{(\boldsymbol{u}_t)^{T(1)}(\boldsymbol{u}_k)^{(1)} - \sum_{h<t,h\neq k}(\boldsymbol{u}_h)^{T(1)}(\boldsymbol{u}_k)^{(1)}(\rho_{th})^{(1)}}{\|\boldsymbol{u}_k\|^{(2)} + (\tau)^{(1)}} \\
\sigma^2_{\rho_{tk}} &= \left((\sigma_t^{-2})^{(1)}\left((\tau)^{(1)} + \|\boldsymbol{u}_k\|^{(2)}\right)\right)^{-1}
\end{aligned}$$

The joint q density is proportional to

$$\begin{aligned}
q(\rho_{tk}, \eta_{tk}) \propto & \left[(2\pi\sigma^2_{\rho_{tk}})^{-1/2}\exp\left\{-\frac{1}{2\sigma^2_{\rho_{tk}}}(\rho_{tk}-\mu_{\rho_{tk}})^2\right\}\right]^{\eta_{tk}} \times \left[\delta_0(\rho_{tk})\right]^{1-\eta_{tk}} \times \\
& \left[\left\{\exp\left((\log\tau)^{(1)} + (\log\sigma_t^{-2})^{(1)}\right)\sigma^2_{\rho_{tk}}\right\}^{\frac{1}{2}}\exp\left\{\frac{\mu^2_{\rho_{tk}}}{2\sigma^2_{\rho_{tk}}}\right\}\exp\left\{(\log\lambda)^{(1)}\right\}\right]^{\eta_{tk}} \times \\
& \left[\exp\left\{(\log(1-\lambda))^{(1)}\right\}\right]^{1-\eta_{tk}}
\end{aligned}$$

and thus by calling

$$(\eta_{tk})^{(1)} = \left[1 + \sqrt{\sigma^{-2}_{\rho_{tk}}} \exp\left((\log(1-\lambda))^{(1)} - \frac{(\log \tau)^{(1)}}{2} - \frac{(\log \sigma_t^{-2})^{(1)}}{2} - (\log \lambda)^{(1)} - \frac{\mu^2_{\rho_{tk}}}{2\sigma^2_{\rho_{tk}}} \right) \right]^{-1}$$

we have under q

$$q(\rho_{tk}|\eta_{tk} = 1) = \mathcal{N}(\mu_{\rho_{tk}}, \sigma^2_{\rho_{tk}}), \quad q(\rho_{tk}|\eta_{tk} = 0) = \delta_0(\rho_{tk})$$
$$q(\eta_{tk}) = Bern((\eta_{tk})^{(1)})$$

Note that now

$$\mathbb{E}_q[\rho_{tk}] = (\rho_{tk})^{(1)} = \mu_{\rho_{tk}}(\eta_{tk})^{(1)}$$
$$\mathbb{E}_q[\eta_{tk}\rho_{tk}] = \mu_{\rho_{tk}}(\eta_{tk})^{(1)}$$
$$(\rho_{tk})^{(2)} = (\mu^2_{\rho_{tk}} + \sigma^2_{\rho_{tk}})(\eta_{tk})^{(1)}.$$

$$\log q(\sigma_t^2) = \mathbb{E}_{-\sigma_t^2}[\log p(\sigma_t^2|\tau, \nu)] + \mathbb{E}_{-\sigma_t^2}[\log p(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \sigma_t^2, \rho_t) + \sum_{k<t} \log \mathbb{E}_{-\sigma_t^2}[p(\rho_{tk}|\sigma^2, \tau, \nu)] + cst$$

$$
\begin{aligned}
\log q(\sigma_t^2) =& \frac{n}{2}\log\sigma_t^{-2} - \frac{1}{2}\sigma_t^{-2}\mathbb{E}_{-\sigma_t^2}\left[\|\boldsymbol{u}_t - \sum_{k<t}\boldsymbol{u}_k\rho_{tk}\|^2\right] + \left(\frac{\nu - T + t}{2} + 1\right)\log\sigma_t^{-2} + \\
&+ \mathbb{E}_{-\sigma_t^2}\left[-\frac{\tau\sigma_t^{-2}}{2} + \sum_{k<t}\eta_{tk}\left(\frac{1}{2}\log\sigma_t^{-2} - \frac{\tau}{2}\rho_{tk}^2\sigma_t^{-2}\right)\right] \\
=& \log\sigma_t^{-2}\left(\frac{n}{2} + \left(\frac{\nu - T + t}{2} + 1\right) + \sum_{k<t}\frac{(\eta_{tk})^{(1)}}{2}\right) - \sigma_t^{-2}\left(\frac{(\tau)^{(1)}}{2} + \frac{\|\boldsymbol{u}_t\|^{(2)}}{2} + \right. \\
&\sum_{k<t}\frac{\|\boldsymbol{u}_k\|^{(2)}}{2}(\rho_{tk})^{(2)} + \sum_{k'<k,\ k<t}(\boldsymbol{u}_k)^{T(1)}(\boldsymbol{u}_{k'})^{(1)}(\rho_{tk})^{(1)}(\rho_{tk'})^{(1)} + \\
&\left. -\sum_{k<t}(\boldsymbol{u}_t)^{T(1)}(\boldsymbol{u}_k)^{(1)}(\rho_{tk})^{(1)} + \sum_{k<t}\frac{(\tau)^{(1)}(\rho_{tk})^{(2)}}{2}\right)
\end{aligned}
$$

$$
q(\sigma_t^2) = Inv - Gamma(a^*_{\sigma^2,t}, b^*_{\sigma^2,t})
$$

$$
a^*_{\sigma^2,t} = \sum_{k<t}\frac{(\eta_{tk})^{(1)}}{2} + \frac{\nu - T + t}{2} + \frac{n}{2} \tag{7.7.58}
$$

$$
\begin{aligned}
b^*_{\sigma^2,t} =& \frac{(\tau)^{(1)}}{2} + \frac{(\tau)^{(1)}}{2}\sum_{k<t}(\rho_{tk})^{(2)} + \frac{||\boldsymbol{u}_t||^{(2)}}{2} + \sum_{k<t}\frac{||\boldsymbol{u}_k||^{(2)}}{2}(\rho_{tk})^{(2)} + \\
&+ \sum_{k'<k,\ k<t}(\boldsymbol{u}_k)^{(1)T}(\boldsymbol{u}_{k'})^{(1)}(\rho_{tk})^{(1)}(\rho_{tk'})^{(1)} - \sum_{k<t}(\boldsymbol{u}_t)^{(1)T}(\boldsymbol{u}_k)^{(1)}(\rho_{tk})^{(1)}
\end{aligned} \tag{7.7.59}
$$

where

$$
(\sigma_t^{-2})^{(1)} = \frac{a^*_{\sigma^2,t}}{b^*_{\sigma^2,t}} \tag{7.7.60}
$$

$$
(\log\sigma_t^{-2})^{(1)} = \Psi(a^*_{\sigma^2,t}) - \log b^*_{\sigma^2,t} \tag{7.7.61}
$$

$$\begin{aligned}\log q(\kappa_j) =& \mathbb{E}_{-\kappa_j}\left[\sum_t \log p(\xi_{tj}|\kappa_j) + \log p(\kappa_j)\right] + cst \\ =& \mathbb{E}_{-\kappa_j}\Bigg[\sum_t \Big(\xi_{tj}\log(\kappa_j) + (1-\xi_{tj})\log(1-\kappa_j)\Big)\mathrm{I}\Big[\sum_j \xi_{tj} \neq 1\Big] + \\ &+ (a_j - 1)\log(\kappa_j) + (b_j - 1)\log(1-\kappa_j)\Bigg] + cst\end{aligned}$$

As the update for $\boldsymbol{\xi}_t$ from the construction of the MCMC and the singular multivariate normal is

$$\mathbb{E}_q[\boldsymbol{\xi}_t] = \mathbb{E}_q\left[\boldsymbol{\xi}_t\ \mathrm{I}\Big[\sum_j \xi_{tj} \neq 1\Big]\right] = (\boldsymbol{\xi}_t)^{(1)}, \tag{7.7.62}$$

the update can be solved in closed form, using the jth element of the MCMC expectations of each vector. The dependency between each of the elements in the vector $\boldsymbol{\xi}_t$ prevents a simple marginal expectation for $\boldsymbol{\xi}_t$.

$$\log q(\kappa_j) = (\sum_t (\xi_{tj})^{(1)} + a_\kappa - 1)\log(\kappa_j) + (T - \sum_t (\xi_{tj})^{(1)} + b_\kappa - 1)\log(1-\kappa_j) + cst$$

$$q(\kappa_j) = Beta(a^*_{\kappa,j}, b^*_{\kappa,j}) \tag{7.7.63}$$

with parameters

$$a^*_{\kappa,j} = a_\kappa + \sum_t (\xi_{tj})^{(1)} \tag{7.7.64}$$

$$b^*_{\kappa,j} = b_\kappa + T - \sum_t (\xi_{tj})^{(1)} \tag{7.7.65}$$

where

$$
\begin{aligned}
(\kappa_j)^{(1)} &= a^*_{\kappa,j} / \left(a^*_{\kappa,j} + b^*_{\kappa,j}\right) = a^*_{\kappa,j} / \left(a_\kappa + b_\kappa + 1\right) \\
(\log \kappa_s)^{(1)} &= \Psi(a^*_{\kappa,j}) - \Psi(a^*_{\kappa,j} + b^*_{\kappa,j}) \\
(\log(1-\kappa_j))^{(1)} &= \Psi(b^*_{\kappa,j}) - \Psi(a^*_{\kappa,j} + b^*_{\kappa,j})
\end{aligned}
\tag{7.7.66}
$$

where $\Psi(\cdot)$ is the digamma function.

$$
\log q(\lambda) = \mathbb{E}_{-\lambda}\left[\sum_t \sum_{k<t} \log p(\eta_{tk}|\lambda)\right] + \mathbb{E}_{-\lambda}[\log p(\lambda)] + cst \tag{7.7.67}
$$

$$
\begin{aligned}
\log q(\lambda) =& \sum_t \sum_{k<t} \Bigg((\eta_{tk})^{(1)} \log(\lambda) + (1-\eta_{tk})^{(1)} (\log(1-\lambda) + \\
& (a_\lambda - 1)\log\lambda + (b_\lambda - 1)\log(1-\lambda)\Bigg) \\
=& \left(\sum_t \sum_{k<t} (\eta_{tk})^{(1)} + a_\lambda - 1\right)\log\lambda + \left(\sum_t \sum_{k<t} (1-\eta_{tk})^{(1)} + b_\lambda - 1\right)\log(1-\lambda)
\end{aligned}
$$

$$
\lambda = Beta(a^*_\lambda, b^*_\lambda)
$$

As $\sum_t \sum_{k<t} 1 = T(T+1)/2$.

$$
\begin{aligned}
a_\lambda =& \sum_t \sum_{k<t} (\eta_{tk})^{(1)} + a_\lambda \\
b_\lambda =& \sum_t \sum_{k<t} (1-\eta_{tk})^{(1)} + b_\lambda
\end{aligned}
$$

with updated expectations

$$(\lambda)^{(1)} = \frac{a_\lambda^*}{a_\lambda^* + b_\lambda^*} \tag{7.7.68}$$

$$(\log \lambda)^{(1)} = \Psi(a_\lambda^*) - \Psi(a_\lambda^* + b_\lambda^*) \tag{7.7.69}$$

$$(\log(1-\lambda))^{(1)} = \Psi(b_\lambda^*) - \Psi(a_\lambda^* + b_\lambda^*). \tag{7.7.70}$$

$$\log q(\tau) = \mathbb{E}_{-\tau}\Big[\sum_{t=1}^{T} \log p(\sigma_t^2|\tau,\nu) + \sum_t \sum_{k<t} \log p(\rho_{tk}|\eta_{tk}\sigma_t^2,\tau) + \log p(\tau)\Big] + cst$$

$$\log q(\tau) \propto \sum_{t=1}^{T} \Big(\frac{\nu - T + t}{2}\log\tau - \frac{\tau}{2}(\sigma_t^{-2})^{(1)}\Big) + \sum_t \sum_{k<t} \Big(\frac{\log\tau}{2} - \frac{\tau(\sigma_t^{-2})^{(1)}}{2}(\rho_{tk})^{(2)}\Big)(\eta_{tk})^{(1)} +$$
$$+ (a_\tau - 1)\log\tau - b_\tau\tau$$

Simplifying

$$\log q(\tau) \propto \Big(\sum_{t=1}^{T} \frac{\nu - T + t}{2} + \sum_t \sum_{k<t} \frac{(\eta_{tk})^{(1)}}{2} + a_\tau - 1\Big)\log\tau +$$
$$+ \Big(\sum_{t=1}^{T} \frac{(\sigma_t^{-2})^{(1)}}{2} + \sum_t \sum_{k<t} \frac{1}{2}(\eta_{tk})^{(1)}(\rho_{tk})^{(2)}(\sigma_t^{-2})^{(1)}\Big) \tag{7.7.71}$$

Since $\sum_t t = \frac{T(T+1)}{2}$,

$$q(\tau) = Gamma\,(a_\tau^*, b_\tau^*)$$

with parameters

$$a_\tau^* = a_\tau + \frac{T(\nu - (T+1)/2)}{2} + \sum_t \sum_{k<t} \frac{(\eta_{tk})^{(1)}}{2} \tag{7.7.72}$$

$$b_\tau^* = b_\tau + \frac{1}{2} \sum_{t=1}^{T} (\sigma_t^{-2})^{(1)} \left(1 + \sum_{k<t} (\rho_{tk})^{(2)} \right) \tag{7.7.73}$$

where

$$(\tau)^{(1)} = a_\tau^* / b_\tau^* \tag{7.7.74}$$

$$(\log \tau)^{(1)} = \Psi(a_\tau^*) - \log b_\tau^* \tag{7.7.75}$$

$$\log q(w_t) = \mathbb{E}_{-w_t} \left[\sum_s \log p(\beta_{ts} | w_t, \gamma_{ts}) + \log p(w_t | a_w, b_w) \right] + cst$$

The update is

$$\begin{aligned} q(w_t) \propto & \mathbb{E}_{-w_t} \Big[\sum_s -\frac{\gamma_{ts}}{2} \Big(\log w_t - w_t^{-1} \frac{\beta_{ts}^2}{2} \Big) \Big] + \mathbb{E}_{-w_t} \Big[(-a_w - 1) \log w_t - b_w w_t^{-1} \Big] \\ & \propto \log w_t \Big(-\frac{1}{2} \sum_s (\gamma_{ts})^{(1)} - a_w - 1 \Big) - w_t^{-1} \Big(\frac{1}{2} \sum_s (\beta_{ts})^{(2)} + (b_w)^{(1)} \Big) \end{aligned} \tag{7.7.76}$$

thus

$$q(w_t) = Inv - Gamma(a_{w,t}^*, b_{w,t}^*) \tag{7.7.77}$$

with parameters

$$a_{w,t}^* = \frac{1}{2} \sum_s (\gamma_{ts})^{(1)} + a_w \tag{7.7.78}$$

$$b_{w,t}^* = \frac{1}{2} \sum_s (\beta_{ts})^{(2)} + (b_w)^{(1)} \tag{7.7.79}$$

where $(\beta_{ts})^{(2)} = \mathbb{E}_q[\beta_{ts}^2 \gamma_{ts}]$. The prior guarantees the the constraint $a_{w,t} > 0$ even if $\sum_s (\gamma_{ts})^{(1)} = 0$.

$$\log q(w_{\alpha_t}) = \mathbb{E}_{-w_{\alpha_t}} \left[\log p(\alpha_t | w_{\alpha_t}) + \log p(w_{\alpha_t} | a_\alpha, b_\alpha)\right] + cst$$

The update is

$$\begin{aligned} q(w_t) &\propto \mathbb{E}_{-w_{\alpha_t}} \Big[-\frac{1}{2} \log w_{\alpha_t} - \frac{w_{\alpha_t}^{-1}}{2} + (-a_\alpha - 1) \log w_{\alpha_t} - b_\alpha w_{\alpha_t}^{-1} \Big] \\ &\propto \log w_{\alpha_t} \Big(-\frac{1}{2} - a_\alpha - 1 \Big) - w_{\alpha_t}^{-1} \Big((\alpha_t)^{(2)} + (b_\alpha)^{(1)} \Big) \end{aligned} \tag{7.7.80}$$

thus

$$q(w_t) = Inv - Gamma(a_{w,t}^*, b_{w,t}^*) \tag{7.7.81}$$

with parameters

$$a_{w,t}^* = \frac{1}{2} + a_\alpha \tag{7.7.82}$$

$$b_{w,t}^* = \frac{1}{2} (\alpha_t)^{(2)} + (b_\alpha)^{(1)} \tag{7.7.83}$$

where $(\alpha_t)^{(2)} = \mathbb{E}_q[\alpha_t^2]$.

$$\begin{aligned} \log q(v_t) &= \mathbb{E}_{(-v_t)} \left[\sum_g \log p(\boldsymbol{\zeta}_{tg} | v_t, \chi_{tg}) + \log p(v_t | a_v, b_v) \right] + cst \\ &= \mathbb{E}_{-v_t} \Big[\sum_g \chi_{tg} \Big(-\frac{m_g}{2} \log v_t - v_t^{-1} \frac{\boldsymbol{\zeta}_{tg}^T \boldsymbol{\zeta}_{tg}}{2} \Big) + (-a_v - 1) \log v_t - b_v v_t^{-1} \Big] + cst \\ &\propto \log v_t \Big(-\frac{1}{2} \Big\{ \sum_g m_g (\chi_{tg})^{(1)} \Big\} - a_v - 1 \Big) - v_t^{-1} \Big(\frac{1}{2} \Big(\sum_g (\boldsymbol{\zeta}_{tg}^T \boldsymbol{\zeta}_{tg})^{(1)} \Big) + (b_v)^{(1)} \Big) \end{aligned}$$

thus

$$q(v_t) = Inv - Gamma(a_v^*, b_v^*) \tag{7.7.84}$$

with parameters

$$a_{v,t}^* = \frac{1}{2}\Big(\sum_g m_g(\chi_{tg})^{(1)}\Big) + a_v \tag{7.7.85}$$

$$b_{v,t}^* = \frac{1}{2}\Big(\sum_g (\boldsymbol{\zeta}_{tg}^T \boldsymbol{\zeta}_{tg})^{(1)}\Big) + (b_v)^{(1)} \tag{7.7.86}$$

$$\log q(b_w) = \mathbb{E}_{-b_w}\Big[\sum_{t=1}^{T} \log p(w_t|a_w, b_w) + \log p(b_w|a_b, b_b)\Big] + cst \tag{7.7.87}$$

$$\begin{aligned}
\log q(b_w) &\propto \mathbb{E}_{-b_w}\Big[\sum_t \Big(a_w \log b_w - b_w w_t^{-1}\Big) + (a_b - 1)\log b_w - b_b b_w\Big] \\
&\propto T a_w \log b_w - b_w \sum_t w_t^{(-1)} + (a_b - 1)\log b_w - b_b b_w \\
&\propto \log b_w (T a_w + a_b - 1) - b_w (\sum_t w_t^{(-1)} + b_b)
\end{aligned} \tag{7.7.88}$$

thus

$$q(b_w) = Gamma(a_b^*, b_b^*)$$

with parameters

$$a_b^* = T a_w + a_b \tag{7.7.89}$$

$$b_b^* = \sum_t w_t^{(-1)} + b_b \tag{7.7.90}$$

where

$$(b_w)^{(1)} = a_b^*/b_b^* \tag{7.7.91}$$

$$(\log b_w)^{(1)} = \Psi(a_b^*) - \log b_b^* \tag{7.7.92}$$

$$\log q(b_v) = \mathbb{E}_{-b_v}\left[\sum_{t=1}^{T} \log p(v_t|a_v, b_v) + \log p(b_v|a_{b_v}, b_{b_v})\right] + cst \tag{7.7.93}$$

$$\begin{aligned}
\log q(b_v) &\propto \mathbb{E}_{-b_v}\left[\sum_t \left(a_v \log b_v - b_v v_t^{-1}\right) + (a_{b_v} - 1)\log b_v - b_{bv} b_v\right] \\
&\propto T a_v \log b_v - b_v \sum_t v_t^{(-1)} + (a_{b_v} - 1)\log b_v - b_{b_v} b_v \\
&\propto \log b_v (T a_v + a_{b_v} - 1) - b_v \left(\sum_t v_t^{(-1)} + b_{b_v}\right)
\end{aligned} \tag{7.7.94}$$

thus

$$q(b_v) = Gamma(a_{b_v}^*, b_{b_v}^*)$$

with parameters

$$a_{b_v}^* = T a_v + a_{b_v} \tag{7.7.95}$$

$$b_{b_v}^* = \sum_t (v_t)^{(-1)} + b_{b_v} \tag{7.7.96}$$

where

$$(b_v)^{(1)} = a^*_{b_v}/b^*_{b_v} \tag{7.7.97}$$

$$(\log b_v)^{(1)} = \Psi(a^*_{b_v}) - \log b^*_{b_v} \tag{7.7.98}$$

$$\log q(b_\alpha) = \mathbb{E}_{-b_\alpha}\left[\sum_{t=1}^{T} \log p(w_{\alpha t}|a_\alpha, b_\alpha) + \log p(b_\alpha|a_{b_\alpha}, b_{b_\alpha})\right] + cst \tag{7.7.99}$$

$$\begin{aligned}
\log q(b_\alpha) &\propto \mathbb{E}_{-b_\alpha}\left[\sum_t \left(a_\alpha \log b_\alpha - b_\alpha w^{-1}_{\alpha_t}\right) + (a_{b_\alpha} - 1)\log b_\alpha - b_{b_\alpha} b_\alpha\right] \\
&\propto T a_\alpha \log b_\alpha - b_\alpha \sum_t (w_{\alpha_t})^{(-1)} + (a_{b_\alpha} - 1)\log b_\alpha - b_{b_\alpha} b_\alpha \\
&\propto \log b_\alpha (T a_\alpha + a_{b_\alpha} - 1) - b_\alpha \left(\sum_t (w_{\alpha_t})^{(-1)} + b_{b_\alpha}\right)
\end{aligned} \tag{7.7.100}$$

thus

$$q(b_\alpha) = Gamma(a^*_{b_\alpha}, b^*_{b_\alpha})$$

with parameters

$$a^*_{b_\alpha} = T a_\alpha + a_{b_\alpha} \tag{7.7.101}$$

$$b^*_{b_\alpha} = \sum_t (w_{\alpha_t})^{(-1)} + b_{b_\alpha} \tag{7.7.102}$$

where

$$(b_\alpha)^{(1)} = a^*_{b_\alpha}/b^*_{b_\alpha} \tag{7.7.103}$$

$$(\log b_\alpha)^{(1)} = \Psi(a^*_{b_\alpha}) - \log b^*_{b_\alpha} \tag{7.7.104}$$

$$\log q(\omega_s) = \mathbb{E}_{-\omega_s}\left[\sum_t \log p(\gamma_{ts}|\omega_s) + \log p(\omega_s)\right] + cst \tag{7.7.105}$$

$$\begin{aligned}\log q(\omega_s) &\propto \sum_t (\gamma_{ts})^{(1)} \log \omega_s + (1-\gamma_{ts})^{(1)} \log(1-\omega_s) + (a_\omega - 1)\log\omega_s + (b_\omega - 1)\log(1-\omega_s) \\ &\propto \left(a_\omega + \sum_t (\gamma_{ts})^{(1)} - 1\right)\log\omega_s + \left(b_\omega + T - \sum_t (\gamma_{ts})^{(1)} - 1\right)\log(1-\omega_s).\end{aligned}$$

which implies that

$$q(\omega_s) = Beta(a^*_{w,s}, b^*_{w,s}) \tag{7.7.106}$$

with parameters

$$a^*_{\omega,s} = a_\omega + \sum_t (\gamma_{ts})^{(1)} \tag{7.7.107}$$

$$b^*_{\omega,s} = b_\omega + T - \sum_t (\gamma_{ts})^{(1)} \tag{7.7.108}$$

where

$$(\omega_s)^{(1)} = a^*_{\omega,s}/\left(a^*_{\omega,s} + b^*_{\omega,s}\right) = a^*_{\omega,s}/\left(a_\omega + b_\omega + T\right) \tag{7.7.109}$$

$$(\log\omega_s)^{(1)} = \Psi(a^*_{\omega,s}) - \Psi(a^*_{\omega,s} + b^*_{\omega,s})$$

$$(\log(1-\omega_s))^{(1)} = \Psi(b^*_{\omega,s}) - \Psi(a^*_{\omega,s} + b^*_{\omega,s})$$

where $\Psi(\cdot)$ is the digamma function.

$$\log q(\varrho_g) = \mathbb{E}_{-\varrho_g}\left[\sum_t \log p(\chi_{tg}|\varrho_g) + \log p(\varrho_g)\right] + cst \tag{7.7.110}$$

$$\begin{aligned}\log q(\varrho_g) &\propto \sum_t (\chi_{tg})^{(1)} \log \varrho_g + (1-\chi_{tg})^{(1)} \log(1-\varrho_g) + (a_\varrho - 1)\log \varrho_g + (b_\varrho - 1)\log(1-\varrho_g) \\ &\propto \left(a_\varrho + \sum_t (\chi_{tg})^{(1)} - 1\right)\log \varrho_g + \left(b_\varrho + T - \sum_t (\chi_{tg})^{(1)} - 1\right)\log(1-\varrho_g)\end{aligned}$$

which implies that

$$q(\varrho_g) = Beta(a^*_{\varrho,g}, b^*_{\varrho,g}) \tag{7.7.111}$$

with parameters

$$a^*_{\varrho,g} = a_\varrho + \sum_t (\chi_{tg})^{(1)} \tag{7.7.112}$$

$$b^*_{\varrho,g} = b_\varrho + T - \sum_t (\chi_{tg})^{(1)} \tag{7.7.113}$$

where

$$\begin{aligned}(\varrho_g)^{(1)} &= a^*_{\varrho,g} / \left(a^*_{\varrho,g} + b^*_{\varrho,g}\right) = a^*_{\varrho,g} / \left(a_{\varrho,g} + b_{\varrho,g} + T\right) \\ (\log \varrho_g)^{(1)} &= \Psi(a^*_{\varrho,g}) - \Psi(a^*_{\varrho,g} + b^*_{\omega,g}) \\ (\log(1-\varrho_g))^{(1)} &= \Psi(b^*_{\varrho,g}) - \Psi(a^*_{\varrho,g} + b^*_{\varrho,g})\end{aligned} \tag{7.7.114}$$

where $\Psi(\cdot)$ is the digamma function.

7.7.2 RJMCMC moves and model proposals

The priors for the parameters associated with the microbiome features, the indicator vectors $\boldsymbol{\xi}_t$ and set of scale parameters $\boldsymbol{\psi}_t$, prevents a conjugate update for $q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t)$. Here, we briefly outline the steps defined in Scott and Lewin (2021), which allow us to introduce an MCMC step

to the CAVI algorithm in the multivariate response model. This provides expectations from the $q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t)$ approximating density which are not available analytically.

Approximating the q variational density to guide the RJMCMC

A univariate approximation of the VI posterior distribution $q(\boldsymbol{\xi}_t, \boldsymbol{\psi}_t|\boldsymbol{Y})$, relative to the jth element, is used to guide the RJMCMC to search the large binary space.

$$
\begin{aligned}
q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t|\boldsymbol{Y}) &\propto q(\boldsymbol{\theta}_t|\boldsymbol{\psi}_t, \boldsymbol{\xi}_t, \boldsymbol{Y}) q(\boldsymbol{\psi}_t, \boldsymbol{\xi}_t|\boldsymbol{Y}) \\
&\propto \text{SMVN}(\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}, \mathbf{T}_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})\delta_0(\theta_{\bar{\xi}_t}) \exp\left(\frac{1}{2}\boldsymbol{\mu}^T_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t}(\mathbf{T}^T_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})^+\mathbf{T}_{\xi_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}} + \right. \\
&\quad + \frac{1}{2}\log\big(\det{}^*(\mathbf{T}_{\xi_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\xi_t})\big) - \frac{1}{2}\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})) + \sum_j \xi_{tj}(\log\kappa_j)^{(1)} + \\
&\quad + \sum_j (1-\xi_{tj})(\log(1-\kappa_j))^{(1)} + (a_{\boldsymbol{\psi}_t}\log(b_{\boldsymbol{\psi}_t}) - \log(\Gamma(a_{\boldsymbol{\psi}_t}))\sum_j \xi_{tj} + \\
&\quad \left. - \sum_j (a_{\boldsymbol{\psi}_t}+1)\xi_{tj}\log(\psi_{tj}) - b_{\boldsymbol{\psi}_t}\sum_j \xi_{tj}\psi^{-1}_{tj}\right). \qquad (7.7.115)
\end{aligned}
$$

These normalised probabilities provide proposal probabilities, informed by the likelihood, in a birth-death and swap sampling scheme.

The pseudo determinant is approximated by removing the constraints $\mathbf{T}_{\boldsymbol{\xi}_t}$ and taking the MCMC expectation conditional on $\xi_{tj} = 1$. So, for the jth element, the approximation is

$$
\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})) \approx \{\log(\psi_{tj})\}^{\{1\}}_{\emptyset}, \qquad (7.7.116)
$$

where the curly brackets $\{\}$ denote an MCMC expectation and $\emptyset$ defines an expectation over all nonzero values. A similar approach can be used to approximate the determinant containing Σ_{θ_ξ}

$$
\log\big(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\boldsymbol{\xi}_t})\big) \approx \log\big(\bar{\sigma}^2_{\theta,tj}\big),
$$

where $\bar{\sigma}^2_{\theta,tj}$ is the non-zero variance average for the j term over the MCMC iterations, obtained

after extracting the diagonal from $\Sigma_{\theta_{(\xi,\psi)}}$, at each iteration. If the jth term has not been included in the model, the term is approximated by

$$\log\big(\det^*(\mathbf{T}_{\boldsymbol{\xi}_t}\Sigma_{\theta_\xi}\mathbf{T}_{\boldsymbol{\xi}_t})\big) \approx \log\left(\left[\|Z_j\|^2\left((\sigma_t^{-2})^{(1)} + \sum_{k>t}(\rho_{kt})^{(2)}(\sigma_k^{-2})^{(1)}\right)\right]^{-1}\right). \tag{7.7.117}$$

This is the variance term for the auxiliary parameter Ω_{tj} when $\mathbb{E}_q[\Delta_{tj}^{-1}|\Upsilon_{tj}=0]$, which is derived in Section 7.7.2. By approximating $\Sigma_{\boldsymbol{\theta}_{\xi_t}}$ to a scalar for each jth element, the matrix dot product reduces to

$$\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}^T\mathbf{T}_{\boldsymbol{\xi}_t}(\mathbf{T}_{\boldsymbol{\xi}_t}^T\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\boldsymbol{\xi}_t})^+\mathbf{T}_{\boldsymbol{\xi}_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}} \approx \bar{\sigma}^2_{\theta,tj}\Big(\sum_j(1-1/d_{\boldsymbol{\xi}_t})\mu^2_{\theta_{\xi_{tj}}} - 2\sum_{j<j'}(\mu_{\theta_{\xi_{tj'}}}\mu_{\theta_{\xi_{tj}}}/d_{\boldsymbol{\xi}_t})\Big). \tag{7.7.118}$$

To account for the cross product terms, a combination of conditional expectations and marginal expectations which shrink the values in proportion to its probability of being zero, is used. As $\boldsymbol{\xi}_{tj}$ can not be separated from the sum in the numerator $d_{\boldsymbol{\xi}_t}$, two approximations of the matrix dot product are used, conditional on the expectation from the previous chain.

Defining the expectations with respect to the parameter currently being updated from the previous MCMC by a curly bracket as:

- $\{\cdot\}_{\not\emptyset}^{\{1\}}$:Conditional expectation $\xi_{tj}=1$, a weighted average of the nonzero terms from previous chain.
- $\{\cdot\}^{\{1\}}$:Expectation wrt q from the previous chain.

The approximation of the dot product $\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}^T\mathbf{T}_{\boldsymbol{\xi}_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}$ is thus approximately equal to

$$\begin{cases}\bar{\sigma}^{-2}_{\theta,tj}\left(\sum_j(1-\frac{1}{\{q_\xi\}^{\{1\}}})\xi_{tj}(\{\mu_{\theta_{tj}}\}_{\not\emptyset}^{\{1\}})^2 - \frac{2}{\{d_{\boldsymbol{\xi}_t}\}^{\{1\}}}\sum_{j<j'}\xi_{tj}\{\mu_{\theta_{\xi_{tj}}}\}_{\not\emptyset}^{\{1\}}\{\mu_{\theta_{\xi_{tj'}}}\}^{\{1\}}\right) & \{d_{\boldsymbol{\xi}_t}\}^{\{1\}}\geq 2\\ \bar{\sigma}^{-2}_{\theta,tj}\sum_j\xi_{tj}(\{\mu_{\theta_{tj}}\}_{\not\emptyset}^{\{1\}})^2 & \{d_{\boldsymbol{\xi}_t}\}^{\{1\}}<2\end{cases}$$

Although $\{d_{\boldsymbol{\xi}_t}\in\mathbb{N}_0|d_{\boldsymbol{\xi}_t}\leq d, d_{\boldsymbol{\xi}_t}\neq 1\}$, the MCMC expectation $\{d_{\boldsymbol{\xi}_t}\}^{\{1\}}$ is in the positive real numbers so we threshold on 2.When $\{d_{\boldsymbol{\xi}_t}\}^{\{1\}}>2$ the probabilities used in the proposal distribution

for the RJMCMC are

$$\tilde{p}(\xi_{tj}=1|\boldsymbol{\vartheta}) \equiv \Bigg[\exp\Bigg\{ -\frac{1}{2\bar{\sigma}^2_{\theta,j}} \Bigg((1-1/\{d_{\boldsymbol{\xi}_t}\}^{\{1\}})(\{\mu_{\theta_{tj}}\}^{\{1\}}_{\emptyset})^2 - \frac{2}{\{d_{\boldsymbol{\xi}_t}\}^{\{1\}}} \{\mu_{\theta_{\xi_j}}\}^{\{1\}}_{\emptyset} \sum_{j' \neq j} \{\mu_{\theta_{\boldsymbol{\xi}_{tj'}}}\}^{\{1\}} \Bigg) + \\ + (\log\Gamma(a_{\boldsymbol{\psi}_t}) - a_{\boldsymbol{\psi}_t}\log b_{\boldsymbol{\psi}_t}) + (a_{\boldsymbol{\psi}_t}+1)(\log\psi_{tj})^{\{1\}}_{\emptyset} + b_{\boldsymbol{\psi}_t}(\psi^{-1}_{tj})^{\{1\}}_{\emptyset} + \\ + (\log(1-\kappa_j))^{(1)} - \frac{1}{2}\log\big(\sigma^2_{\theta,tj}\big) + \frac{1}{2}(\log\psi_{tj})^{\{1\}}_{\emptyset} - (\log\kappa_j)^{(1)} \Bigg\} + 1 \Bigg]^{-1}, \qquad (7.7.119)$$

which contains the free variational expectations and an MCMC condtional expectation from the previous iterations.

Pseudo updates

Samples from the intractable variational approximating posterior $q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t)$ are simulated by an MCMC step. The move types in the RJMCMC use an element-wise approximation of the joint density $q(\boldsymbol{\xi}_t, \boldsymbol{\psi}_t|\boldsymbol{Y})$. For the proposal distribution of $\boldsymbol{\psi}_t$, we use the model likelihood and an unconstrained approximation to the constrained priors by defining auxiliary parameters (upper case Greek letters). These are versions of the constrained parameters, which ignore the sum to zero constraint. We derive pseudo variational updates from an unconstrained model with a simpler prior parameterisation, then use the variational approximating distribution of the relevant auxiliary parameter as our proposal for $\boldsymbol{\psi}_t$. These updates are refined by the full VI updates which account for the constraint at each iteration. The parameter κ_j and the hyperparameters a_{Δ_t} and b_{Δ_t}, which are set to $a_{\boldsymbol{\psi}_t}$ and $b_{\boldsymbol{\psi}_t}$ respectively, provide a link back to the constrained model.

The auxiliary parameters for the unconstrained model are from the following prior parameteri-

sation

$$p(\Omega_{tj}|\Delta_{tj},\Upsilon_{tj}) = \left[\frac{1}{(2\pi\Delta_{tj})^{(-1/2)}}\exp\left(-\frac{1}{2\Delta_{tj}}\Omega_{tj}^2\right)\right]^{\Upsilon_{tj}}\delta_0(\Omega_{tj})^{1-\Upsilon_{tj}} \tag{7.7.120}$$

$$p(\Delta_{tj}|\Upsilon_{tj}) = \left[\frac{b_{\Delta_t}^{a_{\Delta_t}}}{\Gamma(a_{\Delta_t})}(\Delta_{tj})^{-a_{\Delta_t}-1}\exp\{-b_{\Delta_t}\Delta_{tj}^{-1}\}\right]^{\Upsilon_{tj}}\delta_0(\Delta_{tj})^{1-\Upsilon_{tj}} \tag{7.7.121}$$

$$P(\Upsilon_{tj}) = (\kappa_j)^{\Upsilon_{tj}}(1-\kappa_j)^{1-\Upsilon_{tj}} \tag{7.7.122}$$

The pseudo updates are subsequently derived in full. The $q(\Omega_{tj},\Upsilon_{tj})$ update is

$$q(\Omega_{tj},\Upsilon_{tj}) \propto \mathbb{E}_{(-\Omega_{tj},\Upsilon_{tj})}\Big[\log p(\boldsymbol{y}_t|.) + \log p(\Omega_{tj}|\Delta_{tj},\Upsilon_{tj}) + p(\Delta_{tj}|\Upsilon_{tj}) + p(\Upsilon_{tj})\Big]$$

after expanding and rearranging takes the form

$$\begin{aligned} q(\Omega_{tj},\Upsilon_{tj}) \propto &\Big[N(\Omega_{tj}|\mu_{\Omega_{tj}},\sigma^2_{\Omega_{tj}})\Big]^{\Upsilon_{tj}}\left[\delta_0(\Omega_{tj})\right]^{1-\Upsilon_{tj}} \\ &\Big[\exp\Big(\frac{1}{2}\log\sigma^2_{\Omega_{tj}} + (\log\kappa_j)^{(1)} - \frac{1}{2}\mathbb{E}_q(\log\Delta_{tj}|\Upsilon_{tj}) + \frac{1}{2}\mu^2_{\Omega,tj}\sigma^{-2}_{\Omega,tj} + a_\Delta\log(b_{\Delta_t}) + \\ &- \log(\Gamma(a_{\Delta_t})) - (a_{\Delta_t}+1)\mathbb{E}_q(\log\Delta_{tj}|\Upsilon_{tj}) - b_\Delta\mathbb{E}_q[\Delta_{tj}^{-1}|\Upsilon_{tj}]\Big)\Big]^{\Upsilon_{tj}}\times \\ &\Big[(1-\kappa_j)^{(1)} + \delta_0(\Delta_{tj})\Big]^{1-\Upsilon_{tj}} \end{aligned} \tag{7.7.123}$$

Where the mean and variance for Ω_{tj} is

$$\sigma^2_{\Omega,tj} = \left[\|Z_j\|^2\Big((\sigma_t^{-2})^{(1)} + \sum_{k>t}(\rho_{kt})^{(2)}(\sigma_k^{-2})^{(1)}\Big) + \mathbb{E}_q[\Delta_{tj}^{-1}|\Upsilon_{tj}]\right]^{-1} \tag{7.7.124}$$

$$\begin{aligned} \mu_{\Omega,tj} =&\sigma^2_{\Omega,tj}Z_j^T\left[(\sigma_t^{-2})^{(1)}\left((\mathrm{v}_{t-j})^{(1)} - \sum_{k<t}(\mathrm{v}_k)^{(1)}(\rho_{tk})^{(1)}\right) + \right. \\ &\left. + \sum_{k>t}(\sigma_k^{-2})^{(1)}(\rho_{kt})^{(2)}(\mathrm{v}_{t-j})^{(1)} - \sum_{k>t}(\sigma_k^{-2})^{(1)}\rho_{kt}^{(1)}\left((\mathrm{v}_k)^{(1)} - \sum_{h<k\ \ h\neq t}(\mathrm{v}_h)^{(1)}(\rho_{kh})^{(1)}\right)\right] \end{aligned} \tag{7.7.125}$$

with

$$(\mathrm{v}_t)^{(1)} = \boldsymbol{y}_t - \alpha_t \mathbf{1}_n - \sum_s X_s(\beta_{ts})^{(1)} - \sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_{tg})^{(1)} - \sum_j Z_j(\Omega_{tj})^{(1)} \tag{7.7.126}$$

$$(\mathrm{v}_{t-j})^{(1)} = \boldsymbol{y}_t - \alpha_t \mathbf{1}_n - \sum_s X_s(\beta_{ts})^{(1)} - \sum_g \boldsymbol{W}_g(\boldsymbol{\zeta}_{tg})^{(1)} - \sum_{l \neq j} Z_j(\Omega_{tl})^{(1)} \tag{7.7.127}$$

The form of the update in (7.7.123) enables us to determine a value for the conditional expectation of Δ_{tj}. In Equation (7.7.123) we have under q where we condition on the value of Υ_{tj}

$$q(\Omega_{tj}|\Upsilon_{tj} = 1, y) = \mathcal{N}(\mu_{\Omega,tj}, \sigma^2_{\Omega,tj}), \quad q(\Omega_{tj}|\Upsilon_{tj} = 0, y) = \delta_0(\Omega_{tj}) \tag{7.7.128}$$

which gives us the update

$$\sigma^2_{\Omega,tj} = \left[\|Z_j\|^2 \Big((\sigma_t^{-2})^{(1)} + \sum_{k>t} (\rho_{kt})^{(2)} (\sigma_k^{-2})^{(1)} \Big) + \mathbb{E}_q[\Delta_{tj}^{-1}|\Upsilon_{tj} = 1] \right]^{-1} \tag{7.7.129}$$

The terms in the $q(\Upsilon_{tj})$, using $\Delta_{tj} = 0$ when $\Upsilon_{tj} = 0$, are proportional to

$$\begin{aligned} p(\Upsilon_{tj} = 1) \propto & \exp\Big(\frac{1}{2} \log \sigma^2_{\Omega_{tj}} + (\log \kappa_j)^{(1)} + \frac{1}{2} \mu^2_{\Omega,tj} \sigma^{-2}_{\Omega,tj} + a_\Delta \log(b_{\Delta_t}) + \\ & - \log(\Gamma(a_{\Delta_t})) - (a_{\Delta_t} + 3/2) \mathbb{E}_q(\log \Delta_{tj}|\Upsilon_{tj} = 1) - b_\Delta \mathbb{E}_q[\Delta_{tj}^{-1}|\Upsilon_{tj} = 1] \Big) \\ p(\Upsilon_{tj} = 0) \propto & (\log(1 - \kappa_j))^{(1)} \end{aligned}$$

Which after normalisation is

$$\begin{aligned} (\Upsilon_{tj})^{(1)} = \Bigg[1 + \exp\Bigg\{ & \frac{1}{2} \log\big(\sigma^{-2}_{\Omega,ts}\big) + (\log(1 - \kappa_j))^{(1)} - (\log \kappa_j)^{(1)} + \frac{1}{2} \mathbb{E}_q(\log \Delta_{tj}|\Upsilon_{tj} = 1) + \\ & - \frac{1}{2} \mu^2_{\Omega,tj} \sigma^{-2}_{\Omega,tj} - a_{\Delta_t} \log(b_{\Delta_t}) + \log(\Gamma(a_{\Delta_t})) + (a_{\Delta_t} + 1) \mathbb{E}_q(\log \Delta_{tj}|\Upsilon_{tj} = 1) + \\ & + b_{\Delta_t} \mathbb{E}_q[\Delta_{tj}^{-1}|\Upsilon_{tj} = 1] \Bigg\} \Bigg]^{-1} \end{aligned} \tag{7.7.130}$$

The approximating q density for Δ_{tj}, which is proportional to Δ_{tj} but conditional on Υ_{tj} is

$$\begin{aligned}
\log q(\Delta_{tj}|\Upsilon_{tj}) &\propto \mathbb{E}_{q(-\Delta_{tj},-\Upsilon_{tj})}\Big[\log p(\Omega_{tj}|\Upsilon_{tj},\Delta_{tj}) + \log p(\Delta_{tj}|\Upsilon_{tj})\Big] \\
&\propto \mathbb{E}_{q(-\Delta_{tj},-\Upsilon_{tj})}\Big[\frac{1}{2}\log\Delta_{tj}^{-1}\Upsilon_{tj} - \frac{1}{2}\Omega_{tj}^2\Upsilon_{tj}\Delta_{tj}^{-1} + \Upsilon_{tj}(a_{\Delta_t}+1)\log\Delta_{tj}^{-1} - b_{\Delta_t}\Upsilon_{tj}\Delta_{tj}^{-1} + \\
&\quad + (1-\Upsilon_{tj})\delta_0(\Delta_{tj})\Big] \\
&\propto \left[(\log\Delta_{tj}^{-1})\Upsilon_{tj}\left(\frac{1}{2}+a_{\Delta_t}+1\right) - \Delta_{tj}^{-1}\Upsilon_{tj}\left(\frac{1}{2}\Omega_{tj}^2 + b_{\Delta_t}\right)\right]\left[(1-\Upsilon_{tj})\delta_0(\Delta_{tj})\right]
\end{aligned}$$

which gives us

$$q(\Delta_{tj}|\Upsilon_{tj}) \sim \Big[IG(\Delta_{tj}|a^*_{\Delta_{tj}}, b^*_{\Delta_{tj}})\Big]^{\Upsilon_{tj}}\Big[\delta_0(\Delta_j)\Big]^{(1-\Upsilon_j)} \tag{7.7.131}$$

Under q

$$q(\Delta_{tj}|\Upsilon_{tj}=1,\boldsymbol{Y}) \sim IG(\Delta_{tj}|a^*_{\Delta_{tj}}, b^*_{\Delta_{tj}}), \quad q(\Delta_{tj}|\Upsilon_{tj}=0,\boldsymbol{Y}), \sim \delta_0(\Delta_{tj})$$

with updates

$$a^*_{\Delta,tj} = \frac{1}{2} + a_{\Delta_t} \tag{7.7.132}$$

$$\begin{aligned}
b^*_{\Delta,tj} &= \frac{1}{2}\mathbb{E}[\Omega^2_{tj}|\Upsilon_j=1] + b_\Delta \\
&= \frac{1}{2}(\sigma^2_{\Omega,tj} + \mu^2_{\Omega,tj}) + b_{\Delta_t}
\end{aligned} \tag{7.7.133}$$

This gives

$$\mathbb{E}_q(\Delta_{tj}^{-1}|\Upsilon_{tj}=1) = a^*_{\Delta,tj}/b^*_{\Delta,tj} \tag{7.7.134}$$

$$\mathbb{E}_q(\log\Delta_{tj}|\Upsilon_{tj}) = \log\Big(b^*_{\Delta_{tj}}\Big) - \Psi(a^*_{\Delta_{tj}})$$

7.7.3 ELBO calculation

The objective of VI is to find the candidate from a family of densities $\mathcal{D}$ which best approximates, the one closest in KL divergence, to the exact conditional

$$q^*(\boldsymbol{\vartheta}) = \underset{q^*(\boldsymbol{\vartheta})\in\mathcal{D}}{\arg\min} \mathrm{KL}(q(\boldsymbol{\vartheta})||p(\boldsymbol{\vartheta}|\boldsymbol{Y})).$$

This objective is not computable as it requires computing marginal likelihood. If we expand the expression

$$\mathrm{KL}(q(\boldsymbol{\vartheta})||p(\boldsymbol{\vartheta}|\boldsymbol{Y})) = \mathbb{E}_{q(\boldsymbol{\vartheta})}[\log q(\boldsymbol{\vartheta})] - \mathbb{E}_{q(\boldsymbol{\vartheta})}[\log p(\boldsymbol{\vartheta}, \boldsymbol{Y})] + \log p(\boldsymbol{Y})$$

we can identify the elements which are a function of the parameters in the model. As the KL cannot be computed, an alternative objective that is equivalent to the KL up to an added constant is the evidence lower bound (ELBO).

$$\mathcal{L}(q) = \mathbb{E}_{q(\boldsymbol{\vartheta})}[\log p(\boldsymbol{\vartheta}, \boldsymbol{Y})] - \log q(\boldsymbol{\vartheta}) \tag{7.7.135}$$

This function is the negative KL divergence plus the marginal likelihood, and is optimised at each iteration of the CAVI in order to monitor its convergence. The computational details are:

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_{q(\boldsymbol{\vartheta})}[\log p(y, \boldsymbol{\vartheta})] - \mathbb{E}_{q(\boldsymbol{\vartheta})}[\log q(\boldsymbol{\vartheta})] \\
&= \sum_t A(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \boldsymbol{\zeta}_t, \boldsymbol{\theta}_t, \sigma_t^2, \rho_t) + \sum_t B^*(\alpha_t|w_{\alpha_t}) + \sum_t \sum_s B(\beta_{ts}, \gamma_{ts}|w_t, \omega_s) + \\
&\quad + \sum_t \tilde{B}(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t|\kappa) + \sum_t \sum_g \hat{B}(\boldsymbol{\zeta}_{tg}, \chi_{tg}|v_t, \varrho_g) + \sum_s C(\omega_s) + \sum_j \tilde{C}(\kappa_j) + \sum_g \hat{C}(\varrho_g) \\
&\quad + \sum_t D(w_t) + \sum_t D^*(w_{\alpha_t}) + \sum_t \hat{D}(v_t) + \sum_t F(\sigma_t^2|\tau, \nu) + \sum_t \sum_{k<t} G(\rho_{tk}, \eta_{tk}|\sigma_t^2, \tau, \lambda) + \\
&\quad + H(\tau) + I(b_w) + I^*(b_\alpha) + \hat{I}(b_v) + J(\lambda).
\end{aligned}$$

The functions are

$$
\begin{aligned}
A(\boldsymbol{y}_t|\boldsymbol{\beta}_t,\boldsymbol{\zeta}_t,\boldsymbol{\theta}_t,\sigma_t^2,\rho_t) =& \mathbb{E}_q\left[-\frac{n}{2}\log(2\pi)+\frac{n}{2}\log\big(\sigma_t^{-2}\big)-\frac{1}{2\sigma_t^2}\left\|\boldsymbol{u}_t-\sum_{k<t}\boldsymbol{u}_k\rho_{tk}\right\|^2\right] \\
=& -\frac{n}{2}\log(2\pi)+\frac{n}{2}\log\big(\sigma_t^2\big)^{(1)}-(\sigma_t^{-2})^{(1)}\left(b^*_{\sigma^2,t}-\frac{(\tau)^{(1)}}{2}-\frac{(\tau)^{(1)}}{2}\sum_{k<t}(\rho_{tk})^{(2)}\right)
\end{aligned}
$$

$$
\begin{aligned}
B^*(\alpha_t|w_{\alpha_t}) =& \mathbb{E}_q[\log p(\alpha_t|w_{\alpha_t})]-\mathbb{E}_q[\log q(\alpha_t)] \\
=& -\frac{1}{2}\log(2\pi)+\frac{1}{2}(\log w_{\alpha_t}^{-1})^{(1)}-\frac{1}{2(w_{\alpha_t})^{(1)}}(\alpha_t)^{(2)}- \\
& \left(-\frac{1}{2}\log(2\pi)-\frac{1}{2}(\log\sigma^2_{\alpha_t})-\frac{1}{2(\sigma^2_{\alpha_t})}\mathbb{E}_q\left[(\alpha_t-\mu_{\alpha_t})^2\right]\right) \\
=& \frac{1}{2}\log\big(\sigma^2_{\alpha_t}\big)+\frac{1}{2}(\log w_{\alpha_t}^{-1})^{(1)}+\frac{1}{2}-\frac{1}{2}(w_{\alpha_t}^{-1})^{(1)}(\alpha_t)^{(2)}
\end{aligned}
\tag{7.7.136}
$$

$$
\begin{aligned}
B(\beta_{ts},\gamma_{ts}|w_t,\omega_s) =& \mathbb{E}_q[\log p(\beta_{ts},\gamma_{ts})]-\mathbb{E}_q[\log q(\beta_{ts},\gamma_{ts})] \\
=& \frac{(\gamma_{ts})^{(1)}}{2}\Big((\log w_t^{-1})^{(1)}+2(\log\omega_s)^{(1)}+1+\log\sigma^2_{\beta,ts}+1-2\log(\gamma_{ts})^{(1)}\Big)+ \\
& -\frac{(\gamma_{ts})^{(1)}}{2}\Big((\sigma^2_{\beta,ts}+\mu^2_{\beta,ts})(w_t)^{(-1)}\Big)+ \\
& +(1-(\gamma_{ts})^{(1)})\Big(\log(1-\omega_s)^{(1)}-\log\big(1-(\gamma_{ts})^{(1)}\big)\Big)
\end{aligned}
\tag{7.7.137}
$$

$$
\begin{aligned}
\hat{B}(\boldsymbol{\zeta}_{tg}, \chi_{tg}|v_t, \varrho_g) =& \mathbb{E}_q[\log p(\boldsymbol{\zeta}_{tg}|\chi_{tg}, v_t)] + \mathbb{E}_q[\log p(\chi_{tg}|\varrho_g)] - \mathbb{E}_q[\log q(\boldsymbol{\zeta}_{tg}, \chi_{tg})] \\
=& (\chi_{tg})^{(1)}\bigg(- \frac{m_g}{2}\log(2\pi) + \frac{m_g}{2}(\log v_t^{-1})^{(1)}\bigg) - \mathbb{E}_q\bigg[\frac{1}{2v_t}\chi_{gt}\boldsymbol{\zeta}_{tg}^T\boldsymbol{\zeta}_{tg}\bigg] + \\
&+ (1 - (\chi_{tg})^{(1)})\delta_0(\boldsymbol{\zeta}_{tg}) + (\chi_{tg})^{(1)}(\log \varrho_g)^{(1)} + (1 - (\chi_{tg})^{(1)})(\log(1 - \varrho_g))^{(1)} + \\
&+ \frac{1}{2}(\chi_{tg})^{(1)}\bigg(m_g\log(2\pi) + \log\det(\Sigma_{\boldsymbol{\zeta}_{tg}})\bigg) + \\
&+ \mathbb{E}_q\bigg(\frac{1}{2}\chi_{tg}(\boldsymbol{\zeta}_{tg} - \boldsymbol{\mu}_{\zeta_{tg}})^T\Sigma_{\boldsymbol{\zeta}_{tg}}^{-1}(\boldsymbol{\zeta}_{tg} - \boldsymbol{\mu}_{\zeta_{tg}})\bigg) - (1 - (\chi_{tg})^{(1)})\delta_0(\boldsymbol{\zeta}_{tg}) + \\
&- (\chi_{tg})^{(1)}\log(\chi_{tg})^{(1)} - (1 - (\chi_{tg})^{(1)})\log\big(1 - (\chi_{tg})^{(1)}\big)
\end{aligned} \tag{7.7.138}
$$

Simplifying using $\mathbb{E}_q\bigg[\chi_{tg}\left(\boldsymbol{\zeta}_{tg}^T\Sigma_{\boldsymbol{\zeta}_{tg}}^{-1}\boldsymbol{\zeta}_{tg}\right)\bigg] = m_g(\chi_{tg})^{(1)}$

$$
\begin{aligned}
\hat{B}(\boldsymbol{\zeta}_{tg}, \chi_{tg}|v, \varrho_g) =& \frac{(\chi_{tg})^{(1)}}{2}\bigg(m_g(\log v_t^{-1})^{(1)} - \frac{1}{(v_t)^{(1)}}(tr(\Sigma_{\boldsymbol{\zeta}_{tg}}) + \boldsymbol{\mu}_{\zeta_{tg}}^T\boldsymbol{\mu}_{\zeta_{tg}}) + \log\det(\Sigma_{\boldsymbol{\zeta}_{tg}}) + m_g + \\
&+ 2(\log \varrho_g)^{(1)} - 2\log\big((\chi_{tg})^{(1)}\big)\bigg) + \\
&+ (1 - (\chi_{tg})^{(1)})\bigg(\log\big(1 - (\chi_{tg})^{(1)}\big) + (\log(1 - \varrho_g))^{(1)}\bigg)
\end{aligned}
$$

$$
\begin{aligned}
\tilde{B}(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t, \boldsymbol{\psi}_t|\mathbf{T}_\xi, \kappa, a_\psi, b_\psi) =& \mathbb{E}_{q(\boldsymbol{\vartheta})}\Big[\log p(\boldsymbol{\theta}_t|\boldsymbol{\psi}_t, \boldsymbol{\xi}_t) + \log p(\boldsymbol{\psi}_t|\boldsymbol{\xi}_t) + \log p(\boldsymbol{\xi}_t)\Big] + \\
&- \mathbb{E}_{\log q(\boldsymbol{\vartheta})}\Big[\log q(\boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \boldsymbol{\xi}_t)\Big]
\end{aligned} \tag{7.7.139}
$$

The approximating density is only known up to a constant of proportionality but this is sufficient

for the ELBO calculations.

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{\vartheta})}\Big[\log(p(\boldsymbol{\theta}_t,\boldsymbol{\xi}_t,\boldsymbol{\psi}_t))\Big] =& -\frac{1}{2}((d_{\boldsymbol{\xi}_t})^{(1)}-1)\log(2\pi)-\frac{1}{2}(\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t}))^{(1)}+\\ &+\sum_j(\xi_{tj})^{(1)}(\log\kappa_j)^{(1)}-\frac{1}{2}(\boldsymbol{\theta}_{\xi_t}^T(\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})^+\boldsymbol{\theta}_{\xi_t})^{(1)}+\\ &+\sum_j(1-(\xi_{tj})^{(1)})(\log\kappa_j)^{(1)}-b_{\boldsymbol{\psi}_t}\sum_j(\xi_{tj}\psi_{tj}^{-1})^{(1)}+\\ &-\sum_j(a_{\boldsymbol{\psi}_t}+1)(\xi_{tj}\log(\psi_{tj}))^{(1)}+(a_{\boldsymbol{\psi}_t}\log(b_{\boldsymbol{\psi}_t})-\log(\Gamma(a_\psi))\sum_j(\xi_{tj})^{(1)}\end{aligned}\tag{7.7.140}$$

The q expectations $(\xi_{tj}\log(\psi_{tj}))^{(1)}$ and $(\xi_{tj}\psi_{tj}^{-1})^{(1)}$ can be found using the law of iterative expectations but these will cancel. The free parameters are a function of $\boldsymbol{\xi}_t$ so when we take an expectation we have

$$\begin{aligned}\mathbb{E}_{q(\vartheta)}\Big[\log q(\boldsymbol{\theta}_t,\boldsymbol{\xi}_t,\boldsymbol{\psi}_t|\boldsymbol{y}_t)\Big] \propto& \mathbb{E}_{q(\boldsymbol{\vartheta})}\Big[\log(\text{SMVN}(\boldsymbol{\theta}_{\xi_t}))\Big]+\frac{1}{2}(\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}}^T\mathbf{T}_{\boldsymbol{\xi}_t}(\mathbf{T}_{\boldsymbol{\xi}_t}^T\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\boldsymbol{\xi}_t})^+\mathbf{T}_{\boldsymbol{\xi}_t}\boldsymbol{\mu}_{\boldsymbol{\theta}_{\xi_t}})^{(1)}+\\ &+\frac{1}{2}(\log\big(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t}\Sigma_{\boldsymbol{\theta}_{\xi_t}}\mathbf{T}_{\boldsymbol{\xi}_t})\big))^{(1)}-\frac{1}{2}(\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t}D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})))^{(1)}+\\ &+\sum_j\xi_{tj}(\log\kappa_j)^{(1)}+\sum_j(1-\xi_{tj})(\log(1-\kappa_j))^{(1)}+\\ &-\sum_j(a_{\boldsymbol{\psi}_t}+1)(\xi_{tj}\log(\psi_{tj}))^{(1)}-b_{\boldsymbol{\psi}_t}\sum_j(\xi_{tj}\psi_{tj}^{-1})^{(1)}+\\ &+(a_{\boldsymbol{\psi}_t}\log(b_{\boldsymbol{\psi}_t})-\log(\Gamma(a_{\boldsymbol{\psi}_t}))\sum_j(\xi_{tj})^{(1)}\end{aligned}\tag{7.7.141}$$

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{\vartheta})}\Big[\log(\text{SMVN}(\boldsymbol{\theta}_{\xi_t}))\Big] =& -\frac{1}{2}((d_{\boldsymbol{\xi}_t})^{(1)}-1)\log(2\pi)-\frac{1}{2}(\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t}\Sigma_{\boldsymbol{\xi}_t}\mathbf{T}_{\boldsymbol{\xi}_t})))^{(1)}+\\ &-\frac{1}{2}\Big\{(\boldsymbol{\theta}_{\xi_t}^T(\mathbf{T}_{\boldsymbol{\xi}_t}\Sigma_{\boldsymbol{\xi}_t}\mathbf{T}_{\boldsymbol{\xi}_t})^+\boldsymbol{\theta}_{\xi_t})^{(1)}-2(\boldsymbol{\theta}_{\xi_t}^T(\mathbf{T}_{\boldsymbol{\xi}_t}\Sigma_{\boldsymbol{\xi}_t}\mathbf{T}_{\boldsymbol{\xi}_t})^+\mathbf{T}_{\boldsymbol{\xi}_t}\mu_{\theta_{\boldsymbol{\xi}_t}})^{(1)}+\\ &+(\mu_{\theta_{\boldsymbol{\xi}_t}}^T\mathbf{T}_{\boldsymbol{\xi}_t}(\mathbf{T}_{\boldsymbol{\xi}_t}\Sigma_{\boldsymbol{\xi}_t}\mathbf{T}_{\boldsymbol{\xi}_t})^+\mathbf{T}_{\boldsymbol{\xi}_t}\mu_{\theta_{\boldsymbol{\xi}_t}})^{(1)}\Big\}\end{aligned}$$

Bringing together the expression for $\tilde{B}$

$$\begin{aligned}\tilde{B}(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t, \boldsymbol{\psi}_t|.) =& \mathbb{E}_{q(\mathbf{z})}\Big[\log p(\boldsymbol{\theta}_t|\boldsymbol{\xi}_t, \boldsymbol{\psi}_t) + \log p(\boldsymbol{\psi}_t|\boldsymbol{\xi}_t, a_{\boldsymbol{\psi}_t}, b_{\boldsymbol{\psi}_t}) + \log p(\boldsymbol{\xi}_t|\kappa)\Big] - \mathbb{E}_{q(\mathbf{z})}\Big[\log q(\boldsymbol{\theta}_t, \boldsymbol{\xi}_t)\Big] \\ =& -\frac{1}{2}(\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t} D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})))^{(1)} + \frac{1}{2}(\log(\det{}^*(\mathbf{T}_{\boldsymbol{\xi}_t}\Sigma_{\boldsymbol{\xi}_t}\mathbf{T}_{\boldsymbol{\xi}_t})))^{(1)} + \\ & -\frac{1}{2}\Big\{(\boldsymbol{\theta}_{\xi_t}^T(\mathbf{T}_{\boldsymbol{\xi}_t} D(\boldsymbol{\psi}_{\xi_t})\mathbf{T}_{\boldsymbol{\xi}_t})^+\boldsymbol{\theta}_{\xi_t})^{(1)} - (\boldsymbol{\theta}_{\xi_t}^T(\mathbf{T}_{\boldsymbol{\xi}_t}\Sigma_{\boldsymbol{\xi}_t}\mathbf{T}_{\boldsymbol{\xi}_t})^+\boldsymbol{\theta}_{\xi_t})^{(1)}\Big\} + \\ & + (\boldsymbol{\theta}_{\xi_t}^T(\mathbf{T}_{\boldsymbol{\xi}_t}\Sigma_{\boldsymbol{\xi}_t}\mathbf{T}_{\boldsymbol{\xi}_t})^+\mathbf{T}_{\boldsymbol{\xi}_t}\mu_{\theta_{\boldsymbol{\xi}_t}})^{(1)}\end{aligned}$$

$$\begin{aligned}C(\omega_s) =& \mathbb{E}_q[\log p(\omega_s)] - \mathbb{E}_q[\log q(\omega_s)] \\ =& \log B(a^*_{\omega,s}, b^*_{\omega,s}) - \log B(a_{\omega,s}, b_{\omega,s}) + \\ & + (a^*_{\omega,s} - a_{\omega,s})(\log \omega_s)^{(1)} + (b^*_{\omega,s} - b_{\omega,s})(\log[1-\omega_s])^{(1)}\end{aligned} \tag{7.7.142}$$

$$\begin{aligned}\tilde{C}(\kappa_j) =& \mathbb{E}_q[\log p(\kappa_j)] - \mathbb{E}_q[\log q(\kappa_j)] \\ =& \log B(a^*_{\kappa,j}, b^*_{\kappa,j}) - \log B(a_j, b_j) + \\ & + (a^*_{\kappa,j} - a_j)(\log \kappa_j)^{(1)} + (b^*_j - b_j)(\log[1-\kappa_j])^{(1)}\end{aligned} \tag{7.7.143}$$

$$\begin{aligned}\hat{C}(\varrho_g) =& \mathbb{E}_q[\log p(\varrho_g)] - \mathbb{E}_q[\log q(\varrho_g)] \\ =& \log B(a^*_{\varrho,g}, b^*_{\varrho,g}) - \log B(a_\varrho, b_\varrho) + \\ & + (a^*_{\varrho,g} - a_\varrho)(\log \varrho_g)^{(1)} + (b^*_{\varrho,g} - b_\varrho)(\log[1-\varrho_g])^{(1)}\end{aligned} \tag{7.7.144}$$

$$
\begin{aligned}
D(w_t) =& \mathbb{E}_q[\log p(w_t)] - \mathbb{E}_q[\log q(w_t)] \\
&= a_w(\log b_w)^{(1)} - a_w^* \log b_w^* - \log\Gamma(a_w) + \log\Gamma(a_{w,t}^*) + \\
&+ (a_w - a_{w,t}^*)(\log w_t^{-1})^{(1)} + (b_{w,t}^* - (b_w)^{(1)})(w_t^{-1})^{(1)}
\end{aligned} \tag{7.7.145}
$$

$$
\begin{aligned}
D^*(w_{\alpha_t}) =& \mathbb{E}_q[\log p(w_{\alpha_t})] - \mathbb{E}_q[\log q(w_{\alpha_t})] \\
=& \mathbb{E}_q\left[a_\alpha \log b_\alpha - \log\Gamma(a_\alpha) + (a_\alpha + 1)\log w_{\alpha_t}^{-1} - b_\alpha w_{\alpha_t}^{-1}\right] + \\
&- \mathbb{E}_q\left[a_{\alpha_t}^* \log b_{\alpha_t}^* - \log\Gamma(a_{\alpha_t}^*) - (a_{\alpha_t}^* + 1)\log w_{\alpha_t}^{-1} + b_{\alpha_t}^* w_{\alpha_t}^{-1}\right] \\
=& a_\alpha(\log b_\alpha)^{(1)} - a_{\alpha_t}^* \log b_{\alpha_t}^* - \log\Gamma(a_{\alpha_t}) + \log\Gamma(a_{\alpha_t}^*) + \\
&+ (a_\alpha - a_{\alpha_t}^*)(\log w_{\alpha_t}^{-1})^{(1)} + (b_{\alpha_t}^* - (b_\alpha)^{(1)})(w_{\alpha_t}^{-1})^{(1)}
\end{aligned} \tag{7.7.146}
$$

$$
\begin{aligned}
\hat{D}(v_t) =& \mathbb{E}_q[\log p(v)] - \mathbb{E}_q[\log q(v)] \\
=& \mathbb{E}_q\left[a_v \log b_v - \log\Gamma(a_v) + (a_v + 1)\log v_t^{-1} - b_v v_t^{-1}\right] + \\
&- \mathbb{E}_q\left[a_{v,t}^* \log b_{v,t}^* - \log\Gamma(a_{v,t}^*) - (a_{v,t}^* + 1)\log v_t^{-1} + b_{v,t}^* v_t^{-1}\right] \\
=& a_v(\log b_v)^{(1)} - a_{v,t}^* \log b_{v,t}^* - \log\Gamma(a_v) + \log\Gamma(a_{v,t}^*) + \\
&+ (a_v - a_{v,t}^*)(\log v_t^{-1})^{(1)} + (b_{v,t}^* - (b_v)^{(1)})(v_t^{-1})^{(1)}
\end{aligned} \tag{7.7.147}
$$

$$
\begin{aligned}
F(\sigma_t^2|\tau,\nu) =& \mathbb{E}_q[\log p(\sigma_t^2|\tau,\nu)] - \mathbb{E}_q[\log q(\sigma_t^2)] \\
=& \Big(\frac{\nu-T+t}{2}\Big)(\log\tau)^{(1)} - \Big(\frac{\nu-T+t}{2}\Big)\log 2 - \log\Gamma\Big(\frac{\nu-T+t}{2}\Big) + \\
&+ \Big(\frac{\nu-T+1}{2}+1\Big)(\log\sigma_t^{-2})^{(1)} - \frac{\tau^{(1)}}{2}(\sigma_t^{-2})^{(1)} + \\
&- \Big(a^*_{\sigma^2,t}\log b^*_{\sigma^2,t} - \log\Gamma(a^*_{\sigma^2,t}) + (a^*_{\sigma^2,t}+1)(\log\sigma_t^{-2})^{(1)} - b^*_{\sigma^2,t}(\sigma_t^{-2})^{(1)}\Big) \\
=& \Big(\frac{\nu-T+t}{2}\Big)\Big(\log\tau^{(1)} - \log 2\Big) - a^*_{\sigma^2,t}\log b^*_{\sigma^2,t} - \log\Gamma\Big(\frac{\nu-T+t}{2}\Big) + \\
&+ \log\Gamma(a^*_{\sigma^2,t}) + (\log\sigma_t^{-2})^{(1)}\Big(\Big(\frac{\nu-T+t}{2}\Big) - a^*_{\sigma^2,t}\Big) + (\sigma_t^{-2})^{(1)}\Big(b^*_{\sigma^2,t} - \frac{\tau^{(1)}}{2}\Big)
\end{aligned}
\tag{7.7.148}
$$

$$
\begin{aligned}
G(\rho_{tk},\eta_{tk}|\sigma_t^2,\tau,\lambda) =& \mathbb{E}_q[\log p(\rho_{tk},\eta_{tk})] - \mathbb{E}_q[\log q(\rho_{tk},\eta_{tk})] \\
=& \frac{\eta_{tk}^{(1)}}{2}\Bigg((\log\tau)^{(1)} + (\log\sigma_t^{-2})^{(1)} + 2(\log\lambda)^{(1)} + 1 + \log\sigma^2_{\rho_{tk}} + \\
&- 2\log\Big((\eta_{tk})^{(1)}\Big)\Bigg) - \frac{(\rho_{tk})^{(2)}(\tau)^{(1)}(\sigma_t^{-2})^{(1)}}{2} + \\
&+ (1-\eta_{tk})^{(1)}\Bigg((\log(1-\lambda))^{(1)} - \log\Big(1-(\eta_{tk})^{(1)}\Big)\Bigg)
\end{aligned}
\tag{7.7.149}
$$

$$
\begin{aligned}
H(\tau) =& \mathbb{E}_q[\log p(\tau)] - \mathbb{E}_q[\log q(\tau)] \\
=& a_\tau\log b_\tau - a^*_\tau\log b^*_\tau + \log\Gamma(a^*_\tau) - \log\Gamma(a_\tau) + \\
&+ (a^*_\tau - a_\tau)(\log\tau)^{(1)} + (b^*_\tau - b_\tau)(\tau)^{(1)}.
\end{aligned}
\tag{7.7.150}
$$

$$
\begin{aligned}
I(b_w) =& \mathbb{E}_q[\log p(b_w)] - \mathbb{E}_q[\log q(b_w)] \\
=& \mathbb{E}_q\Big[a_b \log b_b - \log\Gamma(a_b) + (a_b - 1)\log b_w - b_b b_w\Big] + \\
& - \mathbb{E}_q\Big[a_b^* \log b_b^* - \log\Gamma(a_b^*) + (a_b^* - 1)\log b_w - b_b^* b_w\Big] \\
=& a_b \log b_b - a_b^* \log b_b^* - \log\Gamma(a_b) + \log\Gamma(a_b^*) + (\log b_w)^{(1)}(a_b - a_b^*) + \\
& + (b_w)^{(1)}(b_b^* - b_b)
\end{aligned} \tag{7.7.151}
$$

$$
\begin{aligned}
I^*(b_\alpha) =& \mathbb{E}_q[\log p(b_\alpha)] - \mathbb{E}_q[\log q(b_\alpha)] \\
=& \mathbb{E}_q\Big[a_{b,\alpha} \log b_{b,\alpha} - \log\Gamma(a_{b,\alpha}) + (a_{b,\alpha} - 1)\log b_\alpha - b_\alpha b_{b,\alpha}\Big] + \\
& - \mathbb{E}_q\Big[a_{b,\alpha}^* \log b_{b,\alpha}^* - \log\Gamma(a_{b,\alpha}^*) + (a_{b,\alpha}^* - 1)\log b_\alpha - b_\alpha b_{b,\alpha}^*\Big] \\
=& a_{b,\alpha} \log b_{b,\alpha} - a_{b,\alpha}^* \log b_\alpha^* - \log\Gamma(a_{b,\alpha}) + \log\Gamma(a_{b,\alpha}^*) + (\log b_\alpha)^{(1)}(a_{b,\alpha} - a_{b,\alpha}^*) + \\
& + (b_\alpha)^{(1)}(b_{b,\alpha}^* - b_{b,\alpha})
\end{aligned} \tag{7.7.152}
$$

$$
\begin{aligned}
\hat{I}(b_v) =& \mathbb{E}_q[\log p(b_v)] - \mathbb{E}_q[\log q(b_v)] \\
=& \mathbb{E}_q\Big[a_{bv} \log b_{bv} - \log\Gamma(a_{bv}) + (a_{bv} - 1)\log b_v - b_{bv} b_v\Big] + \\
& - \mathbb{E}_q\Big[a_{bv}^* \log b_{bv}^* - \log\Gamma(a_{bv}^*) + (a_{bv}^* - 1)\log b_v - b_{bv}^* b_v\Big] \\
=& a_{bv} \log b_{bv} - a_{bv}^* \log b_{bv}^* - \log\Gamma(a_{bv}) + \log\Gamma(a_{bv}^*) + (\log b_v)^{(1)}(a_{bv} - a_{bv}^*) + \\
& + (b_v)^{(1)}(b_{bv}^* - b_{bv})
\end{aligned} \tag{7.7.153}
$$

$$J(\lambda) = \mathbb{E}_q[\log p(\lambda)] - \mathbb{E}_q[\log q(\lambda)]$$
$$= (\log \lambda)^{(1)}(a_\lambda - a_\lambda^*) + (\log(1-\lambda))^{(1)}(b_\lambda - b_\lambda^*) - \log B(a_\lambda, b_\lambda) + \log B(a_\lambda^*, b_\lambda^*) \qquad (7.7.154)$$

CHAPTER 8

Bayesian Hierarchical Mixture of Experts for Multi-dimensional Responses via Variational Inference

8.1 Abstract

We are motivated by clusters of people who exhibit different causal pathways to the same multi-dimensional endpoint. These multi-dimensional biological endpoints are related to each other by a latent structure which will often vary across the clusters, preventing the convenient assumption of independent residuals across the regressions. A hierarchical multivariate response Bayesian mixture of experts model is developed, which captures the different latent structures across the clusters to aid model fitting and understanding. A reparameterisation of the seemingly unrelated regression model with hierarchical priors, assist the model to leverage shared information across the responses, increasing the sensitivity of detecting weaker associations. Cluster specific feature selection within the experts exploits sparsity to facilitate both covariate and covariance selection, where the combination of covariates is free to vary across the experts. The unsupervised learning

of detecting new information in the clustering of individuals is determined by a subset of their predictors. The model is estimated by block-mean-field coordinate ascent variational inference so that it scales efficiently with high-dimensional data.

8.2 Introduction

We are motivated by clusters of people who exhibit different causal pathways to the same multi-dimensional endpoint. Our objective is to cluster multiple response linear regressions with high-dimensional data, with the following constraints:

- The responses are related to each other by a latent structure. This structure is free to vary across the clusters.
- Each regression model is specific to each cluster, in that the covariate for one regression may not be present in another.
- A small set of the covariates can discriminate the clusters.
- Sparsity is expected so, one would like to have relatively few predictors in the model.

This particular scenario is present in many real problems, such as gene expression data. Microarray gene expression studies are performed to measure the transcription levels of an organism's genes. A common aim in the analysis of gene expression measurements observed in a population is the identification of naturally occurring sub-populations. For instance in cancer studies, the identification of sub-groups of tumours having distinct mRNA profiles can help discover molecular fingerprints that will define subtypes of disease (Gosh and Smolkin, 2003).

A variant of the mixture of regressions model is considered from a Bayesian perspective, an area that has been explored by Hurn et al. (2003) and Fruhwirth-Schnatter (2006). This type of model can be useful when there is sudden parameter change after a break point (Goldfeld and Quandt, 1973), an omitted categorical predictor (Hosmer, 1974), segments of individuals within

a population (DeSarbo and Cron, 1988) or the presence of outliers within the data set (Box and Tiao, 1968). In the Bayesian paradigm, estimation of the model can be achieved either using Markov chain Monte Carlo (MCMC) algorithm or variational inference (VI). In MCMC methods, both the Gibbs sampler and the Metropolis-Hastings algorithm are often required (Gormley and Murphy, 2010). As in any mixture model setting, the so called label switching problem (Stephens (2000a) and Frühwirth-Schnatter (2011)) must be considered when employing such algorithms. In order to ensure that the MCMC sampler converges, this is involves either a random permutation of the labels (Frühwirth-Schnatter, 2001) or more sophisticated and complex MCMC methods to improve the mix of the sampler (Celeux et al., 2000). This issue is conveniently bypassed in the VI approach, which relies on scaling the slope of the evidence lower bound (ELBO), to reach a local optimum.

Bayesian variable selection approaches for univariate responses can be framed in terms of the prior specification, specifically “shrinkage priors" or “explicit variable selection" priors. Shrinkage priors, such as the Bayesian lasso (Park and Casella, 2008) or horseshoe prior (Carvalho et al., 2010), encourage the majority of regression coefficients to be shrunk to very small values when an estimator is applied. Explicit variable selection priors (George and McCulloch (1997), Kuo and Mallick (1998)), use augmented latent indicator variables, with respect to the covariates, which regression parameters should be included in the model. Variable selection is often performed by sampling from the posterior distributions of the latent indicators and the posterior distributions of the coefficients of the selected variables. Both approaches have been extended into finite mixture models via MCMC sampling, (Cozzini et al. (2014) and Lee et al. (2016)).

In our applications we expect a small subset of variables to discriminate clusters, which can be achieved using either explicit variable selection or shrinkage priors. Yau and Holmes (2011) build the clusters on the full set of parameters and use shrinkage priors to force the parameters for many variables, to be the same across the clusters. Alternatively Papathomas et al. (2012) use binary indicators, which means that the likelihood factors into a part which is a mixture over groups using the subset of discriminating variables and a part which is common across cluster. Chung and Dunson (2009) also use binary indicators to discriminate variables in a probit regression,

which determine the sticking breaking probabilities and thus clustering properties of a Dirichlet Process mixture model. All of these approaches develop models for data which are limited to a single response.

The mixture of experts model encapsulates a class of mixture models in which the model parameter are modelled as functions of concomitant covariates. While the response variable is modelled via a mixture model, model parameters are modelled as functions of other related covariates from the context under study. The framework facilitates flexible modelling and has been used in numerous classification, regression and fusion applications in healthcare, finance, surveillance and recognition. In 2003 Bishop and Svensen (2003) presented a Bayesian HME where they considered binary trees with softmax functions for the gates. This approach has proved popular and has been used to estimate speech quality (Mossavat et al., 2010), map threads in dynamic runtime environments (Emani and O'Boyle, 2015), model material (Morand and Helm, 2019) and neural connectivity (Bock and Fine, 2014), categorise human behavior (Kanaujia and Metaxas, 2006) and recognize phone activity (Lee and Cho, 2014). Our interest concerns extending finite regression models to a multivariate responses, in settings where latent structure(s) induce a high level of correlation across the responses such as 3D-imaging (Hammond and Suttie, 2012), serum metabolic profiles (Kettunen et al., 2012) or gene expression (Ackermann et al., 2013). Capturing the correlation across the responses has been shown to increase statistical power and improve model estimation and data understanding, offering a considerable improvement to the univariate approach (Inouye et al., 2012).

An explicit variable selection framework for multivariate outcomes was developed by Brown et al. (2002). The posterior space is large and complex for this multivariate model with high-dimensional data. Two alternative simplifying assumptions have been made in applications of this model, to exploit conjugacy with respect to regression coefficients and residual covariance. One of these simplifications is used by Petretto et al. (2010), who assume the same set of covariates is selected in the regression equation for every response. This ensures conjugacy in the model and enables feasible computational time in a high dimensional *omics* setting. The same assumption on covariates is made by Bhadra and Mallick (2013), who extend the model by including sparse

residual covariance selection between regressions, using graphical modelling based on decomposable graphs. An alternative approach is to share information with hierarchical priors, but to assume the residuals are conditionally independent. This enables direct simulation of the posterior probability of covariate inclusion (Scott-Boyer et al. (2012) and Ruffieux et al. (2017)).

Recently, approaches have been developed for a fully Bayesian variable selection model to avoid these simplifying assumptions which are often unrealistic, particularly in the *omics* setting. Banterle and Lewin (2018) reparameterise the Seemingly Unrelated Regression (SUR) model and perform covariance selection via the graphical structure of the precision matrix, using an MCMC augmented with junction trees (Green and Thomas, 2013). This searches the space of decomposable graphs to identify zero entries in the reparameterised covariance space and exploits the expected sparsity in the data, allowing the model to be defined by a subset of the coefficients at each iteration, thus reducing the computational cost of the MCMC sampler. The SUR parameterisation of Banterle and Lewin (2018) is exploited by Scott and Lewin (2022) who use the same SUR model and reparameterisation, but avoid graphical models and MCMC approaches, by using latent indicator variables coupled with fast variational inference computation.

We develop a hierarchical multivariate response model in the mixture of experts framework. This involves integrating the work by Scott and Lewin (2022) in big data regression classification, which relates multivariate outcomes ($\boldsymbol{Y}$) with multivariate predictors ($\boldsymbol{X}$) of high dimension where $n >> p$ and p is sparse, with the unsupervised learning of detecting new information in the clustering of individuals based on their predictors $\boldsymbol{X}$. Through a reparameterisation of the SUR model, the responses are free to be correlated through some latent structure which can vary across the mixtures, but each expert comprises a product of conditionally independent linear regressions. Feature selection which exploits the sparsity in the data, is performed on the parameters within the experts and the mixing coefficients via latent indicator variables. This facilitates both covariate and covariance selection in a model in which, both the design matrix and the number of responses, can be of high dimensions. Hierarchical priors allow the model to leverage shared information across the responses, increasing the sensitivity of detecting weaker associations. The model is estimated by variational inference (VI), with the use of a lower bound on the group local variables which

ensures the conjugate exponential structure is retained for the parameters which determine the probability of a particular cluster.

8.3 Methods

8.3.1 HME likelihood reparameterisation

We are motivated by clusters of individuals who may exhibit different causal pathways to the same multivariate endpoint. The hierarchical mixture of experts (HME) is a machine learning approach which incorporates a mixture of linear regressions within a tree like structure, where the mixing coefficients are themselves a function of the design matrix. In our biological context, the responses T comprise a subset of a system and are related by some latent structure $\boldsymbol{C}_j$, which varies across the $j = 1, ..., J$ clusters. Intuitively, we can think of the multivariate response to be grouped in clusters, and the shape of the data within each cluster to be shaped by $\boldsymbol{C}_j$. Allowing $\boldsymbol{C}_j$ to vary across the clusters ensures the model is able to capture these different shapes. For each cluster j, the linked linear model with the T vectors stacked on top of each other form

$$\begin{aligned}\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\\ \vdots\\ \boldsymbol{y}_T\end{pmatrix} &= \begin{pmatrix}\boldsymbol{X} & \mathbf{0} & \dots & \mathbf{0}\\ \mathbf{0} & \boldsymbol{X} & \dots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{X}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_{1j}\\ \boldsymbol{\beta}_{2j}\\ \vdots\\ \boldsymbol{\beta}_{Tj}\end{pmatrix} + \begin{pmatrix}\mathbf{u}_{1j}\\ \mathbf{u}_{2j}\\ \vdots\\ \mathbf{u}_{Tj}\end{pmatrix} = \tilde{\boldsymbol{X}}\boldsymbol{\beta}_j + \mathbf{u}_j\\ \mathbf{u}_j &\sim N_{n\times T}(\mathbf{0}, \boldsymbol{C}_j \otimes \mathbb{I}_n).\end{aligned} \tag{8.3.1}$$

The error terms $\mathbf{u}_{tj}$ from the same regression are assumed to be independent given the model covariates, and the residual variance is free to change across the models. Importantly, correlation between the error terms of different models is captured in $\boldsymbol{C}_j$, allowing the responses to be correlated between themselves.

This parameterisation is problematic as the computational needed to compute the marginal

conditional or approximate posteriors when the likelihood is in the form of (8.3.1) is prohibitively expensive. The covariance matrix $\boldsymbol{C}_j \otimes \mathbb{I}_n$ is not diagonal and the large design matrix $\tilde{\boldsymbol{X}}$ has to be inverted in the form of $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$, with the hyperparameters, for each cluster j. Imposing feature selection on the covariance parameters is non-trivial, because of the positive definite constraint on the matrix.

We take advantage of the factorisation in Scott and Lewin (2022), who exploit the properties of the conditional bivariate normal to express the linear model for a particular cluster as

$$p(\boldsymbol{Y}|\tilde{\boldsymbol{X}}, \boldsymbol{\beta}_j, \boldsymbol{C}_j) = \prod_{t=1}^{T} \psi(\boldsymbol{y}_t|\boldsymbol{X}\boldsymbol{\beta}_{tj} + \boldsymbol{U}_{(t-1)j}\boldsymbol{\rho}_{tj}, \sigma^2_{tj}\mathbb{I}_n). \tag{8.3.2}$$

where the matrix $\boldsymbol{U}_{(t-1)j} = \boldsymbol{Y}_{(t-1)j} - (\boldsymbol{X}\boldsymbol{\beta}_{1j} \ ... \ \boldsymbol{X}\boldsymbol{\beta}_{(t-1)j})$ consists of the first $t-1$ residuals from the linked regression and $\psi(\mathbf{y}|\boldsymbol{\mu}, \Sigma)$ is the probability density function for the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. The new parameters are defined by

$$\left.\begin{aligned} \sigma^2_{1j} \quad &\equiv c_{1j} \\ \sigma^2_{tj} \quad &\equiv c_{tj} - \mathbf{c}^T_{tj}\boldsymbol{C}^{-1}_{(t-1)j}\mathbf{c}_{tj} \\ \boldsymbol{\rho}_{tj} \quad &\equiv \boldsymbol{C}^{-1}_{(t-1)j}\mathbf{c}_{tj}. \end{aligned}\right\} \quad t = 2, ..., T. \tag{8.3.3}$$

where

$$\boldsymbol{C}_{(t)j} = \begin{pmatrix} \boldsymbol{C}_{(t-1)j} & \mathbf{c}_{tj} \\ \mathbf{c}^T_{tj} & c_{tj} \end{pmatrix}. \tag{8.3.4}$$

The ordering of the decomposition does not affect the joint distribution $p(\boldsymbol{Y}|\tilde{\boldsymbol{X}}, \boldsymbol{\beta}_j, \boldsymbol{C}_j)$ as the factoring is by chain-conditioning. The parameter σ^2_{tj} is the residual variance of the response t conditioned on the $\boldsymbol{U}_{(t-1)j}$ residuals, $\boldsymbol{\rho}_{tj}$ is a real valued vector of regression coefficients.

The HME with the reparameterised likelihood (8.3.2), marginalised over the latent cluster variable, defines a mixture distribution over the response vector $\boldsymbol{y}_i$, conditioned on the parameters

and vector of design points for a a vector of observations $\boldsymbol{x}_{i,.}$

$$p(\boldsymbol{y}_i|\boldsymbol{x}_{i,.},\boldsymbol{\vartheta}) = \sum_{j=1}^{J} g_j(\boldsymbol{x}_{i,.},\boldsymbol{\upsilon}_g)\left(\prod_{t=1}^{T} p(y_{it}|\boldsymbol{x}_{i,.},\boldsymbol{\beta}_{tj},\boldsymbol{u}_{tj},\boldsymbol{\rho}_{tj},\sigma^2_{tj})\right) \tag{8.3.5}$$

Each *expert* corresponds to a multivariate Gaussian linear regression of dimension T. The reparameterisation induces independence across each response, so conveniently $p(\boldsymbol{y}_i|.) = \prod_t p(y_{ti}|.)$, as the multivariate Gaussian distribution is now a product of univariate conditionally independent Gaussian distributions, despite the T responses being related to each other through some latent system which is free to vary across the clusters.

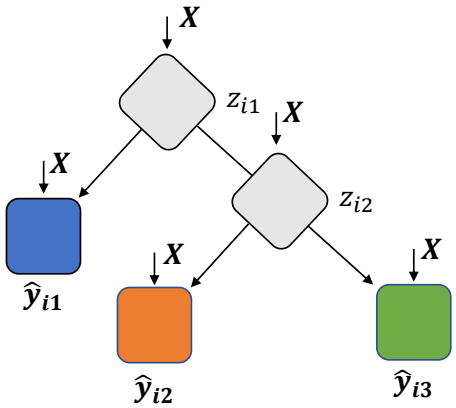

Figure 8.3.1: The gating network of a hierarchical mixture of experts, comprising expert nodes shown as coloured squares, and gating nodes shown as grey diamonds. The binary variables associated with the gating nodes are denoted by z_{ig} and $\hat{\boldsymbol{y}}_{ij}$ is the vector of fitted value from the multivariate response linear regression.

The *experts* are combined in the mixture using weights, called mixing coefficients $g_j(\boldsymbol{x}_{i,.},\boldsymbol{\upsilon}_g)$ which define the probability of an observation belonging to a particular cluster, hence $0 \leq g_j(\boldsymbol{x}_{i,.},\boldsymbol{\upsilon}_g) \leq 1$ and $\sum_{j=1}^{J} g_j(\boldsymbol{x}_{i,.},\boldsymbol{\upsilon}_g) = 1$. Mixing coefficients are conditional on the covariates, and are determined by the gating network: a tree structure with binary classifiers, or gates, at its internal nodes, Figure 8.3.1. Each gating node has an associated binary variable $z_{ig} \in \{0,1\}$, corresponding to the gth gate and the ith data point $\boldsymbol{x}_{i,.}$. A value of 1 for z_{ig} indicates the gth gate left-side branch is chosen for $\boldsymbol{x}_{i,.}$, else $z_{ig} = 0$ indicating the choice of the gth gate right-side branch. To express the likelihood conditional on a particular cluster, rather than the marginal

likelihood in (8.3.5), we define the mixing coefficients as

$$g_j(\boldsymbol{x}_{i,.}, \boldsymbol{v}_g) = p(\zeta_{ij} = 1) \tag{8.3.6}$$

where

$$\zeta_{ij} = \prod_{g=1}^{G} z_{ig}^{S^L(j,g)} (1 - z_{ig})^{S^R(j,g)}. \tag{8.3.7}$$

The latent indicator ζ_{ij} (8.3.7) is thus defined by chain-conditioning clustering probabilities on each other, rather than a direct parameterisation with a Dirichlet prior. The gating network topology is specified by binary matrices $\mathbf{S}^L$ and $\mathbf{S}^R$, where $\mathbf{S}^L(j, g) = 1$ if the jth expert is on the left sub-tree of the gth gate, and zero otherwise. Similarly, $\mathbf{S}^R(j, g) = 1$ if the jth expert is on the right sub-tree of the gth gate, and zero otherwise.

The probability distribution of the binary variable z_{ig} is

$$p(z_{ig}|\boldsymbol{x}_{i,.}, \boldsymbol{v}_g) = \sigma\left(\boldsymbol{v}_g^T \boldsymbol{x}_{i,.}\right)^{z_{ig}} \left[1 - \sigma\left(\boldsymbol{v}_g^T \boldsymbol{x}_{i,.}\right)\right]^{1-z_{ig}} \tag{8.3.8}$$

where $\boldsymbol{x}_{i,.}$ is the vector of design inputs, $\sigma(a) = (1 + \exp(-a))^{-1}$ is the sigmoid function and $\boldsymbol{v}_g$ is the vector of parameters for the gth gate.

A soft (or fuzzy) partitioning of the data is performed in the HME. In the example in Figure 8.3.1 with 3 mixtures $J = 3$, the $\mathbf{S}^L$ and $\mathbf{S}^R$ matrices are

$$\mathbf{S}^L = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{S}^R = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

The latent indicator variable vector $\boldsymbol{\zeta}_i(\boldsymbol{x}_i)$ which determines the likelihood for a particular cluster,

from expanding (8.3.7) for each expert, can be defined as

$$\boldsymbol{\zeta}_i = \begin{bmatrix} z_{i1} \\ z_{i2}(1 - z_{i1}) \\ (1 - z_{i1})(1 - z_{i2}) \end{bmatrix}$$

with a draw from $\boldsymbol{\zeta}_i$ corresponding to a draw from the categorical distribution with probability vector

$$\begin{bmatrix} p(\zeta_{i1} = 1) \\ p(\zeta_{i2} = 1) \\ p(\zeta_{i3} = 1) \end{bmatrix} = \begin{bmatrix} \sigma(\boldsymbol{\upsilon}_1^T \boldsymbol{x}_{i,\cdot}) \\ \sigma(\boldsymbol{\upsilon}_2^T \boldsymbol{x}_{i,\cdot})(1 - \sigma(\boldsymbol{\upsilon}_1^T \boldsymbol{x}_{i,\cdot})) \\ (1 - \sigma(\boldsymbol{\upsilon}_1^T \boldsymbol{x}_{i,\cdot}))(1 - \sigma(\boldsymbol{\upsilon}_2^T \boldsymbol{x}_{i,\cdot})) \end{bmatrix}. \tag{8.3.9}$$

The conditional probability of a multivariate observation from a particular cluster is

$$p(\boldsymbol{y}_i|\cdot) = \left(\prod_t (2\pi)^{-1/2} (\sigma_{tj}^2)^{-1/2} \exp \left\{ -\frac{1}{2\sigma_{tj}^2} \left\| y_{it} - \boldsymbol{x}_{i,\cdot}^T \boldsymbol{\beta}_{tj} - \sum_{k<t} u_{ikj} \rho_{tkj} \right\|^2 \right\} \right)^{\zeta_{ij}}, \tag{8.3.10}$$

where the probability of the latent indicator ζ_{ij} is a function of the design matrix and parameters $\boldsymbol{\upsilon}$ and u_{ikj} is the element in the ith row of the kth column, for the jth cluster in the $\boldsymbol{U}_{(t-1)j}$ matrix.

8.3.2 Priors

We perform explicit variable selection on the covariates for each response within each cluster by positing a "spike-and-slab" prior (George and McCulloch, 1997) on the regression parameters β_{tsj}. The spike is a point mass at 0 (Dirac distribution) with probability $1 - \omega_{sj}$ and the slab is a zero centred Gaussian with variance w_t. An inverse gamma hyperprior is placed on the variance parameter w_t. The binary latent indicator variable γ_{tsj} represents the inclusion of the sth covariate, in the jth cluster, for the tth response. We take advantage of the multiple responses by allowing the sparsity parameter ω_{sj} to vary over the covariate space for each cluster, an option which is

rarely available with a univariate response.

$$p(\beta_{tsj}|\gamma_{tsj}, w_t) = \left[(2\pi)^{-1/2}(w_t)^{-1/2}\exp\left\{-\frac{1}{2w}\|\beta_{tsj}\|^2\right\}\right]^{\gamma_{tsj}} \delta_0(\beta_{tsj})^{1-\gamma_{tsj}}$$
$$p(\gamma_{tsj}|\omega_{sj}) = \omega_{sj}^{\gamma_{tsj}}(1-\omega_{sj})^{1-\gamma_{tsj}}$$

If we posit an inverse Wishart prior on the positive definite matrix $\boldsymbol{C}_j \sim IW(\nu, M_j)$ in the original parameterisation (7.3.5), the priors on the new parameters $\{\sigma^2_{tj}, \boldsymbol{\rho}_{tj}\}$ are suitably defined. As σ^2_{tj} is the Schur complement of c_{tj} in $\boldsymbol{C}_{(t)j}$ and $\boldsymbol{\rho}_{tj} = \boldsymbol{C}^{-1}_{(t-1)j}\boldsymbol{c}_{tj}$, their priors can be calculated using standard matrix properties of the Inverse Wishart (Dawid, 1981). Decomposing $\boldsymbol{M}_j$ conformally with $\boldsymbol{C}_j$ into

$$\boldsymbol{M}_{(t)j} = \begin{pmatrix} \boldsymbol{M}_{(t-1)j} & \mathbf{m}_{tj} \\ \mathbf{m}^T_{tj} & m_{tj} \end{pmatrix}. \tag{8.3.11}$$

for $t = 2, ..., T$, the priors for the new parameters are defined as

$$\sigma^2_{tj} \sim IG\Big(\frac{\nu - T + t}{2}, \frac{m_{tj} - \boldsymbol{m}^T_{tj}\boldsymbol{M}^{-1}_{tj}\boldsymbol{m}_{tj}}{2}\Big) \tag{8.3.12}$$
$$\boldsymbol{\rho}_{tj}|\sigma^2_{tj} \sim N_{t-1}\Big(\boldsymbol{M}^{-1}_{(t-1)j}\boldsymbol{m}_{tj}, \sigma^2_{tj}\boldsymbol{M}^{-1}_{(t-1)j}\Big) \tag{8.3.13}$$

and $\sigma^2_{1j} \sim IG((\nu - T + 1)/2, m_{1j}/2)$. We set $\boldsymbol{M}_j = \tau\boldsymbol{I}_T$ which gives a prior for σ^2_{tj} of

$$\sigma^2_{tj}|\tau, \nu \sim IG\Big(\frac{\nu - T + t}{2}, \frac{\tau}{2}\Big). \tag{8.3.14}$$

As $\boldsymbol{\rho}_{tj}$ can be interpreted as an additional set of regression parameters alongside a design matrix of residuals $\boldsymbol{U}_{(t-1)j}$, we augment the normal prior (8.3.13) with a latent variable η_{tkj}. This serves to reduce the noise in the model by performing a type of covariance selection, conveniently bypassing the difficulties which can be encountered when selecting parameters within a positive definite matrix. Our approach is an alternative to Gaussian graphical models (Wang (2015) and Banterle and Lewin (2018)) which allows us to scale up the model to high dimensions whilst imposing sparsity over the reparameterised space and maintaining computational feasibility. By allowing

the sparsity parameter λ_j to vary across the clusters, the full prior is

$$p(\rho_{tkj}|\sigma_{tj}^2, \tau, \eta_{tkj}) = \left[\frac{1}{\sqrt{2\pi}}\left(\frac{\tau}{\sigma_{tj}^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{\tau}{2\sigma_{tj}^2}\rho_{tkj}^2\right\}\right]^{\eta_{tkj}} \delta_0(\rho_{tkj})^{1-\eta_{tkj}}$$
$$p(\eta_{tkj}|\lambda_j) = \lambda_j^{\eta_{tkj}}(1-\lambda_j)^{1-\eta_{tkj}}, \quad \eta_{tkj} \in \{0,1\}.$$

Feature selection is performed on the $G \times p$ matrix $\boldsymbol{\upsilon}$ matrix, where the latent vector $\boldsymbol{\epsilon}$ ensures the covariates which determine the clustering are the same for each gate but the associated parameters are free to vary, by the multivariate Gaussian "spike-and-slab" prior

$$p(\boldsymbol{\upsilon}|d, \boldsymbol{\epsilon}) = \prod_s \left\{\left[\prod_{g=1}^{G}(2\pi)^{-1/2}(d)^{-1/2}\exp\left\{-\frac{1}{2d}\left\|\upsilon_{gs}\right\|^2\right\}\right]^{\epsilon_s} \delta_0(\boldsymbol{\upsilon}_s)^{1-\epsilon_s}\right\} \quad \upsilon_{gs} \in \mathbb{R}, \tag{8.3.15}$$

with a Bernoulli prior on ϵ_s

$$p(\epsilon_s|\kappa) = \kappa^{\epsilon_s}(1-\kappa)^{1-\epsilon_s}. \tag{8.3.16}$$

The selection of the gating parameters by the latent variable ϵ_s is performed on the individual weight across all gating networks. Thus, it provides a selection mechanism across the clustering.

8.3.3 Variation inference priors

Given the large number of parameters in the model and its potential application on big datasets, we employ Variational Inference (VI) (Blei et al., 2017) as our estimation procedure. The goal is to find a variational distribution $q(\boldsymbol{\vartheta})$ which is closest in Kullback-Leibler (KL) distance to the true posterior distribution, where all of the model parameters are denoted by $\boldsymbol{\vartheta}$. We do this by optimising the evidence lower bound (ELBO) with respect to the approximating density $q(\boldsymbol{\vartheta})$.

We restrict the space of approximating densities to solve the ELBO by using a variant of the mean-field variational family where the latent variables are mutually independent and each governed by a distinct factor in the variational density. The dependencies between the parameters, such as the latent indicator variable and their associated parameter(s), are incorporated within each member

(block). We define our block-mean-field approximation distribution as

$$
\begin{aligned}
q(\boldsymbol{\vartheta}) = & \left\{ \prod_t \prod_s \prod_j q(\beta_{tsj}, \gamma_{tsj}) \right\} \times \left\{ \prod_t q(w_t) \right\} \times \left\{ \prod_s \prod_j q(\omega_{sj}) \right\} \times \\
& \left\{ \prod_s q(\boldsymbol{\upsilon}_s, \epsilon_s) \right\} \times q(d) \times q(\kappa) \times \\
& \left\{ \prod_t \prod_{k<t} \prod_j q(\rho_{tkj}, \eta_{tkj}) \right\} \times \left\{ \prod_t \prod_j q(\sigma^2_{tj}) \right\} \times \\
& \left\{ \prod_j q(\lambda_j) \right\} \times \left\{ \prod_i \prod_g q(z_{ig}) \right\} \times q(b_w) \times q(b_d) \times q(\tau).
\end{aligned} \tag{8.3.17}
$$

We choose to optimise the ELBO using coordinate ascent variational inference (CAVI), which exploits the independence across the approximating densities imposed by the block-mean-field family. The updates take the general form of

$$
q_j(\vartheta_j) \propto \exp\big(\mathbb{E}_{q(\boldsymbol{\vartheta}_{-j}}[\log p(\vartheta_j | \boldsymbol{Y}, \boldsymbol{\vartheta}_{-j})]\big). \tag{8.3.18}
$$

By choosing conditionally conjugate priors, each marginal posterior and the corresponding variational expectation, is available in analytical form.

A difficulty lies with the sigmoid function in (8.3.8), which spoils the conjugate-exponential structure of the model. The variational update for $\boldsymbol{\upsilon}_s$ is not available analytically because our Gaussian spike-and-slab prior for $\boldsymbol{\upsilon}_s$, is not conjugate to the Bernoulli probability of the latent z_{ig} gating variable

$$
\begin{aligned}
p(z_{ig} | \boldsymbol{x}_{i,.}, \boldsymbol{\upsilon}_g, \boldsymbol{\epsilon}) &= \sigma\left(\sum_s x_{i,s} \epsilon_s \upsilon_{gs} \right)^{z_{ig}} \left[1 - \sigma\left(\sum_s x_{i,s} \epsilon_s \upsilon_{gs} \right) \right]^{1 - z_{ig}} \\
&= \sigma\left(\boldsymbol{\upsilon}^T_{g,\epsilon} \boldsymbol{x}_{i,.} \right)^{z_{ig}} \left[1 - \sigma\left(\boldsymbol{\upsilon}^T_{g,\epsilon} \boldsymbol{x}_{i,.} \right) \right]^{1 - z_{ig}}.
\end{aligned} \tag{8.3.19}
$$

The probability distribution in (8.3.19) is augmented with the latent selection variable ϵ_s due to the spike-and-slab prior on $\boldsymbol{\upsilon}_s$.

In order to retain the conjugate-exponential structure for $\boldsymbol{\upsilon}_s$ in the model, we introduce a "local"

lower bound on the group of z_{ig} variables in the model (introduced by Jaakkola and Jordan (1997)) which will combine with its Gaussian spike-and-slab prior. The lower bound is achieved by transforming the sigmoid function so that it is convex, and then approximating it by a first order Taylor series (derived in the Supplementary Section)

$$\sigma(x) \geq T(x, \xi) \equiv \sigma(\xi) \exp\left(\frac{x-\xi}{2} - \lambda_*(\xi)(x^2 - \xi^2)\right) \tag{8.3.20}$$

where $\lambda_*(\xi) = \tanh(\xi/2)/4\xi$.

The prior distribution for $p(z_{ig}|\boldsymbol{x}_{i,\cdot}, \boldsymbol{\upsilon}_g)$ in (8.3.19) is thus replaced by its lower bound

$$p(z_{ig}|\boldsymbol{x}_{i,\cdot}, \boldsymbol{\upsilon}_g, \boldsymbol{\epsilon}) \geq \exp\big(z_{ig}\boldsymbol{\upsilon}_{g,\epsilon}^T\boldsymbol{x}_{i,\cdot}\big)\sigma(\psi_{ig}) \exp\left(\frac{-\boldsymbol{\upsilon}_{g,\epsilon}^T\boldsymbol{x}_{i,\cdot} - \psi_{ig}}{2} - \lambda_*(\psi_{ig})((\boldsymbol{\upsilon}_{g,\epsilon}^T\boldsymbol{x}_{i,\cdot})^2 - \psi_{ig}^2)\right), \tag{8.3.21}$$

for the joint variation update of $q(\boldsymbol{\upsilon}_s, \epsilon_s)$.

The lower bound on the distribution of latent indicator variable z_{ig} introduces an additional variational parameter ψ_{ig}, for each data point and gating node, which is optimised by maximising the lower bound on the marginal likelihood. The approximating densities now maximise $\mathcal{L}(\tilde{q})$, rather than $\mathcal{L}(q)$, as the target density has been approximated,

$$\mathcal{L}(q) \geq \mathcal{L}(\tilde{q}) = \mathbb{E}_{\tilde{q}(\boldsymbol{\vartheta})}[\log p(\mathbf{y}, \boldsymbol{\vartheta}_{-\mathbf{z}})h(\mathbf{z}|\boldsymbol{\psi}, \boldsymbol{\upsilon})] - \mathbb{E}_{\tilde{q}(\boldsymbol{\vartheta})}[\log \tilde{q}(\boldsymbol{\vartheta})]). \tag{8.3.22}$$

8.3.4 Variational inference updates

The variational updates for the approximating densities are all available in closed form. The impact of the linked likelihood factorisation for the multiple responses in (8.3.2) can be seen by the presence of the $\boldsymbol{\rho}$ and $\boldsymbol{u}$ terms in the updates for the parameters directly associated with the multivariate regression $(\sigma_{tj}^2, \boldsymbol{\rho}_{tj}, \boldsymbol{\beta}_{tj})$. Unlike independent updates, information is borrowed across the responses as q expectations from parameters in the other $T-1$ regressions are now included in the analytical update. As expected, these updates are very similar to the multivariate response

regression without clustering Scott and Lewin (2022), but now include the vector of q expectations $(\boldsymbol{\zeta}_j)^{(1)}$. This marginal probability of belonging to a particular cluster performs a type of shrinkage, the nature of which depends on the parameter being approximated. For example, in the case of β_{tsj}

$$\tilde{q}(\beta_{tsj}|\gamma_{tsj}=1)=\mathcal{N}(\mu_{\beta_{tsj}},\sigma^2_{\beta_{tsj}}),\quad \tilde{q}(\beta_{tsj}|\gamma_{tsj}=0)=\delta_0(\beta_{tsj})$$

with the free parameters

$$\sigma^2_{\beta_{tsj}}=\left(||(\boldsymbol{\zeta}_j)^{(1)}\odot\boldsymbol{x}^2_{.,s}||_1\left\{(\sigma_{tj}^{-2})^{(1)}+\sum_{k>t}(\rho_{ktj})^{(1)}(\sigma^2_{kj})^{(1)}\right\}+(w_t^{-1})^{(1)}\right)^{-1} \tag{8.3.23}$$

$$\begin{aligned}\mu_{\beta_{tsj}}=&\sigma^2_{\beta_{tsj}}((\boldsymbol{\zeta}_j)^{(1)}\odot\boldsymbol{x}_{.,s})^T\left[(\sigma_{tj}^{-2})^{(1)}\left((u_{t-sj})^{(1)}-\sum_{k<t}(u_{kj})^{(1)}(\rho_{tkj})^{(1)}\right)+\right.\\&\left.+\sum_{k>t}(\sigma_{kj}^{-2})^{(1)}(u_{t-sj})^{(1)}(\rho_{ktj})^{(2)}-\sum_{k>t}(\sigma_{kj}^{-2})^{(1)}(\rho_{ktj})^{(1)}\left((u_{kj})^{(1)}-\sum_{h<k,h\neq t}(u_{hj})^{(1)}(\rho_{khj})^{(1)}\right)\right].\end{aligned} \tag{8.3.24}$$

The marginal probability of belonging to cluster j for each data point shrinks the sth covariate in the free parameter updates for the mean and variance of $\tilde{q}(\beta_{tsj})$.

The approximating densities for the features all have an approximating density which is in the same form as their prior, a Gaussian spike-and-slab. The latent indicator variables γ_{tsj}, η_{tkj} and ϵ_s all serve to shrink the marginal expectation of the corresponding parameter associated with the covariate, inversely proportional to the marginal probability of inclusion. Hence, the covariates must be standardised. The respective q expectations for the parameters associated with the covariates are

$$\mathbb{E}_q[\beta_{tsj}]=\mu_{\beta_{tsj}}(\gamma_{tsj})^{(1)},\quad \mathbb{E}_q[\rho_{tkj}]=\mu_{\rho_{tkj}}(\eta_{tkj})^{(1)},\quad \mathbb{E}_q[\boldsymbol{\upsilon}_s]=\boldsymbol{\mu}_{\upsilon_s}(\epsilon_s)^{(1)}. \tag{8.3.25}$$

The approximating density for the local variables z_{ig}, which are found in combination within the

cluster variable $\boldsymbol{\zeta}_i$, takes the form

$$\tilde{q}(z_{ig}) = \text{Bernouli}\Big(\sigma((C_{ig})^{(1)})\Big) \tag{8.3.26}$$

where

$$(C_{ig})^{(1)} = (\boldsymbol{\upsilon}_g)^{(1)T}\boldsymbol{x}_{i,\cdot} + \sum_{j\in\mathcal{E}_g^L}(\zeta_{ij}^{\not g})^{(1)}(A_{ij})^{(1)} - \sum_{j\in\mathcal{E}_g^R}(\zeta_{ij}^{\not g})^{(1)}(A_{ij})^{(1)} \tag{8.3.27}$$

and $\mathcal{E}_g^R$ and $\mathcal{E}_g^L$ denote the set of experts on the right-hand-side and left-hand-side of the gth gate respectively and

$$\mathbb{E}_{\tilde{q}(-z_{ig})}[\zeta_{ij}^{\not g}] = \prod_{l=1,l\neq g}^{G}(z_{il})^{(1)S^L(j,l)}(1-(z_{il})^{(1)})^{S^R(j,l)}$$

with $(A_{ij})^{(1)}$ is defined as

$$\begin{aligned}(A_{ij})^{(1)} = &-\frac{T}{2}\log(2\pi) - \sum_t \frac{(\log\sigma_{tj}^{-2})^{(1)}}{2} - \sum_t \frac{(\sigma_{tj}^{-2})^{(1)}}{2}\Bigg((u_{itj})^{(2)} - 2(u_{itj})^{(1)}\sum_{k<t}(u_{ikj})^{(1)}(\rho_{tkj})^{(1)} + \\ &+ 2\sum_{k'\neq k}(u_{ikj})^{(1)}(\rho_{tkj})^{(1)}(u_{ik'j})^{(1)}(\rho_{tk'j})^{(1)} + \sum_{k<t}(u_{ikj})^{(2)}(\rho_{tkj})^{(2)}\Bigg)\end{aligned} \tag{8.3.28}$$

The update requires the computation of the likelihood for each data point $\boldsymbol{y}_i$. Although VI scales to very large datasets, if n is in the order of millions, the time it takes to estimate the model may still be prohibitive.

The update for the local variables z_{ig} is achieved by replacing the prior distribution by the the "local" lower bound (8.3.21). This introduces an additional parameter into the model, ψ_{ig}, which instead of placing a prior on, we treat as a type of tuning parameter and use an empirical Bayes approach. This is the opposite to the frequentist EM algorithm, where the likelihood is augmented with variables to make the computation of the maximum likelihood estimates tractable. Here the variational parameter ψ_{ig} allows us to compute the variational expectations $q(\boldsymbol{\upsilon}_s, \epsilon_s)$ analytically, and we maximise this nuisance parameter at each iteration to ensure $\mathcal{L}(\tilde{q})$ is as close as possible

to $\mathcal{L}(q)$ via the update

$$\psi_{ig}^{(\text{new})} = \sqrt{\boldsymbol{x}_{i,\cdot}^T \mathbb{E}_{\tilde{q}(\upsilon)}[\boldsymbol{\upsilon}_g \boldsymbol{\upsilon}_g^T] \boldsymbol{x}_{i,\cdot}}. \tag{8.3.29}$$

The algorithm, which is run until $\mathcal{L}(\tilde{q})$ indicates convergence to a local optimum, is

Algorithm 9: CAVI in the HME model

Input : A model $p(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\vartheta})$, a data set $\boldsymbol{Y}$, a design matrix $\boldsymbol{X}$

Output : A variational density $q(\boldsymbol{\vartheta}) = \prod_{h=1}^{m} q_h(\vartheta_h)$

Intialize: Variational factors $q_h(\vartheta_h)$

while *the lower bound on the ELBO,* $\mathcal{L}(\tilde{q})$*, has not converged* **do**

  for $h \in \{1, ..., m\}$ **do**

    Set $q_h(\vartheta_h) \propto \exp\{\mathbb{E}_{-h}[\log p(\vartheta_h | \boldsymbol{\vartheta}_{-h}, \boldsymbol{Y})]\}$

  end

  for $i \in \{1, ..., n\}$ **do**

    for $g \in \{1, ..., G\}$ **do**

      Set $\psi_{ig}^{(\text{new})} = \sqrt{\boldsymbol{x}_{i,\cdot}^T \mathbb{E}_{\tilde{q}(\upsilon)}[\boldsymbol{\upsilon}_g \boldsymbol{\upsilon}_g^T] \boldsymbol{x}_{i,\cdot}}$

    end

  end

  Compute $\mathcal{L}(\tilde{q}) = \mathbb{E}_{\tilde{q}(\boldsymbol{\vartheta})}[\log p(\mathbf{y}, \boldsymbol{\vartheta}_{-\mathbf{z}}) h(\mathbf{z} | \boldsymbol{\psi}, \boldsymbol{\upsilon})] - \mathbb{E}_{\tilde{q}(\boldsymbol{\vartheta})}[\log \tilde{q}(\boldsymbol{\vartheta})]$

end

return $q(\boldsymbol{\vartheta})$

8.4 Discussion

Our model extends the big data regression model of Scott and Lewin (2022), which regresses multidimensional responses related through a latent structure with high dimensional multivariate predictors, to a mixture model through the HME framework. The latent structure of the multidimensional response is free to vary across the cluster, enabling the identification of groups of individuals who exhibit different causal pathways to the same endpoint. The unsupervised

learning of detecting the clusters is determined by the multivariate predictors. Covariate selection priors for the parameters within the mixing coefficients identify the important covariates across the gating network. Feature selection priors within the likelihood, exploit the expected sparsity and allow the associated variables to vary across the responses. A hierarchical prior framework enables the leveraging of information across responses within the model, aiding identification of important covariates. The reparameterisation of a matrix normal likelihood alongside feature selection, allows the model to accommodate either sparse or dense residual covariance structures for different clusters, bypassing the considerable computational challenge encountered with Gaussian graphical models. In terms of the mean squared error of a future value (where the expectation is with respect to the data), the shrinkage from the latent indicator variables adds bias to the model estimation, in return for a large reduction in model estimation variance, to ensure the model is generalisable.

The CAVI approach involves iterating though local and global parameter updates, providing fast estimation of the model with very large datasets. The approach can accommodate large biological datasets where $p >> n$ and p is in the order of millions. However, the local updates involves estimating the free parameters for z_{ig}, per data point and gate. This can slow the algorithm when n is of a large orders of magnitude. Our approach can be easily adapted by using stochastic variational inference (SVI) (Hoffman et al., 2013), so that the computational speed is maintained. Rather than ascending $\mathcal{L}(\tilde{q})$ via co-ordinate ascent, SVI uses ascent by natural gradient in a stochastic optimisation algorithm. The result is a minor change to the global updates outlined in the Supplementary Section. A subsample of the data (sample $<< n$) is repeatedly taken to form noisy but cheap to compute estimates of the natural gradient of $\mathcal{L}(\tilde{q})$, which are followed with a decreasing step size. Only the local parameters for the randomly sampled data points are estimated and the global updates are a weighted average of the current and new update. These learning rates can be optimised by allowing them to adapt to the properties of the sampled data (Ranganath et al., 2013).

The number of clusters $(G + 1)$ must be defined by the users. This is expected to be small, and will be optimised over a small set by using a loss function with cross validation. An alternative

approach, which still enables the clustering to be determined by the covariates but allows the data to determine $G+1$, is a Random Partition Model with covariates (PRMx) (Müller and Quintana, 2010). A PRMx is characterized by specifying a Dirichlet Process prior (Ferguson, 1973) on the parameters, alongside the feature selection priors to create a covariate dependent Dirichlet Process Mixture model.

8.5 Appendix

8.5.1 Parameterisation

The following tables provide a summary of the indexes and terms which are used in the derivation of the CAVI updates for the multivariate response HME model. The number of gates G, are defined by the user.

Index	Elements
$t = 1, ..., T$	Responses
$s = 1, ..., p$	Covariates
$k = 1, ..., T-1$	ρ_{tk} elements
$j = 1, ..., J$	Experts
$g = 1, ..., G$	Gating nodes
$i = 1, ..., n$	Data points

Notation	Order of Index	Interpretation
y_{it}	Individual, Response	Data Point
u_{ikj}	Individual, Response, Cluster	Regression Residual
x_{is}	Individual, Covariate	Design Matrix Point
β_{tsj}	Response, Covariate, Cluster	Mean Regression Parameter
γ_{tsj}	Response, Covariate, Cluster	Covariate Indicator for β
ρ_{tkj}, ρ_{ktj}	Response, Covariate, Cluster	Residual Regression Parameter
η_{tkj}	Response, Covariate, Cluster	Covariance Indicator for ρ
υ_{gs}	Gate, Cluster	Cluster Regression Parameter
ϵ_s	Covariate	Covariance Indicator for υ
ζ_{ij}	Individual, Cluster	Cluster Indicator
z_{ig}	Individual, Gate	Latent Tree Indicator

The likelihood for the vector of observations where $\boldsymbol{y}_i \in \mathbb{R}^T$ is

$$p(\boldsymbol{y}_i|\cdot) = \prod_j \left(\prod_t (2\pi)^{-1/2} (\sigma_{tj}^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma_{tj}^2} \left\| y_{it} - \boldsymbol{x}_{i,\cdot}^T \beta_{tj} - \sum_{k<t} u_{ikj} \rho_{tkj} \right\|^2 \right\} \right)^{\zeta_{ij}}. \tag{8.5.1}$$

We assume that each column of the response matrix ($\boldsymbol{Y}$) has been centred, and each column of the design matrix ($\boldsymbol{X}$) has been standardised. The variance term of the data (σ_{tj}^2) is always expressed in this form, rather than in terms of the standard deviation, to avoid confusion with $\sigma(\cdot)$ which represents the sigmoid function.

The prior specification is

$$p(\beta_{tsj}|\gamma_{tsj}, w_t) = \left[(2\pi)^{-1/2} (w_t)^{-1/2} \exp\left\{ -\frac{1}{2w_t} \|\beta_{tsj}\|^2 \right\} \right]^{\gamma_{tsj}} \delta_0(\beta_{tsj})^{1-\gamma_{tsj}} \quad \beta_{tsj} \in \mathbb{R} \tag{8.5.2}$$

$$p(\gamma_{tsj}|\omega_{sj}) = \omega_{sj}^{\gamma_{tsj}} (1-\omega_{sj})^{1-\gamma_{tsj}} \quad \gamma_{tsj} \in \{0,1\} \tag{8.5.3}$$

$$p(\rho_{tkj}|\sigma_{tj}^2, \tau, \eta_{tkj}) = \left[\frac{1}{\sqrt{2\pi}} \left(\frac{\tau}{\sigma_{tj}^2} \right)^{\frac{1}{2}} \exp\left\{ -\frac{\tau}{2\sigma_{tj}^2} \rho_{tkj}^2 \right\} \right]^{\eta_{tkj}} \delta_0(\rho_{tkj})^{1-\eta_{tkj}} \quad \rho_{tkj} \in \mathbb{R} \tag{8.5.4}$$

$$p(\eta_{tkj}|\lambda_j) = \lambda_j^{\eta_{tkj}} (1-\lambda_j)^{1-\eta_{tkj}} \quad \eta_{tkj} \in \{0,1\} \tag{8.5.5}$$

$$p(\sigma_{tj}^2|\tau, \nu) = \frac{1}{\Gamma\left(\frac{\nu-T+t}{2}\right)} \left(\frac{\tau}{2\sigma_{tj}^2} \right)^{\frac{\nu-T+t}{2}} \frac{1}{\sigma_{tj}^2} \exp -\frac{\tau(\sigma_{tj}^2)^{-1}}{2} \quad \sigma_{tj}^2 > 0 \tag{8.5.6}$$

$$p(z_{ig}|\boldsymbol{x}_{i,\cdot}, \upsilon_g) = \sigma(\boldsymbol{\upsilon}_g^T \boldsymbol{x}_{i,\cdot})^{z_{ig}} [1 - \sigma(\boldsymbol{\upsilon}_g^T \boldsymbol{x}_{i,\cdot})]^{1-z_{ig}} \quad z_{ig} \in \{0,1\} \tag{8.5.7}$$

$$p(\boldsymbol{\upsilon}_s|d, \epsilon_s) = \left[\prod_{g=1}^{G} (2\pi)^{-1/2} (d)^{-1/2} \exp\left\{ -\frac{1}{2d} \|\upsilon_{gs}\|^2 \right\} \right]^{\epsilon_s} \delta_0(\boldsymbol{\upsilon}_s)^{1-\epsilon_s} \quad \boldsymbol{\upsilon}_s \in \mathbb{R}^G \tag{8.5.8}$$

$$p(\epsilon_s|\kappa) = \kappa^{\epsilon_s} (1-\kappa)^{1-\epsilon_s} \quad \epsilon_s \in \{0,1\}. \tag{8.5.9}$$

We define ζ_{ij} via the gating network topology as

$$\zeta_{ij} = \prod_{g=1}^{G} z_{ig}^{S^L(j,g)} (1 - z_{ig})^{S^R(j,g)}. \tag{8.5.10}$$

$\mathbf{S}^L$ and $\mathbf{S}^R$ are matrices where $\mathbf{S}^L(j,g) = 1$ if the jth expert is on the left sub-tree of the gth gate, and zero otherwise. Similarly, $\mathbf{S}^R(j,g) = 1$ if the jth expert is on the right sub-tree of the gth

gate, and zero otherwise.

The hierarchical hyperprior specification is

$$p(\omega_{sj}|a_\omega, b_\omega) = \frac{1}{B(a_\omega, b_\omega)}\omega_{sj}^{a_\omega - 1}(1-\omega_{sj})^{b_\omega - 1} \quad 0 \leq \omega_{sj} \leq 1 \tag{8.5.11}$$

$$p(w_t|a_w, b_w) = \frac{b_w^{a_w}}{\Gamma(a_w)}(w_t)^{-a_w - 1}\exp\{-b_w w_t^{-1}\} \quad w_t > 0 \tag{8.5.12}$$

$$p(b_w|a_{b_w}, b_{b_w}) = \frac{b_{b_w}^{a_{b_w}}}{\Gamma(a_{b_w})}(b_w^{a_{b_w}-1})\exp\{-b_{b_w} b_w\} \quad b_w > 0 \tag{8.5.13}$$

$$p(d|a_d, b_d) = \frac{b_d^{a_d}}{\Gamma(a_d)}(d)^{-a_d - 1}\exp\{-b_d d^{-1}\} \quad d > 0 \tag{8.5.14}$$

$$p(b_d|a_{b_d}, b_{b_d}) = \frac{b_{b_d}^{a_{b_d}}}{\Gamma(a_{b_d})}(b_d^{a_{b_d}-1})\exp\{-b_{b_d} b_d\} \quad b_d > 0 \tag{8.5.15}$$

$$p(\lambda_j) = \frac{1}{B(a_\lambda, b_\lambda)}\lambda_j^{a_\lambda - 1}(1-\lambda_j)^{b_\lambda - 1} \quad 0 \leq \lambda \leq 1 \tag{8.5.16}$$

$$p(\tau) = \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)}(\tau)^{a_\tau - 1}\exp\{-b_\tau \tau\} \quad \tau > 0 \tag{8.5.17}$$

$$p(\kappa) = \frac{1}{B(a_\kappa, b_\kappa)}\kappa^{a_\kappa - 1}(1-\kappa)^{b_\kappa - 1} \quad 0 \leq \kappa \leq 1 \tag{8.5.18}$$

with the tuning parameter $\psi_{ig} \in \mathbb{R}^1$, optimised at each iteration of the CAVI.

The joint posterior is

$$p(\boldsymbol{Y}, \boldsymbol{\vartheta}) = \left\{\prod_i \prod_t \prod_j p(y_{it}|\boldsymbol{x}_{i,.}, \boldsymbol{\beta}_{tj}, \sigma_{tj}^2, \boldsymbol{\rho}_{tj}, \zeta_{ij})\right\} \times \tag{8.5.19}$$

$$\left\{\prod_t \prod_s \prod_j p(\beta_{tsj}|w_t, \gamma_{tsj})\right\} \times \left\{\prod_t \prod_s \prod_j p(\gamma_{tsj}|\omega_{sj})\right\} \times$$

$$\left\{\prod_i \prod_g p(z_{ig}|\boldsymbol{x}_{i,.}, \boldsymbol{\upsilon}_i)\right\} \times \left\{\prod_g \prod_s p(\upsilon_{gs}|d, \epsilon_s)\right\} \times \left\{\prod_s p(\epsilon_s|\kappa)\right\} \times$$

$$\left\{\prod_s \prod_j p(\omega_{sj})\right\} \times \left\{\prod_j \prod_t p(\sigma_{tj}^2|\tau, \nu) \prod_{k<t} p(\rho_{tkj}|\sigma_{tj}^2, \tau, \eta_{tkj})\right\} \times \tag{8.5.20}$$

$$\left\{\prod_j \prod_t \prod_{k<t} p(\eta_{tkj}|\lambda)\right\} \times \left\{\prod_s p(\kappa_s)\right\} \times p(d) \times p(w|b_w) \times \times p(\lambda) \times p(b_w) \times p(\tau).$$

Define the block-mean-field approximation distribution as

$$
\begin{aligned}
q(\boldsymbol{\vartheta}) = &\left\{\prod_t \prod_s \prod_j q(\beta_{tsj}, \gamma_{tsj})\right\} \times \left\{\prod_t q(w_t)\right\} \times \left\{\prod_s \prod_j q(\omega_{sj})\right\} \times \\
&\left\{\prod_s q(\boldsymbol{\upsilon}_s, \epsilon_s)\right\} \times q(d) \times q(\kappa) \times \\
&\left\{\prod_t \prod_{k<t} \prod_j q(\rho_{tkj}, \eta_{tkj})\right\} \times \left\{\prod_t \prod_j q(\sigma^2_{tj})\right\} \times \left\{\prod_j q(\lambda_j)\right\} \times \\
&\left\{\prod_i \prod_g q(z_{ig})\right\} \times \times q(b_w) \times q(b_d) \times q(\tau),
\end{aligned}
$$

with $f(\mathbf{z})^{(j)}$ as the j-th moment of $f(\mathbf{z})$ with respect to $q(\mathbf{z})$, $\mathbb{E}_q\left[f(\boldsymbol{z})^j\right]$.

The approximating densities maximise $\mathcal{L}(\tilde{q})$ rather than $\mathcal{L}(q)$ ($\mathcal{L}(\tilde{q}) \leq \mathcal{L}(q)$) because of the lower bound approximation of the distribution for the latent local variable z_{ig}. The block-mean-field distribution remains unchanged, but there is now an additional variational parameter ψ_{ig}.

To simplify notation we introduce the scalar terms

$$(u_{itj})^{(1)} = y_{it} - \boldsymbol{x}^T_{i,\cdot}(\boldsymbol{\beta}_{tj})^{(1)} \tag{8.5.21}$$

$$(u_{itj,-s})^{(1)} = y_{it} - \sum_{l,l\neq s} x_{il}(\beta_{tlj})^{(1)} \tag{8.5.22}$$

$$(u_{itj})^{(1)} = u^{(1)}_{it,-sj} - x_{is}(\beta_{tsj})^{(1)} \tag{8.5.23}$$

$$(u_{itj})^{(2)} = y^2_{it} - 2y_{it}\boldsymbol{x}^T_{i,\cdot}(\boldsymbol{\beta}_{tj})^{(1)} + \sum_s x^2_{is}(\beta_{tsj})^{(2)} - 2\sum_{s<s'} x_{is}x_{is'}(\beta_{tsj})^{(1)}(\beta_{ts'j})^{(1)}, \tag{8.5.24}$$

where $\boldsymbol{x}_{i,\cdot}$ is a row of the design matrix.

The element wise multiplication of a column of the design matrix with the q expectations ($\mathbb{E}_q[\zeta_{ij}]$) of the parameter ζ_{ij} for all units of a particular cluster is

$$\boldsymbol{x}_{\cdot,s} \odot (\boldsymbol{\zeta}_j)^{(1)}. \tag{8.5.25}$$

The vectors are defined as

$$(\boldsymbol{u}_{tj})^{(1)} = \boldsymbol{y}_t - \sum_s \boldsymbol{x}_{.,s}(\beta_{tsj})^{(1)} \tag{8.5.26}$$

$$(\boldsymbol{u}_{t-sj})^{(1)} = \boldsymbol{y}_t - \sum_{l \neq s} \boldsymbol{x}_{.,l}(\beta_{tlj})^{(1)} \tag{8.5.27}$$

$$(\boldsymbol{u}_{tj})^{(1)} = (\boldsymbol{u}_{t-sj})^{(1)} - \boldsymbol{x}_{.,s}(\beta_{tsj})^{(1)}, \tag{8.5.28}$$

where $\boldsymbol{x}_{.,s}$ is a vector from a column of the design matrix.

8.5.2 HME CAVI updates

$$\log \tilde{q}(\beta_{tsj}, \gamma_{tsj}) = \mathbb{E}_{-(\beta_{tsj},\gamma_{tsj})} \left[\log \prod_i \left(\prod_t p(y_{it}|\boldsymbol{x}_{i,.}, u_{i(t-1)j}, \boldsymbol{\beta}_{tj}, \boldsymbol{\rho}_{tj}, \sigma^2_{tj}) \right)^{\zeta_{ij}} \right] + \tag{8.5.29a}$$

$$\mathbb{E}_{-(\beta_{tsj},\gamma_{tsj})} \left[\log p(\beta_{tsj}|\gamma_{tsj}, w_t) \right] + \tag{8.5.29b}$$

$$\mathbb{E}_{-(\beta_{tsj},\gamma_{tsj})} \left[\log p(\gamma_{tsj}|\omega_{sj}) \right] + cst \tag{8.5.29c}$$

Equation (8.5.29b) and (8.5.29c) can be computed easily.

$$(8.5.29b): \quad -\gamma_{tsj} \left(\frac{1}{2} (w_t^{-1})^{(1)} \|\beta_{tsj}\|^2 \right) + (1 - \gamma_{tsj}) \, \delta_0(\beta_{tsj}) + \frac{\gamma_{ts}}{2} \left[(\log w_t^{-1})^{(1)} - \log 2\pi \right]$$

$$(8.5.29c): \quad cst + \gamma_{tsj} \left(\log \omega_{sj} \right)^{(1)} + (1 - \gamma_{tsj}) \left(\log(1 - \omega_{sj}) \right)^{(1)},$$

and we can write Equation (8.5.29a) as

$$
\begin{aligned}
a = &\quad a^1_{itj} + \sum_{k>t} a^2_{ikj} \\
a^1_{itj} = &\quad \mathbb{E}_{-(\beta_{tjs},\gamma_{tsj})}\left[\log \prod_i \left(\prod_t p(y_{it}|\boldsymbol{x}_{i,.}, u_{i(t-1)j}, \boldsymbol{\beta}_{tj}, \boldsymbol{\rho}_{tj}, \sigma^2_{tj})\right)^{\zeta_{ij}}\right] \\
a^2_{ikj} = &\quad \sum_{k>t} \mathbb{E}_{-(\beta_{tsj},\gamma_{tsj})}\left[\log \prod_i \left(\prod_k p(y_{ik}|\boldsymbol{x}_{i,.}, u_{i(k-1)j}, \boldsymbol{\beta}_{kj}, \boldsymbol{\rho}_{kj}, \sigma^2_{kj})\right)^{\zeta_{ij}}\right].
\end{aligned}
$$

Expanding a^1_{itj}

$$
\begin{aligned}
a^1_{itj} = &\quad \mathbb{E}_{-(\beta_{tsj},\gamma_{tsj})}\left[-\sum_i \frac{\zeta_{ij}}{2\sigma^2_{tj}}\left(u_{itj,-s} - x_{is}\gamma_{tsj}\beta_{tsj} - \sum_{k<t} u_{ikj}\eta_{tkj}\rho_{tkj}\right)^2\right] \\
\propto &\quad \mathbb{E}_{-(\beta_{tsj},\gamma_{tsj})}\left[-\frac{\gamma_{tsj}}{2\sigma^2_{tj}}\left(\sum_i \zeta_{ij}x^2_{is}\beta^2_{tsj} - 2\zeta_{ij}\beta_{tsj}x_{is}\left(u_{itj,-s} - \sum_{k<t} u_{ikj}\rho_{tkj}\eta_{tkj}\right)\right)\right].
\end{aligned}
$$

Expanding a^2_{ikj}

$$
\begin{aligned}
a^2_{ikj} = &\quad \mathbb{E}_{-(\beta_{tsj},\gamma_{tsj})}\left[\sum_i \sum_{k>t} -\frac{\zeta_{ij}}{2\sigma^2_{kj}}\left(y_{ik} - \sum_s x_{is}\gamma_{ksj}\beta_{ksj} - \sum_{h<k,h\neq t} u_{ihj}\eta_{khj}\rho_{khj} + \right.\right. \\
&\quad \left.\left. - \left(y_{it} - \sum_{l\neq s} x_{il}\gamma_{tlj}\beta_{tlj}\right)\rho_{ktj} + x_{is}\gamma_{tsj}\beta_{tsj}\rho_{ktj}\right)^2\right] \\
= &\quad \mathbb{E}_{-(\beta_{tsj},\gamma_{tsj})}\left[\sum_i \sum_{k>t} -\frac{\zeta_{ij}}{2\sigma^2_{kj}}\left(u_{ikj} - \sum_{h<k,h\neq t} u_{ihj}\rho_{khj} - u_{itj,-s}\rho_{ktj} + x_{is}\gamma_{tsj}\beta_{tsj}\rho_{ktj}\right)^2\right],
\end{aligned}
$$

$$
\begin{aligned}
a^2_{knj} \propto \mathbb{E}_{-(\beta_{tsj},\gamma_{tsj})}\Bigg[&-\frac{1}{2\sigma^2_{kj}}\Bigg(\sum_{k>t}\sum_i \zeta_{ij}\beta^2_{tsj}x^2_{is}\rho^2_{ktj} + \\
&- 2\sum_{k>t}\sum_i \zeta_{ij}x_{is}\beta_{tsj}\left\{\sum_{h<k,h\neq t} u_{ihj}\rho_{khj}\rho_{ktj} + u_{itj,-s}\rho^2_{ktj} - u_{ikj}\rho_{ktj}\right\}\Bigg)\Bigg].
\end{aligned}
$$

Bring together all the components

$$\begin{aligned}\log\tilde{q}(\beta_{tsj},\gamma_{tsj}) \propto & -\frac{\gamma_{tsj}}{2}\Bigg[\beta_{tsj}^2\Bigg(\sum_i\frac{(\zeta_{ij})^{(1)}x_{is}^2}{(\sigma_{tj}^2)^{(1)}}+\sum_{k>t}\sum_i\frac{(\zeta_{ij})^{(1)}x_{is}^2(\rho_{ktj})^{(2)}}{(\sigma_{kj}^2)^{(1)}}+(w_t^{-1})^{(1)}\Bigg)+\\ & -2\beta_{tsj}\sum_i(\zeta_{ij})^{(1)}x_{is}\Bigg\{\frac{1}{(\sigma_{tj}^2)^{(1)}}\Bigg((u_{itj,-s})^{(1)}-\sum_{k<t}(u_{ikj})^{(1)}(\rho_{tkj})^{(1)}\Bigg)+\\ & -\sum_{k>t}\frac{(\rho_{ktj})^{(1)}}{(\sigma_{kj}^2)^{(1)}}(u_{ikj})^{(1)}+(1-\gamma_{tsj})[\delta_0(\beta_{tsj})+\log(1-\omega_{sj})^{(1)}]+\\ & +\sum_{k>t}(\sigma_{kj}^{-2})^{(1)}\Big(\sum_{h<k,h\neq t}(u_{ihj})^{(1)}(\rho_{khj})^{(1)}(\rho_{ktj})^{(1)}+(u_{itj,-s})^{(1)}(\rho_{ktj})^2\Big)\Bigg\}\Bigg]+\\ & +\gamma_{tsj}\Bigg[-\frac{\log(2\pi)}{2}+(\log w_t^{-1})^{(1)}+(\log\omega_{sj})^{(1)}\Bigg].\end{aligned} \tag{8.5.30}$$

Using completing the square

$$\sigma_{\beta_{tsj}}^2=\Bigg(\sum_i(\zeta_{ij})^{(1)}x_{is}^2\Bigg\{(\sigma_{tj}^{-2})^{(1)}+\sum_{k>t}(\rho_{ktj})^{(2)}(\sigma_{kj}^{-2})^{(1)}\Bigg\}+(w_t^{-1})^{(1)}\Bigg)^{-1} \tag{8.5.31}$$

$$\begin{aligned}\mu_{\beta_{tsj}}=&\sigma_{\beta_{tsj}}^2\Bigg[\sum_i(\zeta_{ij})^{(1)}x_{is}\Bigg\{(\sigma_{tj}^{-2})^{(1)}\Bigg((u_{itj,-s})^{(1)}-\sum_{k<t}(u_{ikj})^{(1)}(\rho_{tkj})^{(1)}\Bigg)-\sum_{k>t}\frac{(\rho_{ktj})^{(1)}}{(\sigma_{kj}^2)^{(1)}}(u_{ikj})^{(1)}+\\ &+\sum_{k>t}(\sigma_{kj}^{-2})^{(1)}\Big(\sum_{h<k,h\neq t}(u_{ihj})^{(1)}(\rho_{khj})^{(1)}(\rho_{ktj})^{(1)}+(u_{itj,-s})^{(1)}(\rho_{ktj})^2\Big)\Bigg\}\Bigg)\Bigg].\end{aligned} \tag{8.5.32}$$

The variational expectation $(\zeta_{ij})^{(1)}$ shrinks the covariates which are not suitable for a particular cluster. Using the vectors defined in (8.5.26) to (8.5.28), the vectorised solution is

$$\sigma_{\beta_{tsj}}^2=\Bigg(||(\boldsymbol{\zeta}_j)^{(1)}\odot\boldsymbol{x}_{.,s}^2||_1\Bigg\{(\sigma_{tj}^{-2})^{(1)}+\sum_{k>t}(\rho_{ktj})^{(1)}(\sigma_{kj}^2)^{(1)}\Bigg\}+(w_t^{-1})^{(1)}\Bigg)^{-1} \tag{8.5.33}$$

$$\begin{aligned}\mu_{\beta_{tsj}}=&\sigma_{\beta_{tsj}}^2((\boldsymbol{\zeta}_j)^{(1)}\odot\boldsymbol{x}_{.,s})^T\Bigg[(\sigma_{tj}^{-2})^{(1)}\Bigg((u_{t-sj})^{(1)}-\sum_{k<t}(u_{kj})^{(1)}(\rho_{tkj})^{(1)}\Bigg)+\\ &+\sum_{k>t}(\sigma_{kj}^{-2})^{(1)}(u_{t-sj})^{(1)}(\rho_{ktj})^{(2)}-\sum_{k>t}(\sigma_{kj}^{-2})^{(1)}(\rho_{ktj})^{(1)}\Bigg((u_{kj})^{(1)}-\sum_{h<k,h\neq t}(u_{hj})^{(1)}(\rho_{khj})^{(1)}\Bigg)\Bigg],\end{aligned} \tag{8.5.34}$$

where $(\boldsymbol{\zeta}_j)^{(1)}$ is an n dimensional vector of variational expectations for the j cluster and $\boldsymbol{x}_{.,s}$ is an

n dimensional vector compising the sth column of the design matrix. The update is the same as the CAVI multivariate regression model, except now we select the appropriate covariates for each cluster in the operation (8.5.25), in the dot product.

Joining all the components together

$$p(\beta_{tsj}, \gamma_{tsj}) \propto -\frac{\gamma_{tsj}}{2}(\beta_{tsj} - \mu_{\beta_{tsj}})^2 + \frac{\gamma_{tsj}}{2\sigma^2_{\beta_{tsj}}}\mu^2_{\beta_{tsj}} + \gamma_{tsj}\Big[-\log\frac{2\pi}{2} + \frac{1}{2}(\log w_t^{-1})^{(1)} - (\log\omega_{sj})^{(1)}\Big] +$$
$$+ (1-\gamma_{tsj})[\delta_0(\beta_{tsj}) + \log(1-\omega_{sj})], \tag{8.5.35}$$

and exponentiating

$$\propto \left[\frac{1}{\sqrt{2\pi\sigma^2_{\beta_{tsj}}}}\exp\left(-\frac{1}{2\sigma^2_{\beta_{tsj}}}(\beta_{tsj} - \mu_{\beta_{tsj}})^2\right)\right]^{\gamma_{tsj}} \delta_0(\beta_{tsj})^{1-\gamma_{tsj}} \times \tag{8.5.36}$$
$$\left[(\sigma^2_{\beta_{tsj}})^{1/2}\exp\left(\frac{\mu^2_{\beta_{tsj}}}{2\sigma^2_{\beta_{tsj}}} + (\frac{1}{2}\log w_t^{-1})^{(1)} + (\log\omega_{sj})^{(1)}\right)\right]^{\gamma_{tsj}} \exp\Big(\log(1-\omega_{sj})^{(1)}\Big)^{1-\gamma_{tsj}}. \tag{8.5.37}$$

The law of iterative expectations is used to obtain the expectation $(\beta_{ts})^{(1)} = \mathbb{E}_{\tilde{q}}[\beta_{ts}]$, given that β_{ts} is parametrised by a mixture distribution

$$\mathbb{E}_{\tilde{q}}[\beta_{tsj}] = \mathbb{E}_{\tilde{q}(\gamma_{tsj})}[\mathbb{E}_{\tilde{q}}[\beta_{tsj}|\gamma_{tsj}]]$$
$$= \mu_{\beta_{tsj}}(\gamma_{tsj})^{(1)} + 0(1-(\gamma_{tsj})^{(1)}) = \mu_{\beta_{tsj}}(\gamma_{tsj})^{(1)},$$

and thus by calling

$$(\gamma_{tsj})^{(1)} = \left[1 + \sqrt{\sigma^{-2}_{\beta_{tsj}}}\exp\left\{(\log 1 - \omega_{sj})^{(1)} - (\log\omega_{sj})^{(1)} - \frac{1}{2}(\log w_t^{-1})^{(1)} - \frac{1}{2}\mu^2_{\beta,ts}\sigma^{-2}_{\beta,ts}\right\}\right]^{-1} \tag{8.5.38}$$

we have that under $\tilde{q}$

$$\beta_{tsj}|\gamma_{tsj} = 1 \sim \mathcal{N}(\mu_{\beta_{tsj}}, \sigma^2_{\beta_{tsj}}), \quad \beta_{tsj}|\gamma_{tsj} = 0 \sim \delta_0(\beta_{tsj})$$
$$\gamma_{tsj} \sim Bern((\gamma_{tsj})^{(1)}).$$

Note that now

$$(\beta_{tsj})^{(1)} = \mu_{\beta,tsj}(\gamma_{tsj})^{(1)} \tag{8.5.39}$$

$$(\beta_{tsj})^{(2)} = (\sigma^2_{\beta_{tsj}} + \mu^2_{\beta_{tsj}})(\gamma_{tsj})^{(1)}. \tag{8.5.40}$$

$$\log \tilde{q}(w_t) = \mathbb{E}_{\tilde{q}(-w_t)}\Big[\sum_s \sum_j \log p(\beta_{tsj}|w_t, \gamma_{tsj}) + \log p(w_t|a_w, b_w)\Big] + cst \tag{8.5.41}$$

$$\begin{aligned}\log \tilde{q}(w_t) \propto &\mathbb{E}_{\tilde{q}(-w_t)}\Big[\sum_s \sum_j \gamma_{tsj}\Big(-\frac{1}{2}\log w_t - \frac{1}{2}\beta^2_{tsj} w_t^{-1}\Big) + (-a_w - 1)\log w - b_w w^{-1}\Big] \\ &- \Big(a_w + \frac{1}{2}\sum_s \sum_j (\gamma_{tsj})^{(1)} + 1\Big)\log w_t - \Big(b_w + \frac{1}{2}\sum_s \sum_j (\beta_{tsj})^{(2)}\Big),\end{aligned}$$

where $\mathbb{E}_{\tilde{q}(-w_t)}[\gamma_{tsj}\beta^2_{tsj}] = (\beta_{tsj})^{(2)}$ from the law of iterative expectations.

Under $\tilde{q}$ we have

$$w_t \sim IG(a^*_{w_t}, b^*_{w_t}), \tag{8.5.42}$$

with

$$a^*_{w_t} = a_w + \frac{1}{2}\sum_s \sum_j (\gamma_{tsj})^{(1)} \tag{8.5.43}$$

$$b^*_{w_t} = (b_w)^{(1)} + \frac{1}{2}\sum_s \sum_j (\beta_{tsj})^{(2)} \tag{8.5.44}$$

where

$$(w_t^{-1})^{(1)} = a^*_{w_t}/b^*_{w_t} \tag{8.5.45}$$

$$(\log w_t^{-1})^{(1)} = \Psi(a^*_{w_t}) - \log b^*_{w_t} \tag{8.5.46}$$

and $\Psi(.)$ is the digamma function.

$$\tilde{q}(\omega_{sj}) \propto \mathbb{E}_{\tilde{q}(-\omega)}\Big[\sum_t \gamma_{tsj} \log \omega_{sj} + \sum_t (1-\gamma_{tsj}) \log(1-\omega_{sj}) + \\ + (a_\omega - 1)\log \omega_{sj} + (b_w - 1)\log(1-\omega_{sj})\Big] \tag{8.5.47}$$

Under $\tilde{q}$ we have

$$\omega_{sj} \sim Beta(a^*_\omega, b^*_\omega)$$

where

$$a^*_\omega = a_\omega + \sum_t (\gamma_{tsj})^{(1)}$$
$$b^*_\omega = a_\omega + T - \sum_t (\gamma_{tsj})^{(1)}$$

with

$$(\omega_{sj})^{(1)} = a^*_\omega / \left(a^*_\omega + b^*_\omega\right)$$
$$(\log \omega_{sj})^{(1)} = \Psi(a^*_\omega) - \Psi(a^*_\omega + b^*_\omega)$$
$$(\log(1-\omega_{sj}))^{(1)} = \Psi(b^*_\omega) - \Psi(a^*_\omega + b^*_\omega),$$

where $\Psi(.)$ is the digamma function.

$$\log \tilde{q}(z_{ig}) \propto \mathbb{E}_{\tilde{q}(z_{ig})}\left[\log p(z_{ig}|\boldsymbol{x}_{i,.}, \boldsymbol{\upsilon}_g) + \sum_j \zeta_{ij} \sum_t \log p(y_{it}|\boldsymbol{x}_{i,.}, \boldsymbol{\beta}_{tj}, u_{i(t-1)j}, \boldsymbol{\rho}_{tj})\right] \tag{8.5.48}$$

The Bernoulli distribution for z_{ig} can be rearranged to

$$p(z_{ig}|\boldsymbol{x}_{i,.}, \boldsymbol{\upsilon}_g, \boldsymbol{\epsilon}) = \exp\big(z_{ig}\boldsymbol{\upsilon}_{g,\epsilon}^T\boldsymbol{x}_{i,.}\big)\sigma\left(-\boldsymbol{\upsilon}_{g,\epsilon}^T\boldsymbol{x}_{i,.}\right).$$

Thus,

$$\begin{aligned}\log\tilde{q}(z_{ig}) \propto \mathbb{E}_{\tilde{q}(-z_{ig})}\Bigg[\sum_j \zeta_{ij}\Bigg\{ -\frac{T}{2}\log(2\pi) - \sum_t \frac{1}{2}\log\sigma_{tj}^{-2} - \sum_t \frac{1}{2\sigma_{tj}^2}\big(u_{itj} - \sum_{k<t} u_{ikj}\rho_{tkj}\big)^2\Bigg\} + \\ + z_{ig}(\boldsymbol{\upsilon}_g^T\boldsymbol{x}_{i,.})\Bigg] \\ \propto \mathbb{E}_{\tilde{q}(-z_{ig})}\Bigg[z_{ig}(\boldsymbol{\upsilon}_g^T\boldsymbol{x}_{i,.}) + \sum_j \zeta_{ij}A_{ij}\Bigg],\end{aligned}$$

where

$$A_{ij} = -\frac{T}{2}\log(2\pi) - \sum_t \frac{1}{2}\log\sigma_{tj}^{-2} - \sum_t \frac{1}{2\sigma_{tj}^2}\big(u_{itj} - \sum_{k<t} u_{ikj}\rho_{tkj}\big)^2. \tag{8.5.49}$$

ζ_{ij} represents the combination of the latent z_{ig} variables, as we move along the tree. For example, in a four expert structure with 3 gates, z_{i1} will appear in ζ_{ij} for every j, where as z_{i3} will only appear twice. Thus the pattern which is proportional to z_{ig} can be defined as

$$\log\tilde{q}(z_{ig}) \propto \ z_{ig}\mathbb{E}_{\tilde{q}(-z_{ig})}\Bigg[(\boldsymbol{\upsilon}_{\boldsymbol{g}}^T\boldsymbol{x}_{i,.}) + \sum_{j\in\mathcal{E}_g^L}\zeta_{ij}^{\not g}A_{ij} - \sum_{j\in\mathcal{E}_g^R}\zeta_{ij}^{\not g}A_{ij}\Bigg],$$

where $\mathcal{E}_g^R$ and $\mathcal{E}_g^L$ denote the set of experts on the right-hand-side and left-hand-side of the gth gate respectively and

$$\zeta_{ij}^{\not g} = \prod_{l=1, l\neq g}^{G} z_{il}^{S^L(j,l)}(1-z_{il})^{S^R(j,l)}.$$

Taking the expectation and exponentiating

$$\tilde{q}(z_{ig}) \propto \ \exp\Bigg(z_{ig}\Bigg((\boldsymbol{\upsilon}_g)^{(1)T}\boldsymbol{x}_{i,.} + \sum_{j\in\mathcal{E}_g^L}(\zeta_{ij}^{\not g})^{(1)}(A_{ij})^{(1)} - \sum_{j\in\mathcal{E}_g^R}(\zeta_{ij}^{\not g})^{(1)}(A_{ij})^{(1)}\Bigg)\Bigg). \tag{8.5.50}$$

Setting

$$(C_{ig})^{(1)} =(\boldsymbol{v}_g)^{(1)T}\boldsymbol{x}_{i,\cdot} + \sum_{j\in\mathcal{E}_g^L}(\zeta_{ij}^{\not g})^{(1)}(A_{ij})^{(1)} - \sum_{j\in\mathcal{E}_g^R}(\zeta_{ij}^{\not g})^{(1)}(A_{ij})^{(1)},$$

with $(A_{ij})^{(1)}$ defined as

$$(A_{ij})^{(1)} = -\frac{T}{2}\log(2\pi) - \sum_t \frac{(\log\sigma_{tj}^{-2})^{(1)}}{2} - \sum_t \frac{(\sigma_{tj}^{-2})^{(1)}}{2}\Bigg((u_{itj})^{(2)} - 2(u_{itj})^{(1)}\sum_{k<t}(u_{ikj})^{(1)}(\rho_{tkj})^{(1)} + \\ + 2\sum_{k'\neq k}(u_{ikj})^{(1)}(\rho_{tkj})^{(1)}(u_{ik'j})^{(1)}(\rho_{tk'j})^{(1)} + \sum_{k<t}(u_{ikj})^{(2)}(\rho_{tkj})^{(2)}\Bigg) \tag{8.5.51}$$

and

$$\mathbb{E}_{\tilde{q}(-z_{ig})}[\zeta_{ij}^{\not g}] = \prod_{l=1,l\neq g}^{G}(z_{il})^{(1)S^L(j,l)}(1-(z_{il})^{(1)})^{S^R(j,l)}.$$

Adding the constant $\sigma(-(C_{ig})^{(1)})$ gives

$$\begin{aligned}\tilde{q}(z_{ig}) &\propto \exp\big(z_{ig}(C_{ig})^{(1)}\big)\sigma(-(C_{ig})^{(1)}) \\ &\propto \sigma((C_{ig})^{(1)})^{z_{ig}}(1-\sigma((C_{ig})^{(1)}))^{1-z_{ig}}.\end{aligned}$$

Thus under $\tilde{q}$,

$$z_{ig} \sim \text{Bernouli}\Big(\sigma((C_{ig})^{(1)})\Big), \tag{8.5.52}$$

with

$$(z_{ig})^{(1)} = \sigma((C_{ig})^{(1)}), \tag{8.5.53}$$

where $\sigma(\cdot)$ is the sigmoid function. Thus

$$\mathbb{E}_{\tilde{q}}[\zeta_{ij}] = \prod_{i=1}^{G}(z_{ig})^{(1)S^L(j,g)}(1-(z_{ig})^{(1)})^{S^R(j,g)}. \tag{8.5.54}$$

A slightly different approach is used for the joint update of $q(\boldsymbol{\upsilon_s}, \epsilon_s)$. The parameters in $p(z_{ig}|\boldsymbol{\upsilon}_g, \boldsymbol{x}_{i,\cdot})$ are in terms of the columns of the matrix $\boldsymbol{\upsilon}$ $(G \times p)$, we define the equation in terms of $(\boldsymbol{\upsilon}, \boldsymbol{\epsilon})$ and then make proportional to $(\boldsymbol{\upsilon}_s, \epsilon_s)$. This enables us to obtain the update in terms of the rows.

$$
\begin{aligned}
p(\boldsymbol{\upsilon}, \boldsymbol{\epsilon}) \propto & \left\{ \prod_i \prod_g \sigma(\boldsymbol{\upsilon}_{g,\epsilon}^T \boldsymbol{x}_{i,\cdot})^{z_{ig}} (1 - \sigma(\boldsymbol{\upsilon}_{g,\epsilon}^T \boldsymbol{x}_{i,\cdot}))^{1-z_{ig}} \right\} \times \\
& \left\{ \prod_s \left[\prod_g (2\pi)^{-1/2} (d)^{-1/2} \exp\left(-\frac{1}{2d} \upsilon_{gs}^2 \right) \right]^{\epsilon_s} \times \delta_0(\boldsymbol{\upsilon}_s)^{1-\epsilon_s} \kappa^{\epsilon_s} (1-\kappa)^{\epsilon_s} \right\}. \qquad (8.5.55)
\end{aligned}
$$

Rearranging the sigmoid function

$$
\begin{aligned}
p(\boldsymbol{\upsilon}, \boldsymbol{\epsilon}) \propto & \left\{ \prod_g \prod_i \exp\big(z_{ig} \boldsymbol{\upsilon}_{g,\epsilon}^T \boldsymbol{x}_{i,\cdot}\big) (\sigma(-\boldsymbol{\upsilon}_{g,\epsilon}^T \boldsymbol{x}_{i,\cdot}))^{1-z_{ig}} \right\} \times \qquad (8.5.56) \\
& \left\{ \prod_s \left[\prod_g (2\pi)^{-1/2} (d)^{-1/2} \exp\left(-\frac{1}{2d} \upsilon_{gs}^2 \right) \right]^{\epsilon_s} \times \delta_0(\boldsymbol{\upsilon}_s)^{1-\epsilon_s} \kappa^{\epsilon_s} (1-\kappa)^{\epsilon_s} \right\}.
\end{aligned}
$$

Introduce the approximate lower bound

$$
\begin{aligned}
p(\boldsymbol{\upsilon}, \boldsymbol{\epsilon}) \geq & \left\{ \prod_g \prod_i \exp\big(z_{ig} \boldsymbol{\upsilon}_{g,\epsilon}^T \boldsymbol{x}_{i,\cdot}\big) \exp\left(-\frac{1}{2} \boldsymbol{\upsilon}_{g,\epsilon}^T \boldsymbol{x}_{i,\cdot} - \lambda_*(\psi_{ig}) (\boldsymbol{\upsilon}_{g,\epsilon}^T \boldsymbol{x}_{i,\cdot})^2 \right) \right\} \\
& \left\{ \prod_s \left[\prod_g (2\pi)^{-1/2} (d)^{-1/2} \exp\left(-\frac{1}{2d} \upsilon_{gs}^2 \right) \right]^{\epsilon_s} \times \right. \\
& \left. \delta_0(\boldsymbol{\upsilon}_s)^{1-\epsilon_s} \kappa^{\epsilon_s} (1-\kappa)^{\epsilon_s} \right\} \qquad (8.5.57)
\end{aligned}
$$

and taking the log

$$
\begin{aligned}
\log p(\boldsymbol{\upsilon}, \boldsymbol{\epsilon}) \geq & \sum_g \sum_i \left\{ z_{ig} \sum_s \upsilon_{gs} \epsilon_s x_{is} - \frac{1}{2} \sum_s \upsilon_{gs} \epsilon_s x_{is} - \lambda_*(\psi_{ig}) \big(\sum_s \upsilon_{gs} \epsilon_s x_{is} \big)^2 \right\} + \\
& + \sum_s \left\{ \epsilon_s \left[\sum_g \Big(-\frac{1}{2} \log 2\pi - \frac{\log d}{2} - \frac{\upsilon_{gs}^2}{2d} \Big) + \log(\kappa) \right] + \right. \\
& \left. + (1 - \epsilon_s) \Big((1 - \log \kappa) + \delta_0(\boldsymbol{\upsilon}_s) \Big) \right\}. \qquad (8.5.58)
\end{aligned}
$$

Taking proportionality with respect to $(\boldsymbol{\upsilon}_s, \epsilon_s)$

$$\log \tilde{q}(\boldsymbol{\upsilon}_s, \epsilon_s) \propto \mathbb{E}_{\tilde{q}(-\upsilon_s, -\epsilon_s)}\Bigg[\epsilon_s\Bigg\{\sum_g\sum_i z_{ig}\upsilon_{gs}x_{is} - \frac{1}{2}\sum_g\sum_i \upsilon_{ig}x_{gs} - \sum_g\sum_i \lambda_*(\psi_{ig})\Big(\upsilon_{gs}^2x_{is}^2 + \\ + 2\upsilon_{gs}x_{is}\sum_{l\neq s}\upsilon_{gl}x_{il}\Big)\Bigg\} - \sum_g \epsilon_s\frac{\upsilon_{gs}^2}{2d} + \epsilon_s\Bigg(-\frac{G}{2}\log(2\pi) - \frac{G\log(d)}{2} + \log\kappa\Bigg) + \\ + (1-\epsilon_s)\Big((1-\log\kappa) + \delta_0(\boldsymbol{\upsilon}_s)\Big)\Bigg], \tag{8.5.59}$$

and rearranging gives

$$\log \tilde{q}(\boldsymbol{\upsilon}_s, \epsilon_s) \propto -\frac{\epsilon_s}{2}\Bigg(\sum_g \upsilon_{gs}^2\Big(2\sum_i \lambda_*(\psi_{ig})x_{is}^2 + \frac{1}{(d)^{(1)}}\Big) - \\ \sum_g 2\Big(\upsilon_{gs}\sum_i (z_{ig})^{(1)}x_{is} - \upsilon_{gs}\sum_i x_{is} - \upsilon_{gs}\sum_n \lambda_*(\psi_{ig})x_{is}\sum_{l\neq s}\upsilon_{gl}x_{il}\Big)\Bigg) + \\ + \epsilon_s\Big(-\frac{G}{2}\log(2\pi) + \frac{G(\log d^{-1})^{(1)}}{2} + (\log\kappa)^{(1)}\Big) + \\ + (1-\epsilon_s)(\log(1-\kappa))^{(1)} + (1-\epsilon_s)\delta_0(\boldsymbol{\upsilon}_s). \tag{8.5.60}$$

Adding

$$\epsilon_s\sum_g \frac{1}{2}\log\Big(\sigma_{\upsilon_{gs}}^2\Big) - \epsilon_s\sum_g \frac{1}{2}\log\Big(\sigma_{\upsilon_{gs}}^2\Big) \tag{8.5.61}$$

and completing the square to define

$$\mu_{\upsilon_{gs}} = \sigma_{\upsilon_{gs}}^2\Bigg[\sum_i x_{is}\Bigg((z_{ig})^{(1)} - 1 - \lambda_*(\psi_{ig})\sum_{l\neq s}(\upsilon_{gl})^{(1)}x_{il}\Bigg)\Bigg] \tag{8.5.62}$$

$$\sigma_{\upsilon_{gs}}^2 = \Bigg(2\sum_i \lambda_*(\psi_{ig})x_{is}^2 + (d^{-1})^{(1)}\Bigg)^{-1}, \tag{8.5.63}$$

gives

$$\begin{aligned}\log \tilde{q}(\boldsymbol{\upsilon}_s, \epsilon_s) \geq & \epsilon_s \left[\sum_g \left\{ -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma^2_{\upsilon_{gs}}\right) - \frac{1}{2\sigma^2_{\upsilon_{gs}}}(\upsilon_{gs} - \mu_{\upsilon_{gs}})^2 + \frac{\mu^2_{\upsilon_{gs}}}{2\sigma^2_{\upsilon_{gs}}} \right\} \right]^{\epsilon_s} + \\ & \epsilon_s \left[(\log\kappa)^{(1)} + \frac{G}{2}(\log d^{-1})^{(1)} + \sum_g \frac{1}{2}\log\left(\sigma^2_{\upsilon_{gs}}\right) \right] + \\ & + (1-\epsilon_s)\Big(\delta_0(\boldsymbol{\upsilon}_s) + (\log(1-\kappa))^{(1)}\Big). \end{aligned} \tag{8.5.64}$$

Exponentiating

$$\begin{aligned}\tilde{q}(\boldsymbol{\upsilon}_s, \epsilon_s) \geq & \left[\prod_{g=1}^{G} \frac{1}{(2\pi\sigma^2_{\upsilon_{gs}})^{1/2}} \exp\left(-\frac{1}{2\sigma^2_{\upsilon_{gs}}}(\upsilon_{gs} - \mu_{\upsilon_{gs}})^2 \right) \right]^{\epsilon_s} \delta_0(\boldsymbol{\upsilon}_s)^{1-\epsilon_s} \\ & \left[\exp\left(\sum_g \left\{ \frac{\mu_{\upsilon_{gs}}}{2\sigma^2_{\upsilon_{gs}}} + \frac{1}{2}\log\sigma^2_{\upsilon_{gs}} + \frac{1}{2}(\log d^{-1})^{(1)} \right\} + (\log\kappa)^{(1)} \right) \right]^{\epsilon_s} \\ & \exp\Big((\log(1-\kappa))^{(1)}\Big)^{1-\epsilon_s} \end{aligned} \tag{8.5.65}$$

and normalising

$$(\epsilon_s)^{(1)} = \left[1 + \exp\left(-\sum_g \left\{ \frac{\mu_{\upsilon_{gs}}}{2\sigma^2_{\upsilon_{gs}}} + \frac{1}{2}\log\sigma^2_{\upsilon_{gs}} \right\} - \frac{G}{2}(\log d^{-1})^{(1)} - (\log\kappa)^{(1)} + (\log(1-\kappa))^{(1)} \right) \right]^{-1}. \tag{8.5.66}$$

We have under $\tilde{q}$

$$\boldsymbol{\upsilon}_s|\epsilon_s = 1 \sim \mathcal{N}_G(\boldsymbol{\mu}_{\upsilon_s}, \Sigma_{\upsilon_s}), \quad \boldsymbol{\upsilon}_s|\epsilon_s = 0 \sim \delta_0(\boldsymbol{\upsilon}_s) \tag{8.5.67}$$

$$\boldsymbol{\upsilon}_s \sim Bern((\epsilon_s)^{(1)}), \tag{8.5.68}$$

where $\boldsymbol{\mu}_{\upsilon_s}$ is a vector with the ith entry equal to $\mu_{\upsilon_{gs}}$ and Σ_{υ_s} is a diagonal matrix with the

(i,i)th entry equal to $\sigma^2_{\upsilon_{gs}}$. Note that now

$$(\boldsymbol{\upsilon}_s)^{(1)} = \boldsymbol{\mu}_{\upsilon_s}(\epsilon_s)^{(1)} \tag{8.5.69}$$

$$(\boldsymbol{\upsilon}_s)^{(2)} = (\Sigma_{\upsilon_s} + \boldsymbol{\mu}^2_{\upsilon_s})(\epsilon_s)^{(1)} \tag{8.5.70}$$

$$(\upsilon_{gs})^{(1)} = \mu_{\upsilon_{gs}}(\epsilon_s)^{(1)} \tag{8.5.71}$$

$$(\upsilon_{gs})^{(2)} = (\sigma_{\upsilon_{gs}} + \mu^2_{\upsilon_{gs}})(\epsilon_s)^{(1)}. \tag{8.5.72}$$

Vector operation updates where $\{(\mathbf{z}_g)^{(1)}, \lambda_*(\psi_g)\} \in \mathbb{R}^n$

$$\mu_{\upsilon_{gs}} = \sigma^2_{\upsilon_{gs}}\left[\boldsymbol{x}^T_{\cdot,s}\left((\mathbf{z}_g)^{(1)} - \mathbf{1}_n - \lambda_*(\boldsymbol{\psi}_g)\sum_{l\neq s}(\upsilon_{il})^{(1)}x_{il}\right)\right]$$

$$\sigma^2_{\upsilon_{gs}} = \left(2\lambda_*(\boldsymbol{\psi}_g)^T\boldsymbol{x}^2_{\cdot,s} + (d^{-1})^{(1)}\right)^{-1}.$$

$$\log\tilde{q}(d) \propto \mathbb{E}_{\tilde{q}(-d)}\Big[\sum_s \epsilon_s\Big(\sum_g\Big\{-\frac{1}{2}\log d - \frac{\upsilon^2_{gs}}{2}d^{-1}\Big\}\Big) + (-a_d - 1)\log d - b_d d^{-1}\Big]$$

$$\propto \Big(-a_d - 1 - \frac{G}{2}\sum_s(\epsilon_s)^{(1)}\Big)\log d - \Big(b_d + \mathbb{E}_{(-d)}\Big[\frac{1}{2}\sum_s\epsilon_s\sum_g\upsilon^2_{gs}\Big]\Big)d^{-1}.$$

Using

$$\mathbb{E}_{\tilde{q}(-d)}\Big[\sum_s\epsilon_s\sum_i\upsilon^2_{gs}\Big] = \sum_s\sum_g(\upsilon_{gs})^{(2)} \tag{8.5.73}$$

under $\tilde{q}$,

$$d \sim Inv - Gamma(a^*_d, b^*_d) \tag{8.5.74}$$

where

$$a_d^* = a_d + \frac{G}{2}\sum_s (\epsilon_s)^{(1)}$$
$$b_d^* = (b_d)^{(1)} + \frac{\sum_s \sum_g (\upsilon_{gs})^{(2)}}{2}$$

where

$$(d^{-1})^{(1)} = a_d^*/b_d^*$$
$$(\log d^{-1})^{(1)} = \Psi(a_d^*) - \log(b_d^*)$$

$$\log \tilde{q}(\sigma_{tj}^2) \propto \mathbb{E}_{\tilde{q}}\Big[\sum_j \sum_i \zeta_{ij} \sum_t \log p(y_{it}|\boldsymbol{x}_{i,.}, u_{i(t-1)j}, \boldsymbol{\beta}_{tj}, \boldsymbol{\rho}_{tj}, \sigma_{tj}^2) + \sum_t \sum_j \log p(\sigma_{tj}^2|\tau,\nu) + \\ + \sum_{k<t} \log p(\rho_{tkj}|\sigma_{tj}^2, \tau, \eta_{tkj})\Big] \tag{8.5.75}$$

$$\log \tilde{q}(\sigma_{tj}^2) \propto \mathbb{E}_{\tilde{q}(-\sigma_{tj}^2)}\Bigg[\sum_i \frac{\zeta_{ij}}{2}\log \sigma_{tj}^{-2} - \sum_i \left\{\frac{\zeta_{ij}}{2}\sigma_{tj}^{-2}\left(u_{itj} - \sum_{k>t} u_{ikj}\rho_{tkj}\right)^2\right\} + \\ + \Big(\frac{\nu - T + t}{2} + 1\Big)\log \sigma_{tj}^2 - \frac{\tau}{2}\sigma_{tj}^{-2} + \sum_{k<t}\left\{\eta_{tkj}\Big(\frac{1}{2}\log \sigma_{tj}^{-2} - \frac{\tau}{2}\rho_{tkj}^2\sigma_{tkj}^{-2}\Big)\right\}\Bigg] \tag{8.5.76}$$

We can define

$$\mathbb{E}_{\tilde{q}}\left[(u_{itj} - \sum_{k<t} u_{ikj}\rho_{tkj})^2\right] = (u_{itj})^{(2)} + \sum_{k<t}(u_{ikj})^{(2)}(\rho_{tkj})^{(2)} + \\ + 2\sum_{k'\neq k}(u_{ikj})^{(1)}(\rho_{tkj})^{(1)}(u_{ik'j})^{(1)}(\rho_{tk'j})^{(1)} - 2\sum_{k<t}(u_{itj})^{(1)}(u_{ikj})^{(1)}\rho_{tkj}^{(1)}.$$

Therefore

$$\begin{aligned}\log \tilde{q}(\sigma_{tj}^2) \propto & -\sigma_{tj}^{-2}\sum_i\left\{\frac{(\zeta_{ij})^{(1)}}{2}\left((u_{itj})^{(2)}+\sum_{k<t}(u_{ikj})^{(2)}(\rho_{tkj})^{(2)}+\right.\right.\\ & \left.\left.+2\sum_{k'\neq k}(u_{ikj})^{(1)}(\rho_{tkj})^{(1)}(u_{ik'j})(\rho_{tk'j})^{(1)}-2\sum_{k<t}(u_{itj})^{(1)}(u_{ikj})^{(1)}(\rho_{tkj})^{(1)}\right)\right\}+\\ & +\log\sigma_{tj}^{-2}\sum_i\frac{(\zeta)_{ij}^{(1)}}{2}+\left(\frac{\nu-T+t}{2}+1\right)\log\sigma_{tj}^{-2}+\\ & -\frac{(\tau)^{(1)}}{2}\sigma_{tj}^{-2}+\sum_{k<t}\frac{(\eta_{tkj})^{(1)}}{2}\log\sigma_{tj}^{-2}-\sum_{k<t}\frac{(\tau)^{(1)}}{2}(\rho_{tkj})^{(2)}\sigma_{tj}^{-2},\end{aligned}$$

which is the inverse gamma kernal

$$\sigma_{tj}^2 \sim Inv-Gamma(a^*_{\sigma^2,tj}, b^*_{\sigma^2,tj})$$

where

$$a^*_{\sigma^2,tj} = \sum_i\frac{(\zeta_{ij})^{(1)}}{2}+\frac{\nu-T+t}{2}+\sum_{k<t}\frac{(\eta_{tkj})^{(1)}}{2} \tag{8.5.77}$$

$$\begin{aligned}b^*_{\sigma^2,tj} = & \sum_i\frac{(\zeta_{ij})^{(1)}}{2}\left((u_{itj})^{(2)}+\sum_{k<t}(u_{ikj})^{(2)}(\rho_{tkj})^{(2)}+2\sum_{k'\neq k}(u_{ikj})^{(1)}(\rho_{tkj})^{(1)}(u_{ik'j})^{(1)}(\rho_{tk'j})^{(1)}+\right.\\ & \left.-2\sum_{k<t}(u_{itj})^{(1)}(u_{ikj})^{(1)}(\rho_{tkj})^{(1)}\right)+\frac{(\tau)^{(1)}}{2}+\sum_{k<t}\frac{(\tau)^{(1)}}{2}(\rho_{tkj})^{(2)}\end{aligned} \tag{8.5.78}$$

with

$$(\sigma_{tj}^{-2})^{(1)} = \frac{a^*_{\sigma^2,tj}}{b^*_{\sigma^2,tj}}$$

$$(\log\sigma_{tj}^{-2}) = \Psi(a^*_{\sigma^2,tj})-\log b^*_{\sigma^2,tj}.$$

Updates from vector operations can be defined using Equations (8.5.26) to (8.5.28)

$$a^*_{\sigma^2,tj} = \sum_i \frac{(\zeta_{ij})^{(1)}}{2} + \frac{\nu - T + t}{2} + \sum_{k<t} \frac{(\eta_{tkj})^{(1)}}{2} \tag{8.5.79}$$

$$\begin{aligned} b^*_{\sigma^2,tj} =& \frac{(\tau)^{(1)}}{2} + \sum_{k<t} \frac{(\tau)^{(1)}}{2} (\rho_{tkj})^{(2)} + (\boldsymbol{\zeta}_j)^{(1)T} \Big(\frac{1}{2} (\boldsymbol{u}_{tj})^{(2)} + \sum_{k<t} (\boldsymbol{u}_{kj})^{(2)} \frac{(\rho_{tkj})^{(2)}}{2} \Big) + \\ & \sum_{k' \neq k} \Big((\boldsymbol{\zeta}_j)^{(1)} \odot (\boldsymbol{u}_{kj})^{(1)} \Big)^T (\boldsymbol{u}_{k'j})^{(1)} (\rho_{tkj})^{(1)} (\rho_{tk'j})^{(1)} + \\ & - \sum_{k<t} \Big((\boldsymbol{\zeta}_j)^{(1)} \odot (\boldsymbol{u}_{tj})^{(1)} \Big)^T (\boldsymbol{u}_{kj})^{(1)} (\rho_{tkj})^{(1)} \Big). \end{aligned} \tag{8.5.80}$$

$$\log \tilde{q}(\rho_{tkj}, \eta_{tkj}) \propto \mathbb{E}_{-(\rho_{tkj}, \eta_{tkj})} \Bigg[\sum_i \zeta_{ij} \log p(y_{it} | \boldsymbol{x}_{i,.}, u_{i(t-1)j}, \boldsymbol{\beta}_{tj}, \boldsymbol{\rho}_{tj}, \sigma^2_{tj}) + \log p(\rho_{tkj} | \sigma^2_{tj}, \tau, \eta_{tkj}) + \\ + \log p(\eta_{tkj} | \lambda_j) \Bigg]$$

$$\begin{aligned} \log \tilde{q}(\rho_{tkj}, \eta_{tkj}) \propto \mathbb{E}_{-(\rho_{tkj}, \eta_{tkj})} \Bigg[& \sum_i \zeta_{ij} \Bigg\{ - \frac{\sigma^{-2}_{tj}}{2} \Big(u_{itj} - \sum_{h<t, h \neq k} u_{ihj} \eta_{hj} \rho_{thj} - u_{ikj} \rho_{tkj} \eta_{tkj} \Big)^2 \Bigg\} + \\ & + \eta_{tkj} \Big(- \frac{1}{2} \log 2\pi + \frac{1}{2} \log \tau + \frac{1}{2} \log \sigma^{-2}_{tj} - \frac{1}{2} \tau \sigma^{-2}_{tj} \rho^2_{tkj} \Big) + \\ & + (1 - \eta_{tkj}) \log(1 - \lambda_j) + (1 - \eta_{tkj}) \delta_0(\rho_{tkj}) \Bigg]. \end{aligned}$$

Using the Equations (8.5.23) and (8.5.24), rearranging, taking the expectation and adding η_{tkj}

to the first component gives

$$\begin{aligned}\log\tilde{q}(\rho_{tkj},\eta_{tkj}) \propto & -\frac{1}{2}\eta_{tkj}(\sigma_{tj}^{-2})^{(1)}\Bigg[\rho_{tkj}^2\Big((\tau)^{(1)}+\sum_i(\zeta_{ij})^{(1)}(u_{ikj})^{(2)}\Big)+\\ & -2\rho_{tkj}\Big(\sum_i(\zeta_{ij})^{(1)}(u_{ikj})^{(1)}\Big((u_{itj})^{(1)}-\sum_{h<t,h\neq k}(u_{ihj})^{(1)}(\rho_{thj})^{(1)}\Big)\Big)\Bigg]+\\ & +\eta_{tkj}\Bigg[-\frac{1}{2}\log 2\pi+\frac{1}{2}(\log\tau)^{(1)}+\frac{1}{2}(\log\sigma_{tj}^{-2})^{(1)}+(\log\lambda_j)^{(1)}\Bigg]+\\ & +(1-\eta_{tkj})\Bigg[\log(1-\lambda_j)^{(1)}+\delta_0(\rho_{tkj})\Bigg].\end{aligned}$$

Setting

$$\mu_{\rho_{tkj}}=\frac{\Bigg[\sum_i(\zeta_{ij})^{(1)}(u_{ikj})^{(1)}\Bigg((u_{itj})^{(1)}-\sum_{h<t,h\neq k}(u_{ihj})^{(1)}(\rho_{thj})^{(1)}\Bigg)\Bigg]}{(\tau)^{(1)}+\sum_i(\zeta_{ij})^{(1)}(u_{ikj})^{(2)}} \tag{8.5.81}$$

$$\sigma^2_{\rho_{tkj}}=\Bigg[(\sigma_{tj}^{-2})^{(1)}\Bigg((\tau)^{(1)}+\sum_i(\zeta_{ij})^{(1)}(u_{ikj})^{(2)}\Bigg)\Bigg]^{-1}. \tag{8.5.82}$$

The joint $\tilde{q}$ density is proportional to

$$\begin{aligned}\tilde{q}(\rho_{tkj},\eta_{tkj}) \propto & \Bigg[(2\pi\sigma^2_{\rho_{tkj}})^{-1/2}\exp\Bigg\{-\frac{1}{2\sigma^2_{\rho_{tkj}}}(\rho_{tkj}-\mu_{\rho_{tkj}})^2\Bigg\}\Bigg]^{\eta_{tkj}}\times\Bigg[\delta_0(\rho_{tk})\Bigg]^{1-\eta_{tkj}}\times\\ & \Bigg[\Bigg\{\exp\big((\log\tau)^{(1)}+(\log\sigma_{tj}^{-2})^{(1)}\big)\sigma^2_{\rho_{tk}}\Bigg\}^{\frac{1}{2}}\exp\Bigg\{\frac{\mu^2_{\rho_{tkj}}}{2\sigma^2_{\rho_{tkj}}}\Bigg\}\exp\Bigg\{(\log\lambda_j)^{(1)}\Bigg\}\Bigg]^{\eta_{tkj}}\times\\ & \Bigg[\exp\Bigg\{(\log(1-\lambda_j))^{(1)}\Bigg\}\Bigg]^{1-\eta_{tkj}},\end{aligned}$$

and thus by calling

$$(\eta_{tkj})^{(1)}=\Bigg[1+\sqrt{\sigma^{-2}_{\rho_{tkj}}}\exp\Bigg\{(\log(1-\lambda_j))^{(1)}-\frac{(\log\tau)^{(1)}}{2}-\frac{(\log\sigma_{tj}^{-2})^{(1)}}{2}-(\log\lambda_j)^{(1)}-\frac{\mu^2_{\rho_{tkj}}}{2\sigma^2_{\rho_{tkj}}}\Bigg\}\Bigg]^{-1}$$

we have under $\tilde{q}$

$$\rho_{tkj}|\eta_{tkj}=1\sim\mathcal{N}(\mu_{\rho_{tkj}},\sigma^2_{\rho_{tkj}}),\quad \rho_{tkj}|\eta_{tkj}=0,\sim\delta_0(\rho_{tkj})$$
$$\eta_{tkj}\sim Bern((\eta_{tkj})^{(1)}).$$

Note that now

$$\mathbb{E}_{\tilde{q}}[\rho_{tkj}]=(\rho_{tkj})^{(1)}=\mu_{\rho_{tkj}}(\eta_{tkj})^{(1)}$$
$$\mathbb{E}_{\tilde{q}}[\eta_{tkj}\rho_{tkj}]=\mu_{\rho_{tkj}}(\eta_{tkj})^{(1)}$$
$$(\rho_{tkj})^{(2)}=(\mu^2_{\rho_{tkj}}+\sigma^2_{\rho_{tkj}})(\eta_{tkj})^{(1)}.$$

The updates can be performed by Vector operations by

$$\mu_{\rho_{tkj}}=\frac{\left[\left((\boldsymbol{\zeta}_j)^{(1)}\odot(\boldsymbol{u}_{kj})^{(1)}\right)^T\left((\boldsymbol{u}_{tj})^{(1)}-\sum_{h<t,h\neq k}(\boldsymbol{u}_{hj})^{(1)}(\rho_{thj})^{(1)}\right)\right]}{(\tau)^{(1)}+(\boldsymbol{\zeta}_j)^{(1)T}(\boldsymbol{u}_{kj})^{(2)}} \tag{8.5.83}$$

$$\sigma^2_{\rho_{tkj}}=\left[(\sigma^{-2}_{tj})^{(1)}\left((\tau)^{(1)}+(\boldsymbol{\zeta}_j)^{(1)T}(\boldsymbol{u}_{kj})^{(2)}\right)\right]^{-1}. \tag{8.5.84}$$

$$\log\tilde{q}(\tau)=\mathbb{E}_{\tilde{q}(-\tau)}\left[\sum_t\sum_j\log p(\sigma^2_{tj}|\tau,\nu)+\sum_t\sum_{k<t}\sum_j\log p(\rho_{tkj}|\sigma^2_{tj},\tau,\eta_{tkj})+\log p(\tau)\right]+cst \tag{8.5.85}$$

$$\log\tilde{q}(\tau)\propto\mathbb{E}_{\tilde{q}(-\tau)}\left[\sum_t\sum_j\left\{\frac{\nu-T+t}{2}\log\tau-\frac{\tau\sigma^{-2}_{tj}}{2}\right\}+\sum_t\sum_{k<t}\sum_j\left\{\frac{\eta_{tkj}}{2}\log\tau-\tau\frac{\eta_{tkj}\rho^2_{tkj}}{2\sigma^2_{tj}}\right\}\right.$$
$$\left.+(a_\tau-1)\log\tau-b_\tau\tau\right]$$

Rearranging and taking the expectation gives

$$\log \tilde{q}(\tau) \propto -\tau\Big(\sum_j \sum_t \frac{(\sigma_{tj}^{-2})^{(1)}}{2} + \sum_j \sum_t \sum_{k<t} \frac{(\rho_{tkj})^{(2)}(\sigma_{tj}^2)^{(1)}}{2} + b_\tau\Big) + \\ + \log\tau\Big(J\sum_t \frac{(\nu - T + t)}{2} + \sum_t \sum_{k<t} \sum_j \frac{(\eta_{tkj})^{(1)}}{2} + a_\tau\Big),$$

where we use $\mathbb{E}_{\tilde{q}}[\eta_{tkj}\rho_{tkj}^2] = (\rho_{tkj})^{(2)}$. Thus, since $\sum_t t = \frac{T(T+1)}{2}$, under $\tilde{q}$

$$\tau \sim Gamma\,(a_\tau^*, b_\tau^*)$$

with parameters

$$a_\tau^* = a_\tau + \frac{JT(\nu - T/2 + 1/2)}{2} + \sum_j \sum_t \sum_{k<t} \frac{(\eta_{tkj})^{(1)}}{2} \tag{8.5.86}$$

$$b_\tau^* = b_\tau + \frac{1}{2}\sum_j \sum_t (\sigma_{tj}^{-2})^{(1)} \left[1 + \sum_{k<t} (\rho_{tkj})^{(2)}\right] \tag{8.5.87}$$

where

$$(\tau)^{(1)} = a_\tau^*/b_\tau^* \tag{8.5.88}$$

$$(\log\tau)^{(1)} = \Psi(a_\tau^*) - \log b_\tau^*. \tag{8.5.89}$$

$$\log \tilde{q}(\lambda_j) \propto \mathbb{E}_{\tilde{q}(-\lambda_j)}\Big[\sum_t \sum_{k<t} \log p(\eta_{tkj}|\lambda_j) + \log p(\lambda_j)\Big] \tag{8.5.90}$$

Expanding gives

$$\log \tilde{q}(\lambda_j) \propto \mathbb{E}_{\tilde{q}(-\lambda_j)}\Big[\sum_t \sum_{k<t} \Big\{\eta_{tkj} \log \lambda_j + (1-\eta_{tkj}) \log(1-\lambda_j)\Big\} + (a_\lambda - 1)\log \lambda_j + (b_\lambda - 1)\log(1-\lambda_j)\Big]$$

under $\tilde{q}$,

$$\lambda_j \sim Beta(a^*_{\lambda_j}, b^*_{\lambda_j}),$$

where

$$a^*_{\lambda_j} = \sum_t \sum_{k<t} (\eta_{tkj})^{(1)} + a_\lambda$$
$$b^*_{\lambda_j} = \sum_t \sum_{k<t} (1-\eta_{tkj})^{(1)} + b_\lambda$$

and

$$(\lambda_j)^{(1)} = a^*_{\lambda_j} / \left(a^*_{\lambda_j} + b^*_{\lambda_j}\right)$$
$$(\log \lambda_j)^{(1)} = \Psi(a^*_{\lambda_j}) - \Psi(a^*_{\lambda_j} + b^*_{\lambda_j})$$
$$(\log(1-\lambda_j))^{(1)} = \Psi(b^*_{\lambda_j}) - \Psi(a^*_{\lambda_j} + b^*_{\lambda_j}).$$

$$\log \tilde{q}(\kappa) \propto \mathbb{E}_{\tilde{q}(-\kappa)}\Big[\sum_s \epsilon_s \log \kappa + \sum_s (1-\epsilon_s)\log(1-\kappa) + (a_\kappa - 1)\log \kappa + b_\kappa - 1(1-\kappa)\Big] \quad (8.5.91)$$

under $\tilde{q}$

$$\kappa \sim Beta(a^*_\kappa, b^*_\kappa)$$

with

$$a_\kappa^* = \sum_s (\epsilon_s)^{(1)} + a_\kappa \tag{8.5.92}$$

$$b_\kappa^* = \sum_s (1 - (\epsilon_s)^{(1)}) + b_\kappa \tag{8.5.93}$$

where

$$\begin{aligned}
(\kappa)^{(1)} &= a_\kappa^* / \left(a_\kappa^* + b_\kappa^*\right) \\
(\log \kappa)^{(1)} &= \Psi(a_\kappa^*) - \Psi(a_\kappa^* + b_\kappa^*) \\
(\log(1-\kappa))^{(1)} &= \Psi(b_\kappa^*) - \Psi(a_\kappa^* + b_\kappa^*).
\end{aligned}$$

$$\log \tilde{q}(b_w) = \mathbb{E}_{\tilde{q}(-b_w)}\left[\sum_{t=1}^{T} \log p(w_t|a_w, b_w) + \log p(b_w|a_{b_w}, b_{b_w})\right] \tag{8.5.94}$$

$$\begin{aligned}
\log \tilde{q}(b_w) &= \mathbb{E}_{\tilde{q}(-b_w)}\left[\sum_t \left\{a_w \log b_w - b_w w_t^{-1}\right\} + (a_{b_w} - 1)\log b_w - b_{b_w} b_w\right] \\
&= (T a_w \log b_w - b_w \sum_t (w_t)^{(-1)} + (a_{b_w} - 1)\log b_w - b_{b_w} b_w \\
&= \log b_w (T a_w + a_{b_w} - 1) - b_w (\sum_t (w_t)^{(-1)} + b_{b_w}).
\end{aligned} \tag{8.5.95}$$

Thus under $\tilde{q}$,

$$b_w \sim Gamma(a_{b_w}^*, b_{b_w}^*),$$

with parameters

$$a^*_{b_w} = Ta_w + a_{b_w} \tag{8.5.96}$$

$$b^*_{b_w} = \sum_t (w_t)^{(-1)} + b_{b_w}, \tag{8.5.97}$$

where

$$(b_w)^{(1)} = a^*_{b_w}/b^*_{b_w} \tag{8.5.98}$$

$$(\log b_w)^{(1)} = \Psi(a^*_{b_w}) - \log b^*_{b_w} \tag{8.5.99}$$

$$\log \tilde{q}(b_d) = \mathbb{E}_{\tilde{q}(-b_d)}\Big[\log p(d|a_d, b_d) + \log p(b_d|a_{b_d}, b_{b_d})\Big] \tag{8.5.100}$$

$$\begin{aligned}\log \tilde{q}(b_d) &= \mathbb{E}_{\tilde{q}(-b_d)}\Big[\Big\{a_d \log b_d - b_d d^{-1}\Big\} + (a_{b_d} - 1)\log b_d - b_{b_d} b_d\Big] \\ &= a_d \log b_d - b_d (d)^{(-1)} + (a_{b_d} - 1)\log b_d - b_{b_d} b_d \\ &= \log b_d (a_d + a_{b_d} - 1) - b_d((d)^{(-1)} + b_{b_d}).\end{aligned} \tag{8.5.101}$$

Thus under $\tilde{q}$

$$b_d \sim Gamma(a^*_{b_d}, b^*_{b_d}),$$

with parameters

$$a^*_{b_d} = a_d + a_{b_d} \tag{8.5.102}$$

$$b^*_{b_d} = (d)^{(-1)} + b_{b_d}, \tag{8.5.103}$$

where

$$(b_d)^{(1)} = a^*_{b_d}/b^*_{b_d} \tag{8.5.104}$$

$$(\log b_d)^{(1)} = \Psi(a^*_{b_d}) - \log b^*_{b_d}. \tag{8.5.105}$$

A lower bound on the probability distribution of the latent indicator variable z_{ig},

$$\begin{aligned} p(z_{ig}|\boldsymbol{x}_{i,\cdot}, \boldsymbol{\upsilon}_g) &= \exp\big(z_{ig}\boldsymbol{\upsilon}_g^T\boldsymbol{x}_{i,\cdot}\big)\sigma(-\boldsymbol{\upsilon}_g^T\boldsymbol{x}_{i,\cdot}) \\ &\geq \exp\big(z_{ig}\boldsymbol{\upsilon}_g^T\boldsymbol{x}_{i,\cdot}\big)\sigma(\psi_{ig})\exp\left(\frac{-\boldsymbol{\upsilon}_g^T\boldsymbol{x}_{i,\cdot} - \psi_{ig}}{2} - \lambda_*(\psi_{ig})((\boldsymbol{\upsilon}_g^T\boldsymbol{x}_{i,\cdot})^2 - \psi_{ig}^2)\right) \end{aligned} \tag{8.5.106}$$

is used to achieve conjugacy for $\boldsymbol{\upsilon}_s$. For each gating node g, there is a separate variational parameter ψ_{ig} for each observation i, which can be optimised to yield the tightest bound. The optimisation of ψ_{ig} is achieved by maximising the lower bound on the marginal likelihood, which is now

$$\begin{aligned} \mathcal{L}(q) &\geq \mathcal{L}(\tilde{q}) \\ &= \mathbb{E}_{\tilde{q}(\boldsymbol{\vartheta})}[\log p(\mathbf{y}, \boldsymbol{\vartheta}_{-\mathbf{z}})h(\boldsymbol{\psi}, \boldsymbol{\upsilon})] - \mathbb{E}_{\tilde{q}(\boldsymbol{\vartheta})}[\log \tilde{q}(\boldsymbol{\vartheta})]. \end{aligned} \tag{8.5.107}$$

As the objective is to choose the parameters which maximise the function, and all updates other then ψ_{ig} have been determined, (8.5.107) can be optimised with respect to ψ_{ig}. Given that we will differentiate with respect to ψ_{ig}, the proportional expression is

$$\begin{aligned} \mathcal{L}(\tilde{q}) &\propto \mathbb{E}_{\tilde{q}(\vartheta)}[\log\sigma(\psi_{ig}) + (-\boldsymbol{\upsilon}_g^T\boldsymbol{x}_{i,\cdot} - \psi_{ig})/2 - \lambda_*(\psi_{ig})(\boldsymbol{\upsilon}_g^T\boldsymbol{x}_{i,\cdot}\boldsymbol{x}_{i,\cdot}^T\boldsymbol{\upsilon}_g - \psi_{ig}^2)] \\ &\propto \log\sigma(\psi_{ig}) - (\sum_s(\upsilon_{gs})^{(1)}x_{is} + \psi_{ig})/2 - \lambda_*(\psi_{ig})(\boldsymbol{x}_{i,\cdot}^T\mathbb{E}_{\tilde{q}(\upsilon)}[\boldsymbol{\upsilon}_g\boldsymbol{\upsilon}_g^T]\boldsymbol{x}_{i,\cdot} - \psi_{ig}^2), \end{aligned}$$

where

$$\boldsymbol{x}_{i,\cdot}^T\mathbb{E}_{\tilde{q}(\upsilon)}[\boldsymbol{\upsilon}_g\boldsymbol{\upsilon}_g^T]\boldsymbol{x}_{i,\cdot} = \sum_s x_{is}^2(\upsilon_{gs})^{(2)} + 2\sum_{s<s',s\neq s} x_{is}x_{is'}(\upsilon_{gs})^{(1)}(\upsilon_{gs'})^{(1)}. \tag{8.5.108}$$

Taking the derivative with respect to ψ_{ig}

$$\begin{aligned}\frac{d}{d\psi_{ig}}\sigma(\psi_{ig}) &= \frac{d}{d\psi_{ig}}(1+\exp(-\psi_{ig}))^{-1} \\ &= (1-\sigma(\psi_{ig}))\sigma(\psi_{ig})\end{aligned}$$

and

$$\tanh(x) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}, \tag{8.5.109}$$

so $\lambda_*(\psi_{ig})$

$$\begin{aligned}\lambda_*(\psi_{ig}) &= \frac{\tanh(\psi_{ig}/2)}{4\psi_{ig}} \\ &= \frac{\exp(\psi_{ig}/2)-\exp(-\psi_{ig}/2)}{4\psi_{ig}(\exp(\psi_{ig}/2)+\exp(-\psi_{ig}/2))} \\ 4\psi_{ig}\lambda_*(\psi_{ig}) &= \frac{1}{1+\exp(-\psi_{ig})} - \frac{\exp(-\psi_{ig})}{\exp(-\psi_{ig})+1} \\ &= 2\sigma(\psi_{ig})-1,\end{aligned}$$

and thus $2\psi_{ig}\lambda_*(\psi_{ig}) = \sigma(\psi_{ig}) - 1/2$.

$$\frac{d}{d\psi_{ig}}\mathcal{L}(\tilde{q}) = \frac{(1-\sigma(\psi_{ig}))\sigma(\psi_{ig})}{\sigma(\psi_{ig})} - \frac{1}{2} - \lambda'_*(\psi_{ig})(\boldsymbol{x}_{i,\cdot}^T\mathbb{E}_{\tilde{q}(v)}[\boldsymbol{v}_g\boldsymbol{v}_g^T]\boldsymbol{x}_{i,\cdot} - \psi_{ig}^2) + 2\psi_{ig}\lambda_*(\psi_{ig}),$$

setting this equal to 0 and rearranging for ψ_{ig}

$$\begin{aligned}0 &= 1-\sigma(\psi_{ig}) - \frac{1}{2} - \lambda'(\psi_{ig})(\boldsymbol{x}_{i,\cdot}^T\mathbb{E}_{\tilde{q}(v)}[\boldsymbol{v}_g\boldsymbol{v}_g^T]\boldsymbol{x}_{i,\cdot} - \psi_{ig}^2) + \sigma(\psi_{ig}) - 1/2 \\ &= \lambda'(\psi_{ig})(\boldsymbol{x}_{i,\cdot}^T\mathbb{E}_{\tilde{q}(v)}[\boldsymbol{v}_g\boldsymbol{v}_g^T]\boldsymbol{x}_{i,\cdot} - \psi_{ig}^2).\end{aligned}$$

As $\lambda'(\psi_{ig}) \neq 0$ we can divide each sides of the equation by this expression to get

$$\begin{aligned}\psi_{ig}^{2(\text{new})} =& \boldsymbol{x}_{i,\cdot}^T \mathbb{E}_{\tilde{q}(\upsilon)}[\boldsymbol{\upsilon}_g \boldsymbol{\upsilon}_g^T] \boldsymbol{x}_{i,\cdot} \\ =& \sum_s x_{is}^2 (\upsilon_{gs})^{(2)} + 2 \sum_{s<s', s\neq s} x_{is} x_{is'} (\upsilon_{gs})^{(1)} (\upsilon_{gs'})^{(1)}.\end{aligned}$$

8.5.3 ELBO

The ELBO is defined as

$$\begin{aligned}\mathcal{L}(\tilde{q}) &= \mathbb{E}_{\tilde{q}(\mathbf{U})}[\log p(\mathbf{y}, \mathbf{U}_{-\mathbf{Z}}) h(\psi, \boldsymbol{\upsilon}, \mathbf{X})] - \mathbb{E}_{\tilde{q}(\mathbf{U})}[\log \tilde{q}(\mathbf{U})] \\ &= \sum_i A(\boldsymbol{y}_i | \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2, \boldsymbol{\zeta}) + \sum_t \sum_s \sum_j B(\beta_{tsj}, \gamma_{tsj} | w_t, \omega_{sj}) + \\ &+ \sum_t \sum_{k<t} \sum_j B^*(\rho_{tkj}, \eta_{tkj} | \sigma_{tj}^2, \tau, \lambda_j) + \sum_i \sum_g C(z_{ig} | \boldsymbol{\upsilon}_g) + \sum_s D(\boldsymbol{\upsilon}_s, \epsilon_s | d, \kappa) + \\ &+ \sum_t \sum_j F(\sigma_{tj}^2 | \tau, \nu) + + \sum_s \sum_j G(\omega_{sj}) + \sum_t H(w_t | b_w) \\ &+ \sum_s I(\kappa) + J(d) + K(\lambda_j) + L(\tau) + M(b_w) + M^*(b_d).\end{aligned}$$

The functions are as follows:

$$\begin{aligned}A(\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2) =& \sum_i \sum_j \zeta_{ij} \sum_t \log p(y_{it} | \cdot) \\ =& \mathbb{E}_q \Bigg[\sum_i \sum_j \zeta_{ij} \sum_t \Bigg\{ -\frac{1}{2} \log(2\pi) + \frac{1}{2} (\log \sigma_{tj}^{-2}) + \\ & - \frac{1}{2\sigma_{tj}^2} \Bigg(y_{it} - \sum_s x_{is} \beta_{tsj} - \sum_{k<t} \big(y_{ik} - \sum_s x_{is} \beta_{ksj} \big) \rho_{tkj} \Bigg)^2 \Bigg\} \Bigg[\big]\end{aligned}$$

Taking the expectation,

$$A(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\rho},\boldsymbol{\sigma}^2) = \sum_i \sum_j \sum_t (\zeta_{ij})^{(1)} \Bigg\{ -\frac{1}{2}\log(2\pi) + \frac{1}{2}(\log \sigma_{tj}^{-2})^{(1)} - \frac{(\sigma_{tj}^{-2})^{(1)}}{2} \Bigg((u_{itj})^{(2)} + \\ + \sum_{k<t} (u_{ikj})^{(2)} (\rho_{tkj})^{(2)} - 2(u_{itj})^{(1)} \sum_{k<t} (u_{ikj})^{(1)} (\rho_{tkj})^{(1)} + \\ + \sum_{k'<k} (u_{ikj})^{(1)} (u_{ik'j})^{(1)} (\rho_{tkj})^{(1)} (\rho_{tk'j})^{(1)} \Bigg) \Bigg\},$$

and vectorising, we get

$$A(\boldsymbol{Y}|\cdot) = \sum_j \sum_t \Bigg\{ -\frac{\log(2\pi)}{2} (\boldsymbol{\zeta}_j)^{T(1)} \mathbf{1}_n + \frac{(\log \sigma_{tj}^{-2})^{(1)}}{2} (\boldsymbol{\zeta}_j)^{T(1)} \mathbf{1}_n - (\sigma_{tj}^{-2})^{(1)} \times \\ \Bigg((\boldsymbol{\zeta}_j)^{(1)T} \Bigg(\frac{1}{2} (\boldsymbol{u}_{tj})^{(2)} + \sum_{k<t} (\boldsymbol{u}_{kj})^{(2)} \frac{(\rho_{tkj})^{(2)}}{2} \Bigg) + \\ + \sum_{k' \neq k} \Big((\boldsymbol{\zeta}_j)^{(1)} \odot (\boldsymbol{u}_{kj})^{(1)} \Big)^T (\boldsymbol{u}_{k'j})^{(1)} (\rho_{tkj})^{(1)} (\rho_{tk'j})^{(1)} + \\ - \sum_{k<t} \Big((\boldsymbol{\zeta}_j)^{(1)} \odot (\boldsymbol{u}_{tj})^{(1)} \Big)^T (\boldsymbol{u}_{kj})^{(1)} (\rho_{tkj})^{(1)} \Bigg),$$

which is simplified by using the update σ_{tj}^2 Equation (8.5.80) to

$$A(\boldsymbol{Y}|\cdot) = \sum_j \sum_t \Bigg\{ -\frac{\log(2\pi)}{2} (\boldsymbol{\zeta}_j)^{T(1)} \mathbf{1}_n + \frac{(\log \sigma_{tj}^{-2})^{(1)}}{2} (\boldsymbol{\zeta}_j)^{T(1)} \mathbf{1}_n + \\ - (\sigma_{tj}^{-2})^{(1)} \Big(b^*_{\sigma^2_{tj}} - \frac{(\tau)^{(1)}}{2} - \frac{(\tau)^{(1)}}{2} \sum_{k<t} (\rho_{tkj})^{(2)} \Big).$$

$$B(\beta_{tsj}, \gamma_{tsj}|w_t, \omega_{sj}) = \mathbb{E}_{\tilde{q}}[\log p(\beta_{tsj}|\gamma_{tsj}, w_t)] + \mathbb{E}_{\tilde{q}}[\log p(\gamma_{tsj}|\omega_{sj})] - \mathbb{E}_{\tilde{q}}[\log \tilde{q}(\beta_{ts}, \gamma_{ts})]$$

$$B(\beta_{tsj},\gamma_{tsj}|w,\omega_{sj}) = \mathbb{E}_{\tilde{q}}\Bigg[\gamma_{tsj}\bigg(-\frac{1}{2}\log(2\pi)+\frac{1}{2}(\log w_t^{-1})-\frac{1}{2w_t}\beta_{tsj}^2+(\log\omega_{sj})\bigg)+$$
$$+(1-\gamma_{tsj})\bigg(\delta_0(\beta_{tsj})+\log(1-\omega_{sj})\bigg)\Bigg]-$$
$$\mathbb{E}_{\tilde{q}}\Bigg[\gamma_{tsj}\bigg(-\frac{1}{2}\log(2\pi)-\frac{1}{2}(\log\sigma^2_{\beta_{tsj}})-\frac{1}{2}\sigma^{-2}_{\beta_{tsj}}(\beta_{tsj}-\mu_{\beta_{tsj}})^2\bigg)+$$
$$+\log\big((\gamma_{tsj})^{(1)}\big)\bigg)+(1-\gamma_{tsj})\bigg(\delta_0(\beta_{tsj})+\log\big(1-(\gamma_{tsj})^{(1)}\big)\bigg)\Bigg].$$

Using $\mathbb{E}_{\tilde{q}}[\gamma_{tsj}\beta_{tsj}^2] = (\mu^2_{\beta_{tsj}}+\sigma^2_{\beta_{tsj}})(\gamma_{tsj})^{(1)}$ and $\mathbb{E}_{\tilde{q}}[\gamma_{tsj}\sigma^{-2}_{\beta_{tsj}}(\beta_{tsj}-\mu_{\beta_{tsj}})^2] = (\gamma_{tsj})^{(1)}$,

$$B(\beta_{tsj},\gamma_{tsj}|w_t,\omega_{sj}) = \frac{(\gamma_{tsj})^{(1)}}{2}\bigg((\log w_t^{-1})^{(1)}+\log\sigma^2_{\beta_{tsj}}+2(\log\omega_{sj})^{(1)}-2\log\big((\gamma_{tsj})^{(1)}\big)+$$
$$-(w_t^{-1})^{(1)}(\mu^2_{\beta_{tsj}}+\sigma^2_{\beta_{tsj}})+1\bigg)+$$
$$(1-(\gamma_{tsj})^{(1)})\bigg((\log(1-\omega_{sj}))^{(1)}-\log\big(1-(\gamma_{tsj})^{(1)}\big)\bigg). \tag{8.5.110}$$

$$B^*(\rho_{tkj},\eta_{tkj}|\cdot) = \mathbb{E}_{\tilde{q}}[\log p(\rho_{tkj}|\sigma^2_{tj},\tau,\eta_{tkj})]+\mathbb{E}_{\tilde{q}}[\log p(\eta_{tkj}|\lambda_j)-\mathbb{E}_{\tilde{q}}[\log\tilde{q}(\rho_{tkj},\eta_{tkj})]$$

$$B^*(\rho_{tkj},\eta_{tkj}|\cdot) = \mathbb{E}_{\tilde{q}}\bigg[\eta_{tkj}\bigg(-\frac{\tau}{2\sigma^2_{tj}}\rho^2_{tkj}\bigg)\bigg]+\frac{(\log\tau)^{(1)}}{2}+\frac{(\log\sigma_{tj}^{-2})^{(1)}}{2}-\frac{\log 2\pi}{2}+$$
$$+(1-(\eta_{tkj})^{(1)})\delta_0(\rho_{tkj})+(\eta_{tkj})^{(1)}(\log\lambda_j)^{(1)}+(1-\eta_{tkj})^{(1)}(\log(1-\lambda_j))^{(1)}+$$
$$-\mathbb{E}_{\tilde{q}}\bigg[\eta_{tkj}\bigg(-\frac{1}{2\sigma^2_{\rho_{tkj}}}(\rho_{tkj}-\mu_{\rho_{tkj}})^2-\frac{\log 2\pi}{2}-\frac{\log\sigma^2_{\rho_{tkj}}}{2}\bigg)+\delta_0(\rho_{tkj})+$$
$$+\eta_{tkj}(\eta_{tkj})^{(1)}+(1-\eta_{tkj})(1-(\eta_{tkj})^{(1)})\bigg]$$
$$=\frac{(\eta_{tkj})^{(1)}}{2}\bigg[(\log\tau)^{(1)}+(\log\sigma_{tj}^{-2})^{(1)}-(\sigma^2_{\rho_{tkj}}+\mu^2_{\rho_{tkj}})(\tau)^{(1)}(\sigma_{tj}^{-2})^{(1)}+$$
$$+2(\log\lambda_j)^{(1)}+1+\log\sigma^2_{\rho_{tkj}}-2\log\big((\lambda_j)^{(1)}\big)\bigg]+$$
$$+(1-\eta_{tkj})^{(1)}\bigg[(\log(1-\lambda_j))^{(1)}-\log\big(1-(\lambda_j)^{(1)}\big)\bigg] \tag{8.5.111}$$

For $C(z_{ig}|\boldsymbol{x}_{i,.}, \boldsymbol{\upsilon}_g)$ the lower bound simplifies the calculation. Defining

$$\mathbb{E}_{\tilde{q}}\Big[(\boldsymbol{\upsilon}_g^T\boldsymbol{x}_{i,.})^2\Big] = \sum_s(\sigma^2_{\upsilon_{gs}} + \mu^2_{\upsilon_{gs}})(\epsilon_s)^{(1)}x_{is}^2 + \sum_{s=1}^{p-1}\mu_{\upsilon_{gs}}(\epsilon_s)^{(1)}x_{is}\sum_{h=s+1}^{p}(\mu_{\upsilon_{gh}})(\epsilon_h)^{(1)}, \tag{8.5.112}$$

$$\begin{aligned}C(z_{ig}|\cdot) =&\mathbb{E}_{\tilde{q}}\Big[\log p(z_{ig}|\boldsymbol{x}_{i,.}, \boldsymbol{\upsilon}_g) - \log\tilde{q}(z_{ig}|\boldsymbol{x}_{i,.}, \boldsymbol{\upsilon}_g)\Big]\\ =&(z_{ig})^{(1)}(\boldsymbol{\upsilon}_g)^{(1)T}\boldsymbol{x}_{i,.} + \log\sigma(\psi_{in}) + \frac{-(\upsilon_g)^{(1)T}X_n - \psi_{ig}}{2} - \Big[\lambda_*(\psi_{ig})(\mathbb{E}_q(\boldsymbol{\upsilon}_g^T\boldsymbol{x}_{i,.})^2 - \psi_{ig}^2)\Big] +\\ &- (z_{ig})^{(1)}\log\sigma((C_{ig})^{(1)}) - (1-(z_{ig})^{(1)})\log\big(1-\sigma((C_{ig})^{(1)})\big)\end{aligned} \tag{8.5.113}$$

$$D(\boldsymbol{\upsilon}_s, \epsilon_s|d,\kappa) = \mathbb{E}_{\tilde{q}}[\log p(\upsilon_s|d,\epsilon_s)] + \mathbb{E}_{\tilde{q}}[\log p(\epsilon_s|\kappa)] - \mathbb{E}_{\tilde{q}}[\log\tilde{q}(\upsilon_s,\epsilon_s)]\Big]$$

$$\begin{aligned}D(\boldsymbol{\upsilon}_s, \epsilon_s|d,\kappa) = \mathbb{E}_{\tilde{q}}\Bigg[&\epsilon_s\sum_g\Big\{-\frac{1}{2}\log 2\pi + \frac{(\log d^{-1})}{2} - \frac{d^{-1}}{2}\upsilon_{gs}^2\Big\} + (1-\epsilon_s)\delta_0(\boldsymbol{\upsilon}_s) +\\ &+\epsilon_s\log(\kappa) + (1-\epsilon_s)\log(1-\kappa) +\\ &-\epsilon_s\sum_g\Big\{-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2_{\upsilon_{gs}} - \frac{1}{2\sigma^2_{\upsilon_{gs}}}(\upsilon_{gs}^2 - \mu^2_{\upsilon_{gs}})\Big\} +\\ &-\epsilon_s\log(\epsilon_s)^{(1)} - \epsilon_s\log\big(1-(\epsilon_s)^{(1)}\big)\Bigg]\end{aligned} \tag{8.5.114}$$

Using $\mathbb{E}_{\tilde{q}}[\upsilon_{gs}^2] = (\sigma^2_{\upsilon_{gs}} + \mu^2_{\upsilon_{gs}})(\epsilon_s)^{(1)}$ and $\mathbb{E}_{\tilde{q}}[\sigma^{-2}_{\upsilon_{gs}}(\upsilon_{gs}^2 - \mu^2_{\upsilon_{gs}})] = (\epsilon_s)^{(1)}$

$$\begin{aligned}D(\boldsymbol{\upsilon}_s, \epsilon_s|d,\kappa) =&\frac{(\epsilon_s)^{(1)}}{2}\Bigg(G(\log d^{-1})^{(1)} - G - \sum_g\Big\{(d^{-1})^{(1)}(\sigma^2_{\upsilon_{gs}} + \mu^2_{\upsilon_{gs}}) + \log\sigma^2_{\upsilon_{gs}}\Big\} +\\ &+2\log(\epsilon_s)^{(1)} + 2(\log\kappa_s)^{(1)}\Bigg) +\\ &+(1-(\epsilon_s)^{(1)})\Bigg((\log(1-\kappa_s))^{(1)} - \log\big(1-(\epsilon_s)^{(1)}\big)\Bigg)\end{aligned} \tag{8.5.115}$$

$$
\begin{aligned}
F(\sigma_{tj}^2|\tau,\nu) =& \mathbb{E}_{\tilde{q}}[\log p(\sigma_{tj}^2|\tau,\nu)] - \mathbb{E}_{\tilde{q}}[\log \tilde{q}(\sigma_{tj}^2)] \\
=& \frac{\nu-T+t}{2}\Big((\log\tau)^{(1)} - \log 2\Big) - \log\Gamma\Big(\frac{\nu-T+t}{2}\Big) - \Big(\frac{\nu-T+1}{2}+1\Big)(\log\sigma_{tj}^2)^{(1)} + \\
& - \frac{(\tau)^{(1)}}{2}(\sigma_{tj}^{-2})^{(1)} - \Big[a^*_{\sigma^2,tj}\log b^*_{\sigma^2,tj} - \log\Gamma(a^*_{\sigma^2,tj}) - (a^*_{\sigma^2,tj}+1)(\log\sigma_{tj}^2)^{(1)} + \\
& - b^*_{\sigma^2,tj}(\sigma_{tj}^{-2})^{(1)}\Big] \\
=& \frac{\nu-T+t}{2}\Big((\log\tau)^{(1)} - \log 2\Big) - a^*_{\sigma^2,tj}\log b^*_{\sigma^2,tj} - \log\Gamma\Big(\frac{\nu-T+t}{2}\Big) + \log\Gamma(a^*_{\sigma^2,tj}) + \\
& + (\log\sigma_{tj}^{-2})^{(1)}\Big[\Big(\frac{\nu-T+1}{2}\Big) - a^*_{\sigma^2,tj}\Big] + (\sigma_{tj}^{-2})^{(1)}\Big(b^*_{\sigma^2,tj} - \frac{(\tau)^{(1)}}{2}\Big). \qquad (8.5.116)
\end{aligned}
$$

$$
\begin{aligned}
G(\omega_{sj}) =& \mathbb{E}_{\tilde{q}}[\log p(\omega_{sj})] - \mathbb{E}_{\tilde{q}}[\log \tilde{q}(\omega_{sj})] \\
=& \log B(a^*_\omega, b^*_\omega) - \log B(a_\omega, b_\omega) + \\
& + (a_\omega - a^*_\omega)(\log\omega_{sj})^{(1)} + (b_\omega - b^*_\omega)(\log[1-\omega_{sj}])^{(1)} \qquad (8.5.117)
\end{aligned}
$$

where $B(\cdot,\cdot)$ is the beta function.

$$
\begin{aligned}
H(w_t) =& \mathbb{E}_{\tilde{q}}[\log p(w_t)] - \mathbb{E}_{\tilde{q}}[\log \tilde{q}(w_t)] \\
=& a_w \log b_w - a^*_{w_t}\log b^*_{w_t} + \log\Gamma(a^*_{w_t}) - \log\Gamma(a_w) + \\
& + (a_w - a^*_{w_t})(\log w_t^{-1})^{(1)} + (b^*_{w_t} - b_w)(w_t^{-1})^{(1)} \qquad (8.5.118)
\end{aligned}
$$

$$
\begin{aligned}
I(\kappa) =& \mathbb{E}_{\tilde{q}}[\log p(\kappa)] - \mathbb{E}_{\tilde{q}}[\log \tilde{q}(\kappa)] \\
=& \log B(a^*_\kappa, b^*_\kappa) - \log B(a_\kappa, b_\kappa) + (a_\kappa - a^*_\kappa)(\log\kappa)^{(1)} + (b_\kappa - b^*_\kappa)(\log(1-\kappa)^{(1)} \qquad (8.5.119)
\end{aligned}
$$

$$
\begin{aligned}
J(d) &= \mathbb{E}_{\tilde{q}}[\log p(d)] - \mathbb{E}_{\tilde{q}}[\log q(\tilde{d})] \\
&= a_d \log b_d - a_d^* \log b_d^* + \log \Gamma(a_d^*) - \log \Gamma(a_d) + (a_d - a_d^*)(\log d^{-1})^{(1)} + (b_d^* - b_d)(d^{-1})^{(1)}
\end{aligned} \tag{8.5.120}
$$

$$
\begin{aligned}
K(\lambda_j) =& \mathbb{E}_{\tilde{q}}[\log p(\lambda_j)] - \mathbb{E}_{\tilde{q}}[\log \tilde{q}(\lambda_j)] \\
&= \log B(a_{\lambda_j}^*, b_{\lambda_j}^*) - \log B(a_\lambda, b_\lambda) + (a_\lambda - a_{\lambda_j}^*)(\log \lambda_j)^{(1)} + (b_\lambda - b_{\lambda_j}^*)(\log(1-\lambda_j))^{(1)}
\end{aligned} \tag{8.5.121}
$$

$$
\begin{aligned}
L(\tau) &= \mathbb{E}_{\tilde{q}}[\log p(\tau)] - \mathbb{E}_{\tilde{q}}[\log q(\tau)] \\
&= a_\tau \log b_\tau - a_\tau^* \log b_\tau^* + \log \Gamma(a_\tau^*) - \log \Gamma(a_\tau) + (a_\tau - a_\tau^*)(\log \tau)^{(1)} + (b_d^* - b_d)(\tau)^{(1)}
\end{aligned} \tag{8.5.122}
$$

$$
\begin{aligned}
M(b_w) =& \mathbb{E}_q[\log p(b_w)] - \mathbb{E}_q[\log \tilde{q}(b_w)] \\
&= \mathbb{E}_q\left[a_{b_w} \log b_{b_w} - \log \Gamma(a_{b_w}) + (a_{b_w} - 1)\log b_w - b_{b_w} b_w\right] + \\
&\quad - \mathbb{E}_q\left[a_{b_w}^* \log b_{b_w}^* - \log \Gamma(a_{b_w}^*) + (a_{b_w}^* - 1)\log b_w - b_{b_w}^* b_w\right] \\
&= a_{b_w} \log b_{b_w} - a_{b_w}^* \log b_{b_w}^* - \log \Gamma(a_{b_w}) + \log \Gamma(a_{b_w}^*) + (\log b_w)^{(1)}(a_{b_w} - a_{b_w}^*) + \\
&\quad + (b_w)^{(1)}(b_{b_w}^* - b_{b_w})
\end{aligned} \tag{8.5.123}
$$

$$
\begin{aligned}
M^*(b_d) =& \mathbb{E}_q[\log p(b_d)] - \mathbb{E}_q[\log \tilde{q}(b_d)] \\
=& \mathbb{E}_q\Big[a_{b_d} \log b_{b_d} - \log \Gamma(a_{b_d}) + (a_{b_d} - 1) \log b_d - b_{b_d} b_d\Big] + \\
& - \mathbb{E}_q\Big[a^*_{b_d} \log b^*_{b_d} - \log \Gamma(a^*_{b_d}) + (a^*_{b_d} - 1) \log b_d - b^*_{b_d} b_d\Big] \\
=& a_{b_d} \log b_{b_d} - a^*_{b_d} \log b^*_{b_d} - \log \Gamma(a_{b_d}) + \log \Gamma(a^*_{b_d}) + (\log b_d)^{(1)}(a_{b_d} - a^*_{b_d}) + \\
& + (b_d)^{(1)}(b^*_{b_d} - b_{b_d})
\end{aligned} \tag{8.5.124}
$$

8.5.4 Lower bound on the sigmoid function

We obtain a lower bound on the sigmoid function $g(x) = \sigma(x)$ so the functional form will combine with a Gaussian prior. As the sigmoid function is neither convex nor concave we perform a transformation on both the input variable and of the function itself. The sigmoid function can be expressed as

$$
\begin{aligned}
\log g(x) &= \log\big(1 + e^{-x}\big)^{-1} \\
&= -\log\left(\left(1 + e^{-x}\right)\frac{e^{\frac{x}{2}}}{e^{\frac{x}{2}}}\right) \\
&= \frac{x}{2} - \log\big(e^{\frac{x}{2}} + e^{-\frac{x}{2}}\big),
\end{aligned} \tag{8.5.125}
$$

where $f(x) = -\log\big(e^{\frac{x}{2}} + e^{-\frac{x}{2}}\big)$. An important aspect of (8.5.125) is that the $f(x)$ term is convex in x^2. Thus any first order Taylor approximation of $g(x)$ will be a lower bound. Setting $y = x^2$ and performing the expansion at ϵ

$$
f(y) \approx -\log\left(e^{\frac{\sqrt{\epsilon}}{2}} + e^{-\frac{\sqrt{\epsilon}}{2}}\right) - \frac{1}{4}\epsilon^{-\frac{1}{2}}\frac{e^{\frac{\sqrt{\epsilon}}{2}} - e^{-\frac{\sqrt{\epsilon}}{2}}}{e^{\frac{\sqrt{\epsilon}}{2}} + e^{-\frac{\sqrt{\epsilon}}{2}}}(y - \epsilon). \tag{8.5.126}
$$

Setting $\xi^2 = \epsilon$ and returning to the parameterisation with respect to x

$$\begin{aligned} f(x) &\approx -\log\left(e^{\frac{\xi}{2}} + e^{-\frac{\xi}{2}}\right) - \frac{1}{4\xi}\frac{e^{\frac{\xi}{2}} - e^{-\frac{\xi}{2}}}{e^{\frac{\xi}{2}} + e^{-\frac{\xi}{2}}}(x^2 - \xi^2) \\ &\approx -\log\left(e^{\frac{\xi}{2}} + e^{-\frac{\xi}{2}}\right) - \lambda_*(\epsilon)(x^2 - \xi^2), \end{aligned}$$

where

$$\lambda_*(\xi) = \frac{1}{4\xi}\frac{e^{\frac{\xi}{2}} - e^{-\frac{\xi}{2}}}{e^{\frac{\xi}{2}} + e^{-\frac{\xi}{2}}} = \frac{\tanh\left(\frac{\xi}{2}\right)}{4\xi}. \tag{8.5.127}$$

Thus, using the lower bound for $g(x)$

$$\log g(x) \geq \frac{x}{2} + f(\xi) - \lambda_*(\xi)(x^2 - \xi^2)$$

and $f(\xi) = \log g(\xi) - \xi/2$ from (8.5.125)

$$\log g(x) \geq \frac{x}{2} + \log g(\xi) - \xi/2 - \lambda_*(\xi)(x^2 - \xi^2).$$

Exponentiating gives

$$\sigma(x) \geq \sigma(\xi)\exp\left(\frac{x - \xi}{2} - \lambda_*(\xi)(x^2 - \xi^2)\right) \qquad \square \tag{8.5.128}$$

CHAPTER 9

Discussion

In this last chapter we conclude with a general discussion on possible future extensions. This is split into two areas, the feature selection methods for compositional covariates and the HME model for multidimensional responses.

9.1 Compositional Feature Selection

The compositional feature selection models are accompanied by a series of routines, programmed in Python, to perform the modelling. Our aim is to publish a stand alone Python package to accompany the two publications, freely available for practitioners. There is also scope to incorporate a reduced multiple response model within the software (or a separate package), with a simple design matrix of continuous covariates. This would offer a fast tool for integrated multivariate Quantitative Trait Loci (QTL), particularly aimed at highly correlated molecular phenotypes. In the search for molecular mechanisms mediating the effects of genetic variants, integrating high-

dimensional molecular biomarker data sets is a fundamental problem in bioinformatics. Our VI model, which incorporates correlations across the responses, would be a powerful approach for identifying genes associated with metabolic markers of diseases, where the multivariate response is generally in the order of hundreds.

The ability of both the univariate and multivariate response Bayesian hierarchical linear log-contrast model in detecting the correct compositional covariates to include, as is the case for all regression models, suffers when there is a large degree of multicollinearity. With microbiome data the raw number of OTUs represent organisms that are phenotypically similar and have a related function. The "relatedness" is captured by mapping the OTUs to the taxonomic tree structures using bacterial 16S rRNA databases. This grouping of the microorganisms is then used in the model, reducing the correlation across the compositional covariates. However, these groupings can still be highly correlated even at the phylum level, where there can be as little as six covariates.

Intuitively, this is a problem because we are trying to estimate the effect of changes in the explanatory variable upon the dependent variable. If two explanatory variables exhibit a large correlation, the attempt to isolate the effect of one variable, all other things held constant, is made difficult by the fact that in the sample the variable exhibits little independent variation. The correlation between two explanatory variables implies that changes in one are linked to changes in the other, and thus separating out their individual effects may be difficult.

In the Bayesian approach, multicollinearity can be accounted for in the prior specification. We review three approaches in the literature which adjust either the latent indicator variable or the regression coefficient to perform linear regression in the presence of correlated predictors, and thus imply a possible extension to our models.

9.1.1 Markov random field prior

In variable selection with microbiome data, Zhang et al. (2020) address the correlation of the features by using the raw OTUs in the design matrix (after a suitable transformation). The phylogenetic tree is used to integrate prior information on the similarity of the taxa into a Markov

Random Field (MRF) prior on the variable inclusion indicators. The covariates $i = 1, ..., d$ are assumed to lie in an undirected graph which can be represented by an edge set $\mathcal{E} = \{(i, j) : 1 \leq i \neq j \leq d\}$. Given this graph, let $\boldsymbol{a} = (a_1, ..., a_d)^T$ be a vector and $\boldsymbol{Q} = (q_{ij})_{d\times d}$ by a symmetric matrix of real numbers where $q_{ij} = 0$ for all $(i, j) \notin \mathcal{E}$. The MRF (or Ising) prior distribution for $\boldsymbol{\gamma}$ is thus defined by

$$p(\boldsymbol{\gamma}) = \exp\big(\boldsymbol{a}^T\boldsymbol{\gamma} + \boldsymbol{\gamma}^T\boldsymbol{Q}\boldsymbol{\gamma} - \psi(\boldsymbol{a}, \boldsymbol{Q})\big), \tag{9.1.1}$$

where $\psi(\boldsymbol{a}, \boldsymbol{Q})$ is the normalizing constant. The hyperparameters $\boldsymbol{a}$ control the sparsity of $\boldsymbol{\gamma}$ and $\boldsymbol{Q}$ the smoothness of $\boldsymbol{\gamma}$ over $\mathcal{E}$ (the larger q_{ij}, the greater the probability of the ith and jth covariate being jointly selected). When $q_{ij} = 0$ for all pairs (i, j) the covariates are independent and the prior reduces to an independent Bernoulli prior (Appendix 9.3.1).

The key idea is that the MRF prior increases the likelihood of joint covariates being selected, relative to the correlation between them. The incorporation of biological information on the structured dependence through the $\boldsymbol{Q}$ matrix in a MRF is a popular approach (Lee et al. (2017), Li and Zhang (2010), Vannucci et al. (2012)).

If $\gamma(-i) = \{\gamma_j : j \neq i\}$ and $G_{(-i)}$ be $\{\gamma_j = 1 : j \neq i\}$, the set of indices for the selected variables other than i. The conditional distribution of γ_i is given by

$$p(\gamma_i|\gamma_{(-i)}) = \frac{\exp\Big(\gamma_i a_i + \sum_{j\in G_{(-i)}} q_{ij}\gamma_i\gamma_j\Big)}{1 + \exp\Big(a_i + \sum_{j\in G_{(-i)}} q_{ij}\gamma_j\Big)}, \tag{9.1.2}$$

which can be combined with the marginal likelihood (if available) in an MCMC sampler to obtain the marginal model posterior.

To maintain sparsity, $\boldsymbol{a} = a(1, ..., 1)^T$ with a fixed to a negative integer between (0 to -30) since the smaller a_i is, the more likely it is a priori that the ith covariate will be omitted. The matrix $\boldsymbol{Q}$ is set to the inverse of the phlyogeny-induced correlation matrix (Euclidean of exponential). The prior is sensitive to the choice of hyperparameters (Li and Zhang, 2010), although empirical Bayes is used to choose $\boldsymbol{Q}$, the value of $\boldsymbol{a}$ is not obvious and will require sensitivity analysis from multiple runs of the algorithm. (Zhao et al., 2021) suggests specifying a range for the model sparsity (c_1, c_2)

and specifying the hyper-parameter $a = \text{logit}(c_1)$. Then, c_1 represents a lower bound for sparsity which is reached when the covariates are all independent.

9.1.2 Gram matrix

Yuan and Lin (2005) also adjust the prior on the latent indicator variable $\boldsymbol{\gamma}$ to account for correlation between the predictors, if two predictors are highly correlated only one is included in the selected model (rather than both, with the MRF prior). In a simple linear regression framework, with a normal spike-and-slab prior in the form of (2.1.4), the standard product of Bernoulli priors can be multiplied by the square root of the determinant of the Gram matrix of predictors $|\boldsymbol{X}^T\boldsymbol{X}|$,

$$p(\boldsymbol{\gamma}) \propto \omega^{|\gamma|}(1-\omega)^{d-|\gamma|}\sqrt{\det(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma)} \tag{9.1.3}$$

where $\det(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma) = 1$ if $|\boldsymbol{\gamma}| = 0$. When the correlation between the two covariates goes to 1, the prior converges to a prior that only allows one of the two variables in the model. This can be observed from the conditional prior odds ratio for $\gamma_j = 1$

$$\frac{p(\gamma_j = 1|\boldsymbol{\gamma}_{(-j)})}{p(\gamma_j = 0|\boldsymbol{\gamma}_{(-j)})} = \frac{\omega}{1-\omega}\sqrt{\frac{\det(\boldsymbol{X}^T_{\boldsymbol{\gamma}_{(-j)},\gamma_j=1}\boldsymbol{X}_{\boldsymbol{\gamma}_{(-j)},\gamma_j=1})}{\det(\boldsymbol{X}^T_{\boldsymbol{\gamma}_{(-j)},\gamma_j=0}\boldsymbol{X}_{\boldsymbol{\gamma}_{(-j)},\gamma_j=0})}}, \tag{9.1.4}$$

where the design matrix in the numerator and denominator of the ratio of determinants, either includes or excludes the $\boldsymbol{X}_j$ covariate alongside the other selected covariates respectively. If the $\boldsymbol{X}_j$ is the last column identified by the index j, the ratio can be expressed as (Appendix 9.3.2)

$$\frac{p(\gamma_j = 1|\boldsymbol{\gamma}_{(-j)})}{p(\gamma_j = 0|\boldsymbol{\gamma}_{(-j)})} \propto \sqrt{\det(\boldsymbol{X}_j^T(\boldsymbol{X}_j - \hat{\boldsymbol{X}}_j))}, \tag{9.1.5}$$

where $\hat{\boldsymbol{X}}_j$ are the fitted values from the OLS regression of $\boldsymbol{X}_j$ on $\boldsymbol{X}_{\boldsymbol{\gamma}_{(-j)},\gamma_j=0}$. Clearly, if $\boldsymbol{X}_j$ is highly correlated with the current covariates $\boldsymbol{X}_{\boldsymbol{\gamma}_{(-j)},\gamma_j=0}$, the residuals will be small and the conditional prior odds ratio for $\gamma_j = 1$ will be low. Therefore it is more likely that $\boldsymbol{X}_j$ will be removed from the full model, in direct contrast to the MRF prior.

9.1.3 Dirichlet process

When the covariates exhibit multicollinearity, the data is deficient for determining the independent effects of a covariate on the responses as the covariates are not independent, so the covariates can be considered to "move together". Rather than express this via the latent indicator variable (Zhang et al. (2020), Yuan and Lin (2005)), Curtis and Ghosh (2011) propose a prior distribution on the space of all linear regression coefficient restrictions of the form $\beta_j = \beta_{j'}$ $(j \neq j')$ and $\beta_j = 0$, where the linear restrictions on the coefficient parameters are determined by the data. The prior is based on the Dirichlet process (Ferguson, 1973), indexed by a base distribution $H(\cdot)$ and precision parameter α. The base distribution can be thought of informally as the center of the random distributions from the Dirichlet process, and the precision parameter α controls how "close" the random distributions from the Dirichlet process are to the base distribution $H(\cdot)$. As in the elastic net, the authors aim to select groups of variables that are highly correlated.

The clustering properties of the Dirichlet process ensure a positive probability to events $\theta_i = \theta_j$ $(i \neq j)$, for a sequence of random draws $\theta_1, ..., \theta_d$ from a realization of the Dirichlet process $\mathcal{D}(\cdot)$. The simple linear model with the presence of multicollinearity in the design matrix is

$$y_i|\boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2). \tag{9.1.6}$$

The prior on the regression coefficients β_j is induced by combining random draws $\theta_1, ..., \theta_d$ from $\mathcal{D}(\cdot)$, where $\mathcal{D}(\cdot)$ is a random distribution from a Dirichlet process, and random draws $\gamma_1, ..., \gamma_d$ from a Bernoulli distribution. The key aspects of the Bayesian model can be summarised as

$$\beta_j = \gamma_j\theta_j \qquad j = 1, ..., d \tag{9.1.7}$$

$$\gamma_j \sim \omega^{\gamma_j}(1-\omega)^{1-\gamma_j} \qquad j = 1, ..., d \tag{9.1.8}$$

$$\theta_j|\mathcal{D} \sim \mathcal{D}(\cdot) \qquad j = 1, ..., d \tag{9.1.9}$$

$$\mathcal{D} \sim \mathrm{DP}(\alpha, N(0, \tau^2)) \tag{9.1.10}$$

The independence of the indicator variable and regression coefficient in the prior parameterisation (Kuo and Mallick, 1998), implies $\boldsymbol{\gamma}$ only enters the model via the likelihood. Covariates are removed from the model when $\gamma_j = 0$. The normal distribution $N(0, \tau^2)$ is used for the base distribution of $\mathcal{D}$, which allows for the clustering of the predictors.

9.2 Mixture of Experts

We plan to assess the performance of the proposed HME model and develop accompanying software, before submitting the article to an appropriate journal.

9.2.1 Simulation

The feature selection performance of the model will be compared to existing cluster regression models (frequentist) which have freely available software. The expectation is that by incorporating the latent structure of the response within our approach, the model will outperform those methods that assume independent responses. The R package *flexmix* provides infrastructure for the flexible fitting of finite mixtures models, estimated via the EM algorithm. The E-step is handled by the routine, where as the M-step can be adapted for feature selection, by incorporating a penalisation term of the linear regression coefficients (by adaptive lasso or elastic net with *glmnet*).

The simulation study will be set up by randomly subsampling $p = 50$ single nucleotide polymorphisms (SNPs) from our real omics data set (Golub et al., 1999). This forms our covariate set $\boldsymbol{X}$ and allows us to mimic correlation effects and linkage disequilibrium between genetic markers that would be difficult to simulate artificially.

For each observation, first an indicator variable of three levels will be drawn from a multivariate

distribution with probability vector

$$\boldsymbol{\zeta}_i = \begin{bmatrix} \sigma(\boldsymbol{\upsilon}_{1,\boldsymbol{\epsilon}}^T \boldsymbol{x}_{i,.}) \\ \sigma(\boldsymbol{\upsilon}_{2,\boldsymbol{\epsilon}}^T \boldsymbol{x}_{i,.})(1 - \sigma(\boldsymbol{\upsilon}_1^T \boldsymbol{x}_{i,.})) \\ (1 - \sigma(\boldsymbol{\upsilon}_{1,\boldsymbol{\epsilon}}^T \boldsymbol{x}_{i,.}))(1 - \sigma(\boldsymbol{\upsilon}_{2,\boldsymbol{\epsilon}}^T \boldsymbol{x}_{i,.})) \end{bmatrix}. \tag{9.2.1}$$

The vector of indicator variables $\boldsymbol{\epsilon}$ are fixed, so the same sparse subset of covariates determine each cluster probability. The different vector values for $\boldsymbol{\upsilon}_1$ and $\boldsymbol{\upsilon}_2$ are chosen so that the probability for each cluster is similar.

Each $T = 5$ dimensional response, given the sampled cluster identifier, will be drawn from a multivariate normal

$$p(\boldsymbol{y}_i | \boldsymbol{x}_{i,.}, \mathbf{B}_j, \boldsymbol{C}_j, \boldsymbol{\zeta}_i = j) \sim N_T(\boldsymbol{x}_{i,.}^T \mathbf{B}_j, \boldsymbol{C}_j), \tag{9.2.2}$$

where $\boldsymbol{B}_j$ and $\boldsymbol{C}_j$ are the cluster specific parameters. To present a range of possible association patterns between outcomes and predictors, we fix the binary indicators $\boldsymbol{\gamma}_j$ so that different set of predictors display a variety of associations for each cluster. The total number of "significant" predictors for each response will be small, to reflect the presence of sparsity so common in omics data. Given the small number of responses, we specify a sparse inverse error covariance matrix (which will lead to a dense covariance matrix) for each cluster $\boldsymbol{C}_j^{-1}$.

Two summaries of signal to noise for each cluster will be considered, constructed to detect information contained in the predictors and covariance matrix respectively

$$\text{SNR}_{\beta_j} = \frac{1}{T} \sum_t \frac{\frac{1}{n-1} \sum_{i=1}^n (\boldsymbol{x}_{i,\boldsymbol{\gamma}_{tj}}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}_{tj}})(\boldsymbol{x}_{i,\boldsymbol{\gamma}_{tj}}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}_{tj}})}{\sigma_{tj}^2}, \tag{9.2.3}$$

$$\text{SNR}_{C_j} = \frac{1}{T} \sum_t \frac{\frac{1}{n-1} \sum_i (\boldsymbol{u}_{ij}^T \boldsymbol{\rho}_{\boldsymbol{\eta}_{tj}})(\boldsymbol{u}_{ij}^T \boldsymbol{\rho}_{\boldsymbol{\eta}_{tj}})}{\sigma_{tj}^2}. \tag{9.2.4}$$

The performance of the models will be compared by three different criteria, sensitivity/specificity, parameter estimation, and clustering performance. The sensitivity/specificity is defined by:

- *Sensitivity*: proportion of correctly estimated zero regression coefficients,

- *Specificity*: proportion of correctly estimated non-zero regression coefficients.

Variable selection within the variational Bayes HME model can be performed by thresholding the marginal approximate posterior distribution of each latent inclusion indicator variable $\mathbb{E}_q[q(\gamma_{tsj}|\boldsymbol{Y})]$ at 0.5.

For the clustering criterion, once the model has been estimated, $q(\boldsymbol{\zeta}|\boldsymbol{Y})$ represents a soft partition of the data. A hard partition of the can be performed by finding the maximum element of the expectation of the approximate marginal posterior

$$\hat{c}_i = \arg\max_{j=1}^{J} \mathbb{E}_q[\zeta_{ij}] \tag{9.2.5}$$

where $\hat{c}_i$ represents the estimated cluster level for the ith observation. Given the estimated and true cluster labels, we can compute the correct classification rate and the adjusted rand index.

9.2.2 Application on dataset

To demonstrate the clustering accuracy of the approach, the HME model will be applied to the data set in Golub et al. (1999) which contains measurements of leukaemia patients' gene expression levels from 38 bone marrow samples. Acute leukemia can be classified into acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML), depending on whether the cancer arises from lymphoid precursor cells or myeloid precursor cells. Twenty-seven of the patients have ALL and eleven have AML. Each bone marrow sample provides the quantitative expression levels of 6817 genes, but a subset of 50 genes most highly correlated with ALL-AML class distinction has been identified by Golub et al. (1999).

The 50 genes within the subset are highly correlated with one-another. We will capitalise on the multicollinearity and fit a regression model in which a small subset of genes serve as the multivariate response and the rest as the explanatory variables. The goal is asses the performance the HME model to classify the 38 samples into the ALL and AML subgroups.

By thresholding the marginal approximate posterior expectation for each indicator variable with respect to the mixture covariate $\mathbb{E}_q[q(\gamma_{tsj}|\boldsymbol{Y})]$ at 0.5, we can determine the explanatory variables selected for each expert. Performing the same thresholding of $\mathbb{E}_q[q(\epsilon_s|\boldsymbol{Y})]$, will reveal the exploratory variables which determine the two clusters.

9.2.3 Software options

Despite the local and global variable structure in the HME model, the model can be scaled to massive data sets by employing stochastic VI (explained in Section 4.6). The approach requires a modest change in the local updates, which achieves large computational savings when n (number of samples) is massive. The speed of this approach can be improved by using the method developed in Ranganath et al. (2013) (and explained in Section 4.7) by optimally adapting the learning rate. The HME software will incorporate the option of estimating the model by either CAVI or SVI, depending on the size of the input dataset.

9.3 Appendix

9.3.1 Markov Random Field Prior

The Markov Random Field (MRF) prior on the variable inclusion indicators is defined as

$$p(\boldsymbol{\gamma}) \propto \exp\big(\boldsymbol{a}^T\boldsymbol{\gamma} + \boldsymbol{\gamma}^T\boldsymbol{Q}\boldsymbol{\gamma}\big), \tag{9.3.1}$$

where $\boldsymbol{a}$ is a d dimensional vector and $\boldsymbol{Q}$ is a matrix, with elements $\{q_{ij}\}$ set to some constants for the connected nodes and to 0 for the non-connected ones. If $\gamma(-i) = \{\gamma_j : j \neq i\}$ and $G_{(-i)}$ be $\{\gamma_j = 1 : j \neq i\}$, the set of indices for the selected variables other than i. The conditional

distribution of γ_i is given by

$$p(\gamma_i|\gamma_{(-i)}) = \frac{\exp\left(\gamma_i a_i + \sum_{j\in G_{(-i)}} q_{ij}\gamma_i\gamma_j\right)}{1+\exp\left(a_i + \sum_{j\in G_{(-i)}} q_{ij}\gamma_j\right)}, \tag{9.3.2}$$

as the normalising constant is equal to the sum of the two proportional probabilities

$$\tilde{p}(\gamma_i = 1) + \tilde{p}(\gamma_i = 0) = \exp\left(a_i + \sum_{j\in G_{(-i)}} q_{ij}\gamma_j\right) + 1. \tag{9.3.3}$$

If there are no connected nodes to i, then prior distribution for γ_i reduces to a Bernoulli distribution where the parameter η is

$$\eta = \frac{\exp(a_i)}{1+\exp(a_i)}, \tag{9.3.4}$$

the logistic transformation of a_i. This motivates the choice of hyper-parameter by Zhao et al. (2021), where η is specified in terms of the expected sparsity then back transformed for a sparsity scalar parameter of $a = \text{logistic}^{-1}(\eta)$.

9.3.2 Determinant of the Gram matrix

The ratio of the determinants of the Gram matrices from (9.1.4), the conditional prior odds ratio for $\gamma_j = 1$, is derived. This is used in the prior of Yuan and Lin (2005) for the latent vector of indicator variables $\boldsymbol{\gamma}$

$$p(\boldsymbol{\gamma}) = \omega(1-\omega)\sqrt{\frac{\det\left(\boldsymbol{X}^T_{\boldsymbol{\gamma}_{(-j)},\gamma_j=1}\boldsymbol{X}_{\boldsymbol{\gamma}_{(-j)},\gamma_j=1}\right)}{\det\left(\boldsymbol{X}^T_{\boldsymbol{\gamma}_{(-j)},\gamma_j=0}\boldsymbol{X}_{\boldsymbol{\gamma}_{(-j)},\gamma_j=0}\right)}}, \tag{9.3.5}$$

where superscript $(-j)$ indicates the jth component is removed. Defining the $n \times m$ matrix,

$$\boldsymbol{X}_{\boldsymbol{\gamma}_{(-j)},\gamma_j=1} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 & \ldots & \boldsymbol{X}_j & \ldots & \boldsymbol{X}_p \end{pmatrix}, \tag{9.3.6}$$

and the $n \times (m-1)$ matrix with the jth column removed as

$$\boldsymbol{X}_{\boldsymbol{\gamma}_{(-j)},\gamma_j=0} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 & \ldots & \boldsymbol{X}_p \end{pmatrix}. \tag{9.3.7}$$

Rearranging the columns, the design matrix $\boldsymbol{X}_{\boldsymbol{\gamma}_{(-j)},\gamma_j=1}$ can be expressed as

$$\boldsymbol{X}_{\boldsymbol{\gamma}_{(-j)},\gamma_j=1} = \begin{pmatrix} \boldsymbol{X}_{\boldsymbol{\gamma}_{(-j)},\gamma_j=0} & \boldsymbol{X}_j \end{pmatrix}. \tag{9.3.8}$$

Thus, dropping $\boldsymbol{\gamma}_{(-j)}$ from the subscript notation for clarity, the Gram matrix of $\boldsymbol{X}^T_{\gamma_j=1}\boldsymbol{X}_{\gamma_j=1}$ is a block matrix of $((m-1)+1) \times ((m-1)+1)$

$$\boldsymbol{X}^T_{\gamma_j=1}\boldsymbol{X}_{\gamma_j=1} = \begin{pmatrix} \boldsymbol{X}^T_{\gamma_j=0}\boldsymbol{X}_{\gamma_j=0} & \boldsymbol{X}^T_{\gamma_j=0}\boldsymbol{X}_j \\ \boldsymbol{X}^T_j\boldsymbol{X}_{\gamma_j=0} & \boldsymbol{X}^T_j\boldsymbol{X}^T_j \end{pmatrix}. \tag{9.3.9}$$

The lower-diagonal-upper (LDU) decomposition of a block matrix $\mathbf{M}$, provides the determinant property of (Ouellette, 1981)

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \qquad \det(\mathbf{M}) = \det(\mathbf{A})\det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}). \tag{9.3.10}$$

The matrices on the right side of the determinant equation are the matrix $\mathbf{A}$ and the Schur complement with respect to $\mathbf{A}$. Thus,

$$\det(\boldsymbol{X}^T_{\gamma_j=1}\boldsymbol{X}_{\gamma_j=1}) = \det(\boldsymbol{X}^T_{\gamma_j=0}\boldsymbol{X}_{\gamma_j=0})\det(\boldsymbol{X}^T_j\boldsymbol{X}^T_j - \boldsymbol{X}^T_j\boldsymbol{X}_{\gamma_j=0}(\boldsymbol{X}^T_{\gamma_j=0}\boldsymbol{X}_{\gamma_j=0})^{-1}\boldsymbol{X}^T_{\gamma_j=0}\boldsymbol{X}_j). \tag{9.3.11}$$

The conditional prior odds ratio for γ_j is

$$\begin{aligned} \frac{p(\gamma_j = 1|\boldsymbol{\gamma}_{(-j)})}{p(\gamma_j = 0|\boldsymbol{\gamma}_{(-j)})} &= \frac{\omega}{1-\omega}\sqrt{\frac{\det(\boldsymbol{X}^T_{\gamma_j=1}\boldsymbol{X}_{\gamma_j=1})}{\det(\boldsymbol{X}^T_{\gamma_j=0}\boldsymbol{X}_{\gamma_j=0})}} \\ &\propto \sqrt{\det(\boldsymbol{X}^T_j\boldsymbol{X}^T_j - \boldsymbol{X}^T_j\boldsymbol{X}_{\gamma_j=0}(\boldsymbol{X}^T_{\gamma_j=0}\boldsymbol{X}_{\gamma_j=0})^{-1}\boldsymbol{X}^T_{\gamma_j=0}\boldsymbol{X}_j)}. \end{aligned} \tag{9.3.12}$$

In OLS regression of a response $\mathbf{y}$ on $\boldsymbol{X}$, the fitted values are

$$\hat{\mathbf{y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y}, \tag{9.3.13}$$

and if the residuals $\boldsymbol{\epsilon} = (\mathbf{y} - \hat{\mathbf{y}})$, then

$$\begin{aligned}\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y} &= \mathbf{y}^T(\mathbf{y} - \hat{\mathbf{y}}) \\ &= \mathbf{y}^T\boldsymbol{\epsilon}\end{aligned} \tag{9.3.14}$$

Equation (9.3.11) can be interpreted in terms of the regression of $\boldsymbol{X}_j$ on the design matrix $\boldsymbol{X}_{\gamma_j=0}$. Defining the fitted values as $\hat{\boldsymbol{X}}_j$, from (9.3.14) the conditional prior odds are proportional to

$$\frac{p(\gamma_j = 1|\boldsymbol{\gamma}_{(-j)})}{p(\gamma_j = 0|\boldsymbol{\gamma}_{(-j)})} \propto \sqrt{\det(\boldsymbol{X}_j^T(\boldsymbol{X}_j - \hat{\boldsymbol{X}}_j))}. \tag{9.3.15}$$

Clearly, the more correlated $\boldsymbol{X}_j$ is with $\hat{\boldsymbol{X}}_j$, the smaller the residuals and the lower the conditional prior odds ratio.

CHAPTER 10

Conclusion

Throughout the thesis we have explored an assortment of Bayesian approaches for feature selection. A variety of fast variational inference algorithms have been presented, for manageable model computation in high-volume data. Our hope is that the methods we have outlined and the software we plan to produce, will be used by practitioners to develop biological understanding and insight.

We introduce a Bayesian linear variable selection model that identifies compositional covariates and effect sizes associated with a response of interest. This is particularly useful for data sets generated from genome sequencing technology such as human microbiome, as these only contain information on the relative magnitudes of the compositional components. Our approach fully accounts for: the parameter constraints imposed by transforming the data onto the real line and the capacity of the proportions to differ by several orders of magnitude.

We extend this approach to a multivariate response, where different compositional regressors are free to be associated with different responses. This allows the relationship between the microbiome and complex phenotypes such as lipids or metabolites to be explored in one model, facilitating a

“system genetics" approach to understanding the flow of biological information. By incorporating the latent structure of the responses within a hierarchical framework, we leverage information across the responses, increasing statistical power and improving model estimation. Through a reparameterisation of the likelihood we are able to perform fast covariance and covariate selection despite the vast model space, ensuring the model is generalisable.

A hierarchical Bayesian model is developed for clusters of people who exhibit different causal pathways to the same multi-dimensional endpoint. We capture the different latent structures across the clusters to aid model fitting and understanding. Again, we are able to reparametrise the likelihood to incorporate fast predictor and covariance selection within a large model space. Sparse feature selection is performed both within each expert and in the unsupervised learning of cluster detection.

References

Ackermann, M., Sikora-Wohlfeld, W., and Beyer, A. (2013). Impact of natural genetic variation on gene expression dynamics. *PLoS Genetics*, 9(6).

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall ltd, Caldwell, New Jersey, 1st edition.

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data*. Blackburn Press: Caldwell, NJ, USA.

Aitchison, J. and Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika*, 71(2):323–330.

Aitchison, J. and Shen, S. M. (1980). Logistic-normal distributions: some properties and uses. *Biometrika*, 67(2):261–272.

Amari, S. (1982). Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*, 10(2):357–385.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276.

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119–143.

Armougom, F., Henry, M., Vialettes, B., Raccah, D., and Raoult, D. (2009). Monitoring bacterial community of human gut microbiota reveals an increase in Lactobacillus in obese patients and methanogens in anorexic patients. *PLoS ONE*, 4(9):1–8.

Bander, Z. A., Nitert, M. D., Mousa, A., and Naderpoor, N. (2020). The gut microbiota and inflammation: An overview. *International Journal of Environmental Research and Public Health*, 17(20):1–22.

Banterle, M. and Lewin, A. (2018). Sparse variable and covariance selection for high-dimensional seemingly unrelated Bayesian regression. *bioRxiv*.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897.

Baydian, A. G., Pearlmutter, B. A., Radual, A. A., and Siskind, J. M. (2018). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18:1–43.

Beal, M. J. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics 7: proceedings of the seventh Valencia International Meeting, June 2-6, 2003*, page 453.

Bernado, J. and Smith, A. (1994). *Bayesian Theory*. John Wiley & Sons Ltd.

Betancourt, B., Rodríguez, A., and Boyd, N. (2017). Bayesian Fused Lasso regression for dynamic binary networks. *Journal of Computational and Graphical Statistics*, 26(4):840–850.

Bhadra, A. and Mallick, B. K. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics*, 69(2):447–457.

Bharti, R. and Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, 22(1):178–193.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991.

Bishop, C. and Winn, J. (2006). Variational message passing. *Journal of Machine Learning Research*, 6(1):661.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer New York, 1 edition.

Bishop, C. M. and Svensen, M. (2003). *Bayesian hierarchical mixtures of experts*. UAI.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Bock, A. S. and Fine, I. (2014). Anatomical and functional plasticity in early blind individuals and the mixture of experts architecture. *Frontiers in Human Neuroscience*, 8(DEC):1–13.

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., and Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, 6(90):1–17.

Bottolo, L., Petretto, E., Blankenberg, S., Cambien, F., Cook, S. A., Tiret, L., and Richardson, S. (2011). Bayesian detection of expression quantitative trait loci hot-spots. *Genetics*, 189(4):1449–1459.

Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, 5(3):583–618.

Box, G. E. and Tiao, G. C. (1968). A Bayesian approach to some outlier problems. *Biometrika*, 55(1):119–129.

Brooks, S. P., Giudici, P., and Roberts, G. O. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 65(1):3–39.

Brown, P. J., Vannucci, M., and Fearn, T. (1998). Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics*, 12(1):173–182.

Brown, P. J., Vannucci, M., and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(3):519–536.

Butler, J. C. (1979). The effects of closure on the moments of a distribution. *Journal of the International Association for Mathematical Geology*, 11(7):75–84.

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):473–484.

Carvalho, C. M., Chang, J., Lucas, J., Nevins, J., Wang, Q., and West, M. (2011). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 23(1):1–7.

Carvalho, C. M., Massam, H., and West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika*, 94(3):647–659.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. *Journal of Machine Learning Research*, 5:73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Casella, G. and Robert, C. P. (1996). Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94.

Castillo, I., Schmidt-Hieber, J., and Van Der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Annals of Statistics*, 43(5):1986–2018.

Celeux, G. (1998). Bayesian inference for mixture: the label switching problem. In *Compstat*, pages 227–232.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970.

Challis, E. and Barber, D. (2013). Gaussian Kullback-Leibler approximate inference. *Journal of Machine Learning Research*, 14:2239–2286.

Chen, Y. and Welling, M. (2012). Bayesian structure learning for Markov random fields with a spike and slab prior. *Uncertainty in Artificial Intelligence - Proceedings of the 28th Conference, UAI 2012*, pages 174–184.

Chen, Y., Yang, F., Lu, H., Wang, B., Chen, Y., Lei, D., Wang, Y., Zhu, B., and Li, L. (2011). Characterization of fecal microbial communities in patients with liver cirrhosis. *Hepatology*, 54(2):562–572.

Chen, Y., Zheng, H., xia Zhang, G., lan Chen, F., dan Chen, L., and cong Yang, Z. (2020). High Oscillospira abundance indicates constipation and low BMI in the Guangdong gut microbiome project. *Scientific Reports*, 10(1):1–8.

Chipman, H., George, E. I., and McCulloch, R. E. (2001). The practical implementation of Bayesian model selection. *IMS Lecture Notes*, 38:65–116.

Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488):1646–1660.

Civelek, M. and Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, 15(1):34–38.

Clyde, M. A. and Parmigiani, G. (1998). Protein construct storage: Bayesian variable selection and prediction with mixtures. *Journal of Biopharmaceutical Statistics*, 8(3):431–443.

Combettes, P. L. and Müller, C. L. (2021). Regression models for compositional data: general log-contrast formulations, proximal optimization, and microbiome data applications. *Statistics in Biosciences*, 13(2):217–242.

Cook, S., Malyutina, S., Kudryavtsev, A. V., Averina, M., Bobrova, N., Boytsov, S., Brage, S., Clark, T. G., Benavente, E. D., Eggen, A. E., Hopstock, L. A., Hughes, A., Johansen, H., Kholmatova, K., Kichigina, A., Kontsevaya, A., Kornev, M., Leong, D., Magnus, P., Mathiesen, E., McKee, M., Morgan, K., Nilssen, O., Plakhov, I., Quint, J. K., Rapala, A., Ryabikov, A., Saburova, L., Schirmer, H., Shapkina, M., Shiekh, S., Shkolnikov, V. M., Stylidis, M., Voevoda, M., Westgate, K., and Leon, D. A. (2018). Know your heart: rationale, design and conduct of a cross-sectional study of cardiovascular structure, function and risk factors in 4500 men and women aged 35-69 years from two Russian cities. *Wellcome Open Research*, 3:1–29.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (2007). *Probabilistic networks and expert systems: exact computational methods for Bayesian networks*. Springer.

Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–283.

Cozzini, A., Jasra, A., Montana, G., and Persing, A. (2014). A Bayesian mixture of lasso regressions with t-errors. *Computational Statistics and Data Analysis*, 77:84–97.

Curtis, S. M. K. and Ghosh, S. K. (2011). A Bayesian approach to multicollinearity and the simultaneous selection and clustering of predictors in linear regression. *Journal of Statistical Theory and Practice*, 5(4):715–735.

Davis, C. D. (2016). The gut microbiome and its role in obesity. *Nutrition Today*, 51(4):167–174.

Davis, J. C. (2002). *Statistics and Data Analysis in Geology*. Wiley, 3rd edition.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper-Markov laws in the statistcal analysis of analysis of decomposable graphical models. *Annals of Statistics*, 21:1272–1317.

de Oliveira Neves, V. G., de Oliveira, D. T., Oliveira, D. C., Oliveira Perucci, L., dos Santos, T. A. P., da Costa Fernandes, I., de Sousa, G. G., Barboza, N. R., and Guerra-Sá, R. (2020). High-sugar diet intake, physical activity, and gut microbiota crosstalk: implications for obesity in rats. *Food Science and Nutrition*, 8(10):5683–5695.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12(1):27–36.

Demmie, M., Mcleod, G., and Hall, P. (2015). The CAGE questionnaire: validation of a new alcoholism screening instrumenttle. *The American Journal of Psychiatry*, 131(10):1121–1123.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B: Methodological*, 39(1):1–38.

DeSarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5(2):249–282.

Dominianni, C., Sinha, R., Goedert, J. J., Pei, Z., Yang, L., Hayes, R. B., and Ahn, J. (2015). Sex, body mass index, and dietary fiber intake influence the human gut microbiome. *PLoS ONE*, 10(4):1–14.

Drummond, C. A., Brewster, P. S., He, W., Ren, K., Xie, Y., Tuttle, K. R., Haller, S. T., Jamerson, K., Dworkin, L. D., Cutlip, D. E., Murphy, T. P., D'Agostino, R. B., Henrich, W. L., Tian, J., Shapiro, J. I., and Cooper, C. J. (2017). Cigarette smoking and cardio-renal events in patients with atherosclerotic renal artery stenosis. *PLoS ONE*, 12(3):1–15.

Duvenaud, D. and Adams, R. P. (2016). Black-box stochastic variational inference in five lines of Python. In *Black Box Learning and Inference*, pages 1–4.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7):795–828.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300.

Emani, M. K. and O'Boyle, M. (2015). Celebrating diversity: A mixture of experts approach for runtime mapping in dynamic environments. *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*, 2015-June:499–508.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Filzmoser, P., Hron, K., and Reimann, C. (2010). The bivariate statistical analysis of environmental (compositional) data. *Science of the Total Environment*, 408(19):4230–4238.

Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., and Huttenhower, C. (2015). Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nature Reviews Microbiology*, 13:360–372.

Frühwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453):194–209.

Fruhwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer, New York, 1st edition.

Frühwirth-Schnatter, S. (2011). Dealing with label switching under model uncertainty. In Robert, C. P., Mengersen, K. L., and Titterington, D. M., editors, *Mixture Estimation and Applications*, chapter 10, pages 213–239. John Wiley & Sons Ltd.

Funamoto, M., Shimizu, K., Sunagawa, Y., Katanasaka, Y., Miyazaki, Y., Komiyama, M., Yamakage, H., Satoh-Asahara, N., Takahashi, Y., Wada, H., Hasegawa, K., and Morimoto, T. (2019).

Serum cystatin C, a sensitive marker of renal function and cardiovascular disease, decreases after smoking cessation. *Circulation Reports*, 1(12):623–627.

Gelman, A. (2006). Conservative prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 34(3):515–533.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 1(7):339–373.

Ghosh, S., Yao, J., and Doshi-Velez, F. (2018). Structured variational learning of Bayesian neural networks with horseshoe priors. *35th International Conference on Machine Learning, ICML 2018*, 4:2819–2832.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 8:1–6.

Goldfeld, S. M. and Quandt, R. E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, 1(1):3–15.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gassenbeck, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., and Others (1999). Molecular classification of cancer: class discovery. *Science*, 286(October):531–537.

Gormley, I. C. and Murphy, T. B. (2010). A mixture of experts latent position cluster model for social network data. *Statistical Methodology*, 7(3):385–405.

Gosh, D. and Smolkin, M. (2003). Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, 4(1):36.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Green, P. J. and Thomas, A. (2013). Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, 100(1):91–110.

Griffin, J. E. and Brown, P. J. (2005). Alternative prior distributions for variable selection with very many more variables than observations.

Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics*, 5(3):1780–1815.

Hammond, P. and Suttie, M. (2012). Large-scale objective phenotyping of 3D facial morphology. *Human Mutation*, 33(5):817–825.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102(478):507–516.

Haro, C., Rangel-Zúñiga, O. A., Alcalá-Díaz, J. F., Gómez-Delgado, F., Pérez-Martínez, P., Delgado-Lista, J., Quintana-Navarro, G. M., Landa, B. B., Navas-Cortés, J. A., Tena-Sempere, M., Clemente, J. C., López-Miranda, J., Pérez-Jiménez, F., and Camargo, A. (2016). Intestinal microbiota is influenced by gender and body mass index. *PLoS ONE*, 11(5):1–16.

Harrell, F. E. (2021). rms: regression modeling strategies.

Hathaway, R. J. (1985). A constrained formulation of the maximum-likielihood estimation for normal mixture distributions. *Annals of Statistics*, 13(2):795–800.

Hobby, G. P., Karaduta, O., Dusio, G. F., Singh, M., Zybailov, B. L., and Arthur, J. M. (2019). Chronic kidney disease and the gut microbiome. *American Journal of Physiology - Renal Physiology*, 316(6):F1211–F1217.

Hoffman, M. D. and Blei, D. M. (2015). Structured stochastic variational inference. *Journal of Machine Learning Research*, 38:361–369.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning*, 14(1):1303–1347.

Holmes, C. C., Denison, D. G., and Mallick, B. K. (2002). Accounting for model uncertainty in seemingly unrelated regressions. *Journal of Computational and Graphical Statistics*, 11(3):533–551.

Honkela, A. and Valpola, H. (2005). Unsupervised variational Bayesian learning of nonlinear models. In *Advances in Neural Information Processing Systems*.

Hosmer, D. W. (1974). Maximum likelihood estimates of the parameters of a mixture of two regression lines. *Communications in Statistics, Part A - Theory and Methods*, 3:995–1006.

Hurn, M., Justel, A., and Robert, C. P. (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79.

Ignacio, A., Fernandes, M. R., Rodrigues, V. A., Groppo, F. C., Cardoso, A. L., Avila-Campos, M. J., and Nakano, V. (2016). Correlation between body mass index and faecal microbiota from children. *Clinical Microbiology and Infection*, 22(3):258.e1–258.e8.

Inouye, M., Ripatti, S., Kettunen, J., Lyytikäinen, L. P., Oksala, N., Laurila, P. P., Kangas, A. J., Soininen, P., Savolainen, M. J., Viikari, J., Kähönen, M., Perola, M., Salomaa, V., Raitakari, O., Lehtimäki, T., Taskinen, M. R., Järvelin, M. R., Ala-Korpela, M., Palotie, A., and de Bakker, P. I. (2012). Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS Genetics*, 8(8):e1002907.

Jaakkola, T. S. and Jordan, M. I. (1997). A variational approach to Bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37.

Johndrow, J., Orenstein, P., and Bhattacharya, A. (2020). Scalable approximate MCMC algorithms for the horseshoe prior. *Journal of Machine Learning Research*, 21:1–61.

Kanaujia, A. and Metaxas, D. (2006). Learning ambiguities using Bayesian mixture of experts. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, (November):436–440.

Kettunen, J., Tukiainen, T., Sarin, A. P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L. P., Kangas, A. J., Soininen, P., Würtz, P., Silander, K., Dick, D. M., Rose, R. J., Savolainen, M. J., Viikari, J., Kähönen, M., Lehtim Currency Signki, T., Pietil Currency Signinen, K. H., Inouye, M., McCarthy, M. I., Jula, A., Eriksson, J., Raitakari, O. T., Salomaa, V., Kaprio, J., J Currency Signrvelin, M. R., Peltonen, L., Perola, M., Freimer, N. B., Ala-Korpela, M., Palotie, A., and Ripatti, S. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature Genetics*, 44(3):269–276.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27(4):887–906.

Kiefer, N. M. (1978). Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica*, 46(2):427–434.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, (Ml):1–14.

Koliada, A., Moseiko, V., Romanenko, M., Lushchak, O., Kryzhanovska, N., Guryanov, V., and Vaiserman, A. (2021). Sex differences in the phylum-level human gut microbiota composition. *BMC Microbiology*, 21(1):1–9.

Koslovsky, M. D., Hoffman, K. L., Daniel, C. R., and Vannucci, M. (2020). A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. *Annals of Applied Statistics*, 14(3):1471–1492.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *The Indian Journal of Statistics*, 60(1):65–81.

Kurihara, K. and Sato, T. (2006). Variational Bayesian grammar induction for natural language. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 84–96.

Lamnisos, D., Griffin, J. E., and Steel, M. F. (2009). Transdimensional sampling algorithms for Bayesian variable selection in classification problemswith many more variables than observations. *Journal of Computational and Graphical Statistics*, 18(3):592–612.

Lamnisos, D., Griffin, J. E., and Steel, M. F. (2013). Adaptive Monte Carlo for Bayesian variable selection in regression models. *Journal of Computational and Graphical Statistics*, 22(3):729–748.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Lee, K. H., Tadesse, M. G., Baccarelli, A. A., Schwartz, J., and Coull, B. A. (2017). Multivariate Bayesian variable selection exploiting dependence structure among outcomes: application to air pollution effects on DNA methylation. *Biometrics*, 73(1):232–241.

Lee, K. J., Chen, R. B., and Wu, Y. N. (2016). Bayesian variable selection for finite mixture model of linear regressions. *Computational Statistics and Data Analysis*, 95:1–16.

Lee, S. Y. (2021). Gibbs sampler and coordinate ascent variational inference: A set-theoretical review. *Communications in Statistics - Theory and Methods*, 1(1):1–19.

Lee, Y. S. and Cho, S. B. (2014). Activity recognition with android phone using mixture-of-experts co-trained with labeled and unlabeled data. *Neurocomputing*, 126:106–115.

Leng, C., Tran, M. N., and Nott, D. (2014). Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, 66(1):221–244.

Lewin, A., Saadi, H., Peters, J. E., Moreno-Moral, A., Lee, J. C., Smith, K. G., Petretto, E., Bottolo, L., and Richardson, S. (2016). MT-HESS: An efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues. *Bioinformatics*, 32(4):523–532.

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105(491):1202–1214.

Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94.

Li, Q., Jiang, S., Koh, A. Y., Xiao, G., and Zhan, X. (2019). Bayesian modeling of microbiome data for differential abundance analysis. *arXiv:1902.08741*.

Liang, F. and Wong, W. H. (2000). Evolutionary Monte Carlo: applications to Cp model sampling and change point problem. *Statistica Sinica*, 10(2):317–342.

Likas, A. C. and Galatsanos, N. P. (2004). A variational approach for Bayesian blind image deconvolution. *IEEE Transactions on Signal Processing*, 52(8):2222–2233.

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika*, 101(4).

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review / Revue Internationale de Statistique International Statistical Institute (ISI)*, 63(2):215–232.

Makalic, E. and Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Marlin, B. M., Khan, M. E., and Murphy, K. P. (2011). Piecewise bounds for estimating bernoulli-logistic latent Gaussian models. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 633–638.

Matey-Hernandez, M. L., Williams, F. M., Potter, T., Valdes, A. M., Spector, T. D., and Menni, C. (2018). Genetic and microbiome influence on lipid metabolism and dyslipidemia. *Physiological Genomics*, 50(2):117–126.

Mohammadi, A. and Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138.

Morand, L. and Helm, D. (2019). A mixture of experts approach to handle ambiguities in parameter identification problems in material modeling. *Computational Materials Science*, 167(May):85–91.

Mossavat, S. I., Amft, O., De Vries, B., Petkov, P. N., and Kleijn, W. B. (2010). A Bayesian hierarchical mixture of experts approach to estimate speech quality. *2010 2nd International Workshop on Quality of Multimedia Experience, QoMEX 2010 - Proceedings*, (2010):200–205.

Müller, P. and Quintana, F. (2010). Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, 140(10):2801–2808.

Muntner, P., Winston, J., Uribarri, J., Mann, D., and Fox, C. S. (2008). Overweight, obesity, and elevated serum cystatin C levels in adults in the United States. *American Journal of Medicine*, 121(4):341–348.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics*, 42(2):789–817.

Nearing, J. T., Douglas, G. M., Hayes, M., Macdonald, J., Desai, D., Allward, N., Jones, C. M. A., Wright, R., Dhanani, A., Comeau, A. M., and Langille, M. G. I. (2021). Microbiome differential abundance methods produce disturbingly different results across 38 datasets. *bioRxiv*, 13(1):342.

Niu, Y., Guha, N., De, D., Bhadra, A., Baladandayuthapani, V., and Mallick, B. K. (2020). Bayesian variable selection in multivariate nonlinear regression with graph structures. *arXiv:2010.14638*.

Nott, D. J. and Kohn, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika*, 92(4):747–763.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *American Statistician*, 64(2):154.

Ouellette, D. V. (1981). Schur complements and statistics. *Linear Algebra and Its Applications*, 36(C):187–295.

Paisley, J., Blei, D. M., and Jordan, M. I. (2012). Variational Bayesian inference with stochastic search. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2(2000):1367–1374.

Papathomas, M., Molitor, J., Hoggart, C., Hastie, D., and Richardson, S. (2012). Exploring data from genetic association studies using Bayesian variable selection and the Dirichlet process: application to searching for gene x gene patterns. *Genetic Epidemiology*, 36(6):663–674.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). The Aitchison geometry. In *Modeling analysis of compositional data*, chapter 3, pages 23–31. John Wiley & Sons Ltd., 1st edition.

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413.

Petretto, E., Bottolo, L., Langley, S. R., Heinig, M., McDermott-Roe, C., Sarwar, R., Pravenec, M., Hübner, N., Aitman, T. J., Cook, S. A., and Richardson, S. (2010). New insights into the genetic control of gene expression using a bayesian multi-tissue approach. *PLoS Computational Biology*, 6(4):e1000737.

Polson, N. G. and Scott, J. G. (2011). Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9(1):501–538.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Powell, Philip, D. (2011). Calculating determinants of block natrices. *arXiv:1112.4379*, pages 1–11.

R, T. (2017). R: A Language and Environment for Statistical Computing.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. *Journal of Machine Learning Research*, 33:814–822.

Ranganath, R., Wang, C., Blei, D. M., and Xing, E. P. (2013). An adaptive learning rate for stochastic variational inference. In *30th International Conference on Machine Learning, ICML 2013*.

Ray, K. and Szabó, B. (2021). Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*.

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing Flows. *ICML 15*, 37:1530–1538.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *31st International Conference on Machine Learning, ICML 2014*, 4:3057–3070.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer.

Ross, S. M. (2006). *Simulation*. Elsevier Academic Press, San Diego, 4th edition.

Rothenberg, T. (1971). Identification in parametric models. *Econometrica*, 39:577–591.

Rue, H. (2002). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(2):325–338.

Ruffieux, H., Davison, A. C., Hager, J., and Irincheeva, I. (2017). Efficient inference for genetic association studies with multiple outcomes. *Biostatistics*, 18(4):618–636.

Salimans, T. and Knowles, D. A. (2013). Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882.

Sandhu, H. S., Koley, S., and Sandhu, K. S. (2008). A study of correlation between lipid profile and body mass index (BMI) in patients with diabetes mellitus. *Journal of Human Ecology*, 24(3):227–229.

Schaul, T., Zhang, S., and LeCun, Y. (2013). No more pesky learning rates. In *30th International Conference on Machine Learning, ICML 2013*.

Schwiertz, A., Taras, D., Schäfer, K., Beijer, S., Bos, N. A., Donus, C., and Hardt, P. D. (2010). Microbiota and SCFA in lean and overweight healthy subjects. *Obesity*, 18(1):190–195.

Scott, D. and Lewin, A. (2021). Single response variational inference compositional regression.

Scott, D. and Lewin, A. (2022). Multivariate response variational inference compositional regression.

Scott-Boyer, M. P., Imholte, G. C., Tayeb, A., Labbe, A., Deschepper, C. F., and Gottardo, R. (2012). An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Statistical Applications in Genetics and Molecular Biology*, 11(4):1–30.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*, 14(8):1–14.

Shamai, L., Lurix, E., Shen, M., Novaro, G. M., Szomstein, S., Rosenthal, R., Hernandez, A. V., and Asher, C. R. (2011). Association of body mass index and lipid profiles: evaluation of a broad spectrum of body mass index patients including the morbidly obese. *Obesity Surgery*, 21(1):42–47.

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *Annals of Applied Statistics*, 10(2):1019–1040.

Shin, M., Bhattacharya, A., and Johnson, V. E. (2020). Functional horseshoe priors for subspace shrinkage. *Journal of the American Statistical Association*, 115(532):1784–1797.

Silva, R. and Ghahramani, Z. (2009). The hidden life of latent variables: Bayesian learning with mixed graph models. *Journal of Machine Learning Research*, 10:1187–1238.

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97(460):1141–1153.

Stephens, M. (2000a). Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Annals of Statistics*, 28(1):40–74.

Stephens, M. (2000b). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 62(4):795–809.

Stingo, F. C. and Vannucci, M. (2011). Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics*, 27(4):495–501.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Titsias, M. K. and Lázaro-Gredilla, M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. *31st International Conference on Machine Learning, ICML 2014*, 5:4056–4069.

Tokui, S. and Sato, I. (2016). Reparameterization trick for discrete variables.

Tran, M. N., Nguyen, N., Nott, D., and Kohn, R. (2020). Bayesian deep net GLM and GLMM. *Journal of Computational and Graphical Statistics*, 29(1):97–113.

Tseng, C. H. and Wu, C. Y. (2019). The gut microbiome in obesity. *Journal of the Formosan Medical Association*, 118:S3–S9.

Vannucci, M., Stingo, F. C., and Berzuini, C. (2012). Bayesian models for variable selection that incorporate biological information. *Bayesian Statistics 9*, 9780199694(May 2014).

Verdam, F. J., Fuentes, S., De Jonge, C., Zoetendal, E. G., Erbil, R., Greve, J. W., Buurman, W. A., De Vos, W. M., and Rensen, S. S. (2013). Human intestinal microbiota composition is associated with local and systemic inflammation in obesity. *Obesity*, 21(12):607–615.

Wang, H. (2015). Scaling it up: stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377.

Wermuth, N. (1976). Model search among multiplicative models. *Biometrics*, 32(2):253–263.

Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *NeuroImage*, 21:1732–1747.

Xing, E. P., Jordan, M. I., and Russell, S. (2002). A generalized mean field algorithm for variational inference in exponential families. In *UAI 03: Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 583–591.

Xu, X. and Ghosh, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936.

Yang, Y., Zou, H., and Bhatnagar, S. (2020). gglasso: group lasso penalized learning using a unified BMD algorithm, R package.

Yau, C. and Holmes, C. (2011). Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. *Bayesian Analysis*, 6(2):329–352.

Ye, L., Beskos, A., De Iorio, M., and Hao, J. (2020). Monte Carlo co-ordinate ascent variational inference. *Statistics and Computing*, 30:887–905.

Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100(472):1215–1225.

Yuan, X., Chen, R., McCormick, K. L., Zhang, Y., Lin, X., and Yang, X. (2021). The role of the gut microbiota on the metabolic status of obese children. *Microbial Cell Factories*, 20(1):1–13.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques.*, 6(6):233–243.

Zhang, L., Shi, Y., Jenq, R. R., Do, K. A., and Peterson, C. B. (2020). Bayesian compositional regression with structured priors for microbiome feature selection. *Biometrics*, 77(3):824–838.

Zhao, Z., Banterle, M., Lewin, A., and Zucknick, M. (2021). Structured Bayesian variable selection for multiple related response variables and high-dimensional predictors.