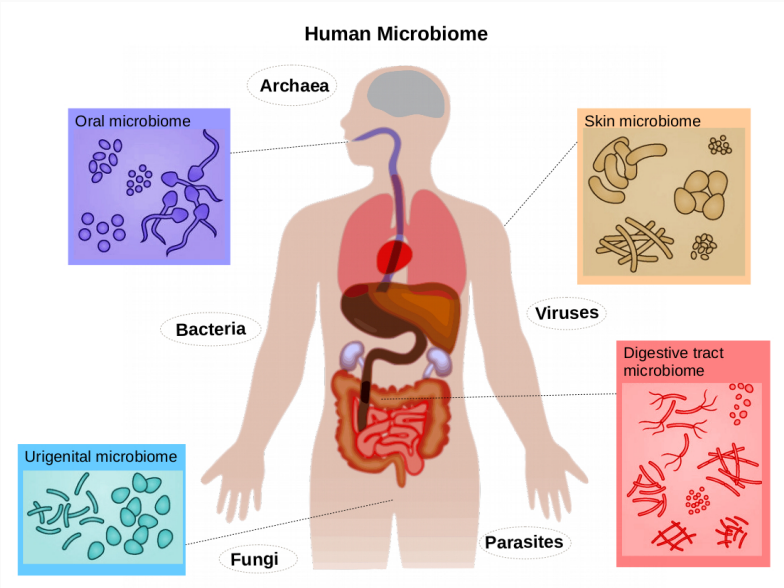
Variational Bayes for Model Averaging for Multivariate models using Compositional predictors

Alex Lewin

(Work with Darren Scott, LSHTM)

December 2020

Microbiome

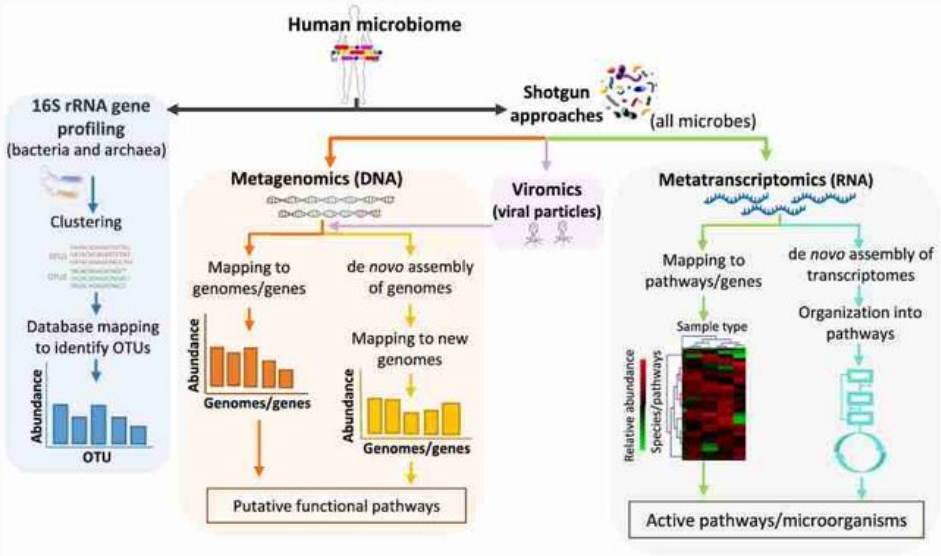


Microbiome: collection of all the microbes inside and on surface of the human body

Microbes interact with immune system, weight regulation etc. etc.

Quantification: next generation sequencing of microbiome species

Obtain relative abundances of different species



Pictures from www.whatisbiotechnology.org, Bikel et al. Computational and Structural Biotechnology Journal 2015

Compositional data as Covariates

Microbiome: p variables $X_1, \cdots, X_p$ with $\sum_j X_j = 1$

Usual to use some form of log transform for microbiome variables as predictors.

Eg, with reference category, use **log-ratio** transform:

$$y = \sum_{j=1}^{p-1} \theta_j log(X_j/X_p) + \epsilon$$

We use an alternative parametrisation:

$$y = \sum_{j=1}^{p} \theta_j log(X_j) + \epsilon$$

with **constraint** $\sum_{j=1}^{p} \theta_j = 0$

Lin et al. Biometrika 2013.

Variable selection with Compositional data

Spike-and-slab prior:

Variable selection performed through binary indicators

$$\xi_j = \begin{cases} 1 & \implies \theta_j \neq 0 \\ 0 & \implies \theta_j = 0 \end{cases}$$

Two issues regarding priors:

1. Prior on non-zero $\boldsymbol{\theta}$?
   Need constraint $\sum_{j=1}^{p} \theta_j = 0$.
2. Prior on $\boldsymbol{\xi}$?
   We cannot have $|\xi| = 1$.

Priors for Variable selection with Compositional data

Singular MVN on non-zero $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_{\boldsymbol{\xi}} | \boldsymbol{\xi} \sim N(T_{\boldsymbol{\xi}} \boldsymbol{\mu}_{\boldsymbol{\xi}}, \sigma^2 T_{\boldsymbol{\xi}} T_{\boldsymbol{\xi}}^T)$$

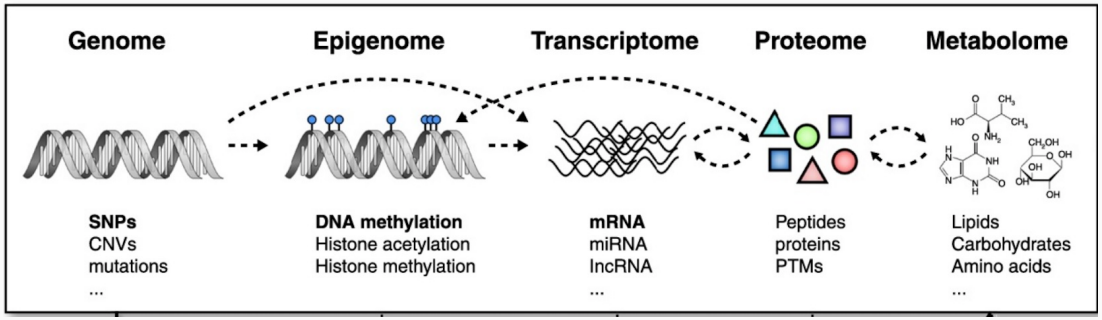
where $T_{\boldsymbol{\xi}} = \mathbb{I}_{|\xi|} - \frac{1}{|\xi|}\mathbb{J}_{|\xi|}$ ensures the sum to zero constraint.

For binary indicators, use a truncated distribution:

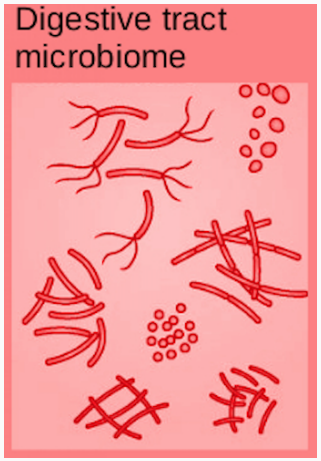
$$\boldsymbol{\xi} \propto \prod_{j=1}^{p} \kappa_j^{\xi_j} (1-\kappa_j)^{(1-\xi_j)} \times \mathrm{I}[|\boldsymbol{\xi}| \neq 1]$$

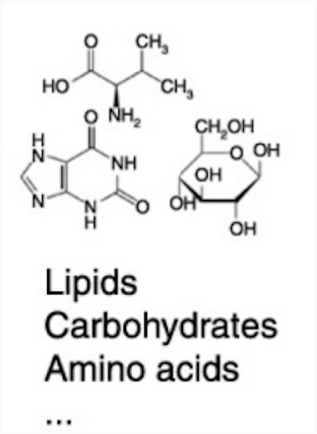
High-dimensional outcomes, high-dimensional predictors

Metabolomics: small molecules, products of cellular processes



Our aim: linking metabolome $\sim$ microbiome





Pictures from Krumsiek et al. Current Opinion in Biotechnology 2016

High-dimensional outcomes, high-dimensional predictors

Frame the problem as a multivariate linear regression model:

$$Y|X, Z \sim \mathcal{MN}(X_\gamma B_\gamma + Z_\xi \Theta_\xi, \mathbb{I}, C_\eta)$$

- Sparse variable selection on matrices of associations (B_γ and Θ_ξ)
- Sparse covariance selection (C_η)
- Model averaging over all combinations of γ, ξ, η provides flexible modelling of B, Θ, C

Our previous work on sparse variable and covariance selection using MCMC:

Bottolo, Banterle, et al. doi: 10.1101/467019 on bioRxiv

Estimating the posterior: Variational Bayes

Full model

$$
\begin{aligned}
p(y,\theta) = &\left\{\prod_t p(y_t|\beta_t,\sigma_t^2,\rho_t)\right\} \times \left\{\prod_t\prod_s p(\beta_{ts}|w_t,\gamma_{ts})\right\} \times \left\{\prod_t p(\theta_t|\Sigma_t(w_t,T),\xi_t)\right\} \times \\
&\left\{\prod_t\prod_s p(\gamma_{ts}|\omega_s)\right\} \times \left\{\prod_t\prod_j p(\xi_{tj}|\kappa_j)\right\} \times \left\{\prod_s p(\omega_s)\right\} \times \left\{\prod_j p(\kappa_j)\right\} \times \\
&\left\{\prod_t p(\sigma_t^2|\tau,\nu)\prod_{k<t} p(\rho_{tk}|\sigma_t^2,\tau,\eta_{tk})\right\} \times \\
&\left\{\prod_t\prod_{k<t} p(\eta_{tk}|\lambda)\right\} \times \left\{\prod_t p(w_t|a_w,b_w)\right\} \times p(\lambda)p(b_w)p(\tau)
\end{aligned}
$$

Mean field approximation

$$
\begin{aligned}
q(\mathbf{z}) = &\left\{\prod_t\prod_s q(\beta_{ts},\gamma_{ts})\right\} \times \left\{\prod_t q(\theta_t,\xi_t)\right\} \times \left\{\prod_s q(\omega_s)\right\} \times \left\{\prod_j q(\kappa_j)\right\} \\
&\left\{\prod_t q(\sigma_t^2)\prod_{k<t} q(\rho_{tk},\eta_{tk}|\sigma_t^2)\right\} \times \\
&\left\{\prod_t q(w_t|a_w,b_w)\right\} \times q(\lambda)q(b_w)q(\tau)
\end{aligned}
$$

Mean field approximation for variable selection

We use CAVI (co-ordinate ascent variational inference).

Unconstrained	Constrained
Need joint distribution for β, γ	Need joint distribution for θ, ξ
Product of Normal and Bernoulli	Product of Singular Normal and Truncated Bernoulli
Treat variables (j) independently	Variables (j) are dependent
Mean field: $\prod_{j=1}^{p} q(\beta_j \vert \gamma_j) q(\gamma_j)$	Mean field: $q(\boldsymbol{\theta} \vert \boldsymbol{\xi}) q(\boldsymbol{\xi})$
VB (CAVI) obtains $E_q(\beta_j \mid \gamma_j)$, $Var_q(\beta_j \mid \gamma_j)$, $E_q(\gamma_j)$	VB (CAVI) obtains $E_q(\boldsymbol{\theta} \mid \boldsymbol{\xi})$, $Var_q(\boldsymbol{\theta} \mid \boldsymbol{\xi})$
	MCMC needed to get $E_q(\boldsymbol{\xi})$

Hybrid VB-MCMC for fitting with constraints

Algorithm 1 Pseudo-Algorithm for Hybrid VB-MCMC

for $i = 1, \ldots, n_{VB}$ **do**

CAVI update obtains $E_q(\boldsymbol{\beta})$

CAVI update obtains $E_q(\boldsymbol{\gamma})$

...

for $j = 1, \ldots, n_{MCMC}$ **do**

Reversible Jump Birth/Death move propose update to $\boldsymbol{\xi}$

Accept/reject using Metropolis-Hastings step

end for

Calculate $\hat{E}_q(\boldsymbol{\xi})$ Monte Carlo average

end for

Ye et al. Statistics and Computing 2020.

Bayesian Model Averaging

Bayesian framework allows for model averaging over all explored models.

Posterior means of binary indicators γ, ξ, η are marginal posterior probabilities of inclusion (MPPI):

> *Using mean field approximation, $E_q(\boldsymbol{\gamma})$ are estimates of posterior means.*
>
> *Monte Carlo average for compositional indicators:*
> $\hat{E}_q(\boldsymbol{\xi}) = \frac{1}{N_{iter}} \sum_{t=1}^{N_{iter}} \boldsymbol{\xi}^{(t)}$

Also obtain shrunk estimates of regression coefficients (include uncertainty from model selection):

> *Mean field approximation: $E_q(\boldsymbol{\beta})$, $E_q(\boldsymbol{\theta})$*

Simulation Study

Predictors (real data from n=514 adult cohort):
- 40 microbiome species
- sex, alcohol intake, diet intake
- 60 noise variables

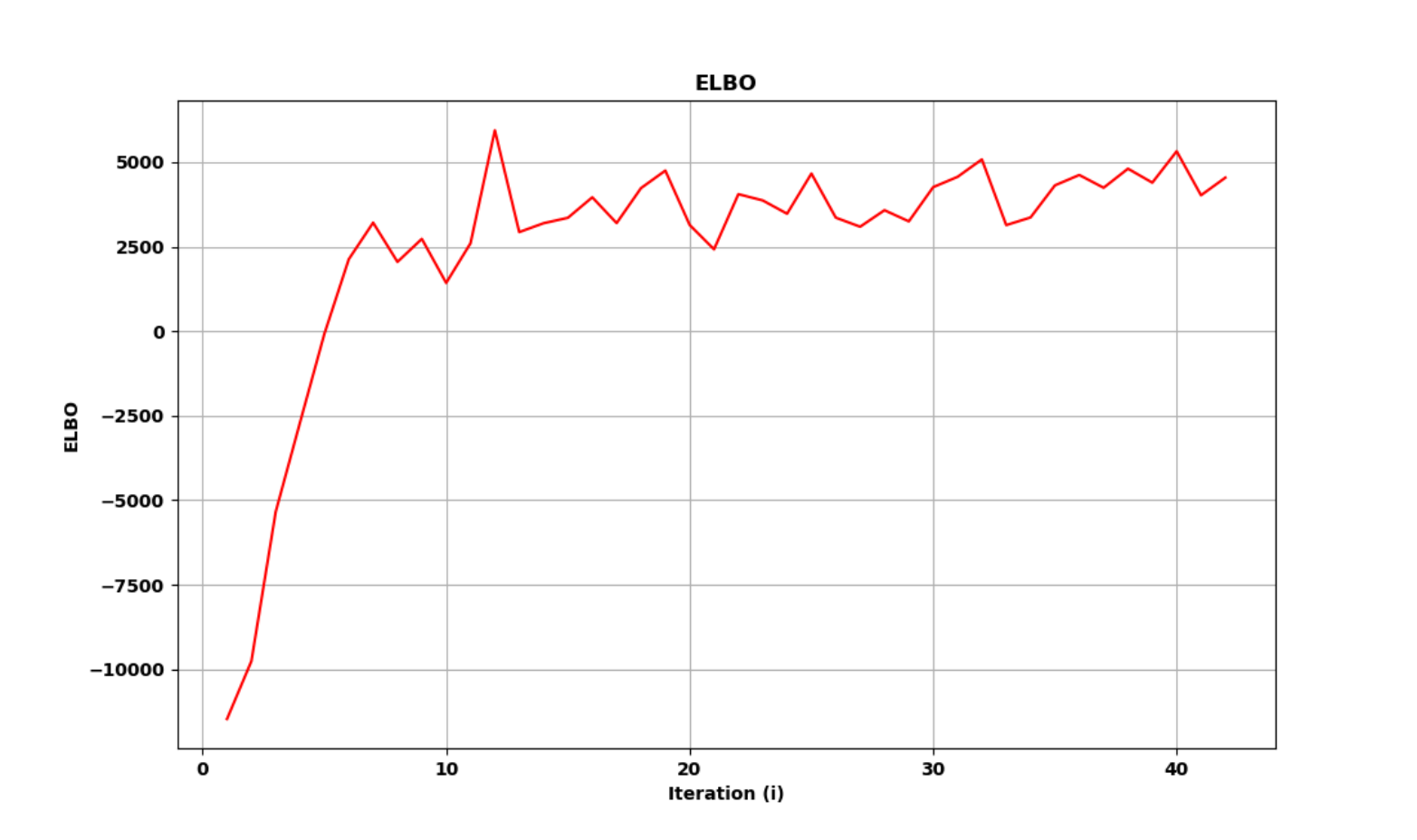
Simulate 10 response variables from multi-variate Normal

Associations:
- 8 microbiome species associated with 5 responses each
- sex, alchol, diet associated with 10, 8, 4 responses

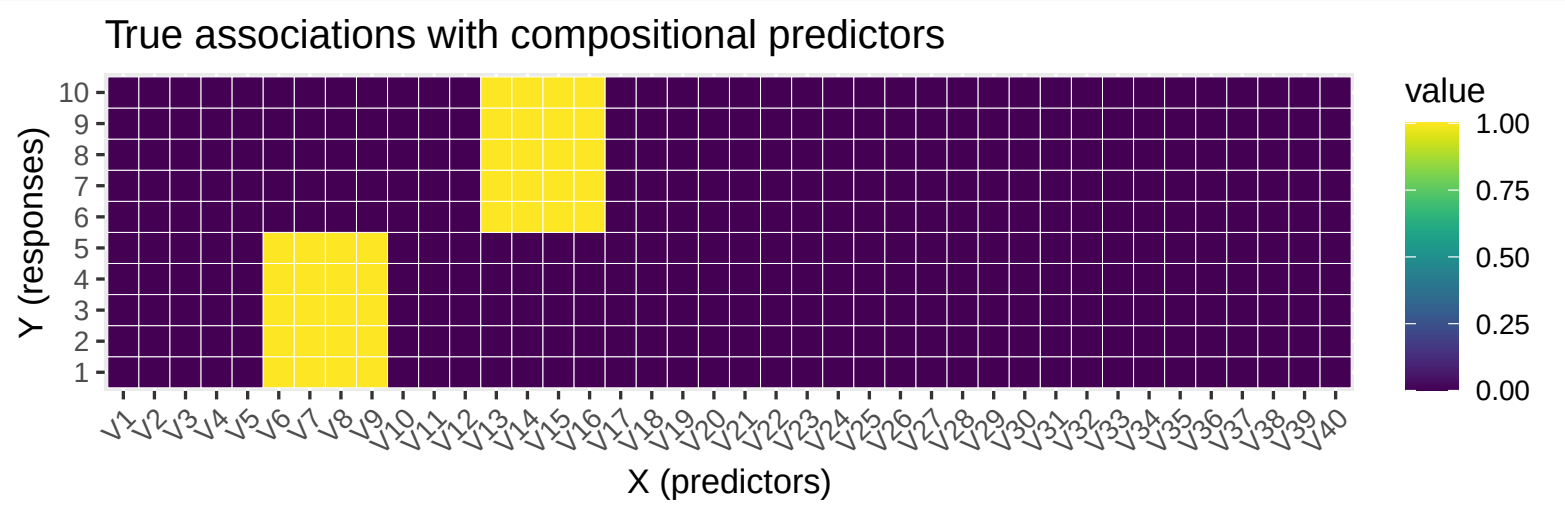
Simulation Study Parameters:
- SNR $= 2$
- residual correlation $= 0.3$
- residual dependence block diagonal

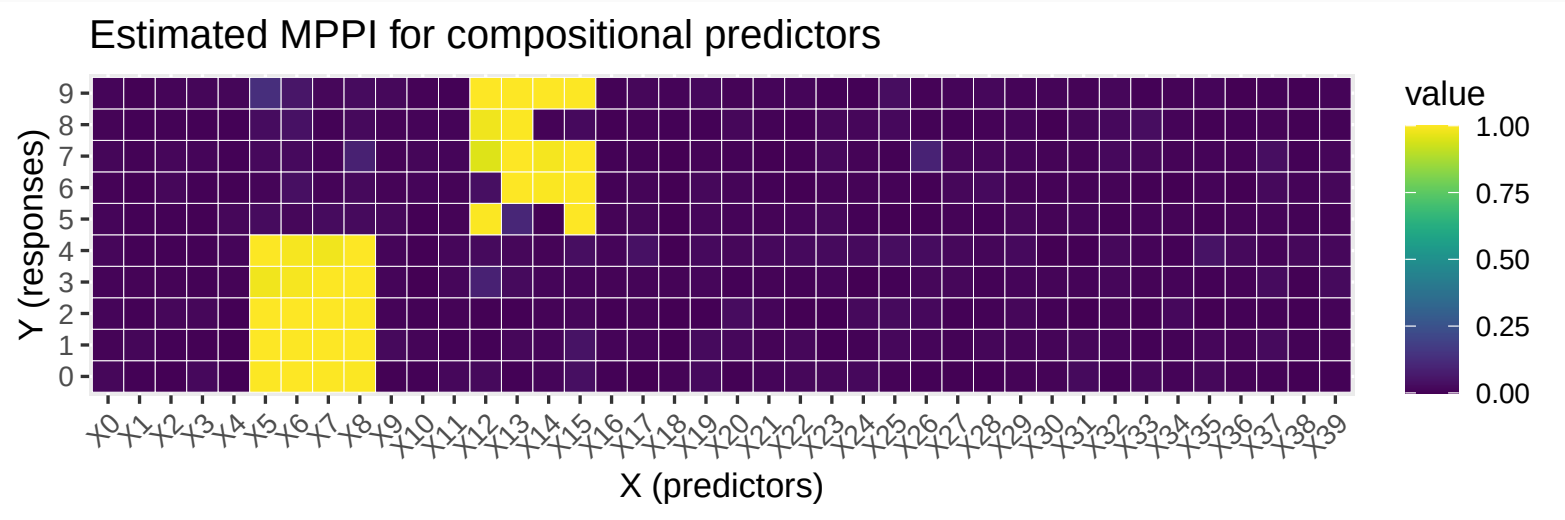
Results

Monitoring convergence of Variational updates:

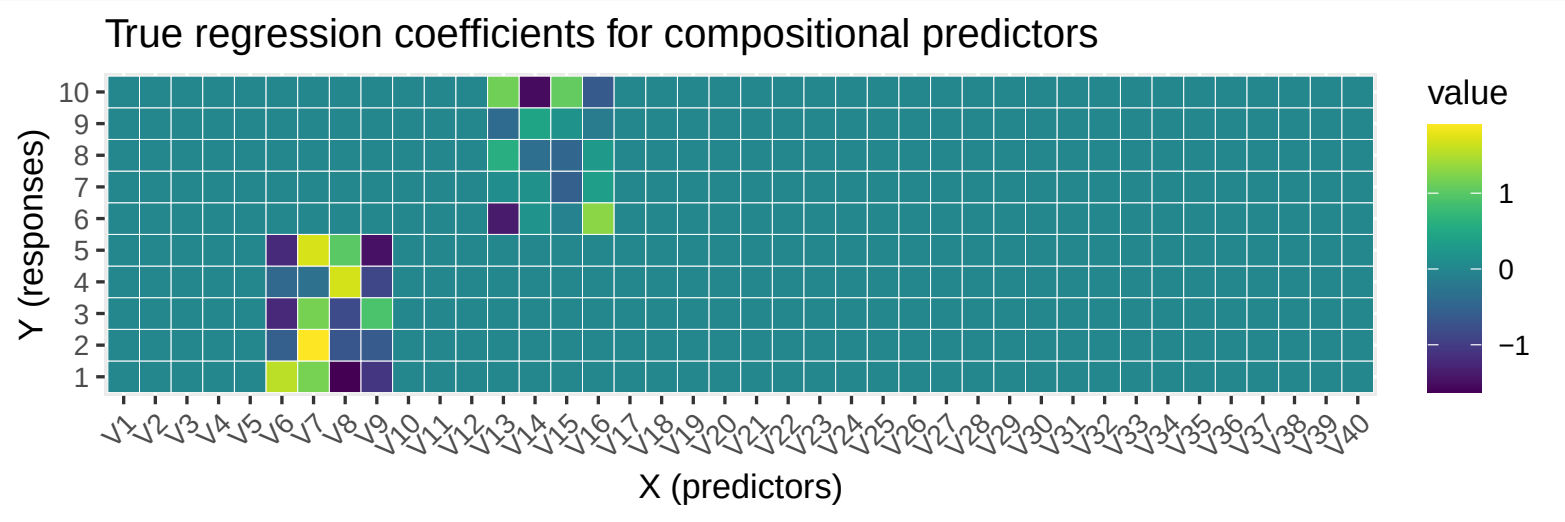


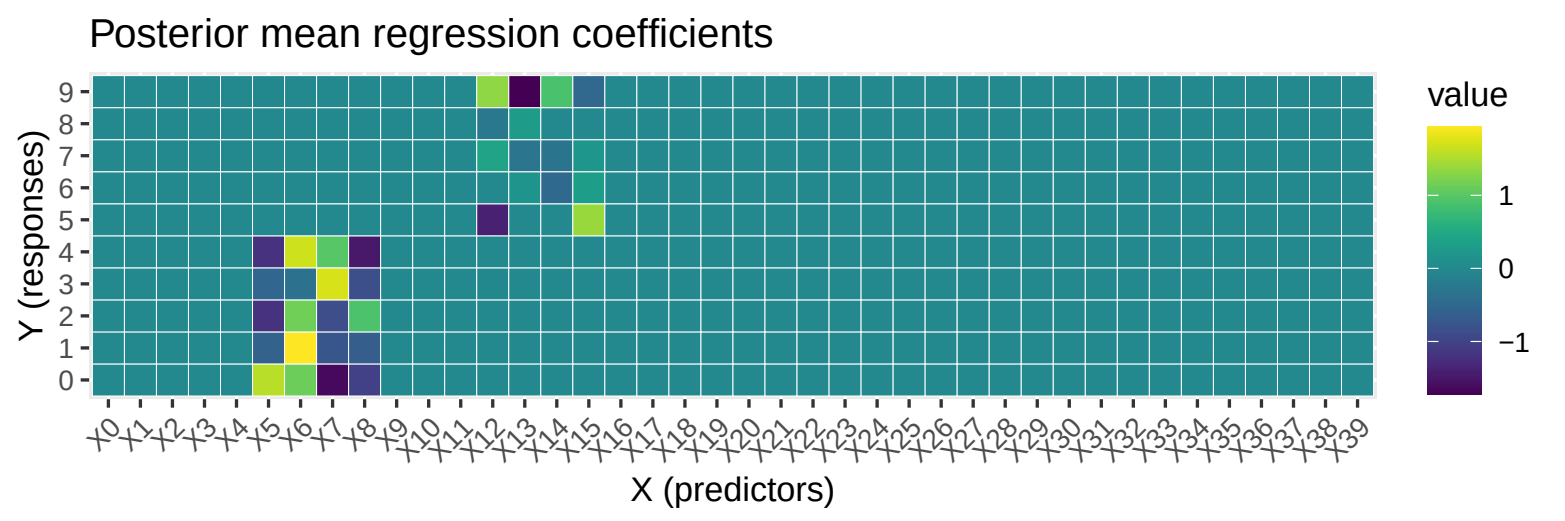
Results


True associations with compositional predictors
Y (responses)
X (predictors)
value
1.00
0.75
0.50
0.25
0.00


Estimated MPPI for compositional predictors
Y (responses)
X (predictors)
value
1.00
0.75
0.50
0.25
0.00

Results


True regression coefficients for compositional predictors
Y (responses)
10
9
8
7
6
5
4
3
2
1
V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38 V39 V40
X (predictors)
value
1
0
−1


Posterior mean regression coefficients
Y (responses)
9
8
7
6
5
4
3
2
1
0
X0 X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 X27 X28 X29 X30 X31 X32 X33 X34 X35 X36 X37 X38 X39
X (predictors)
value
1
0
−1

Results

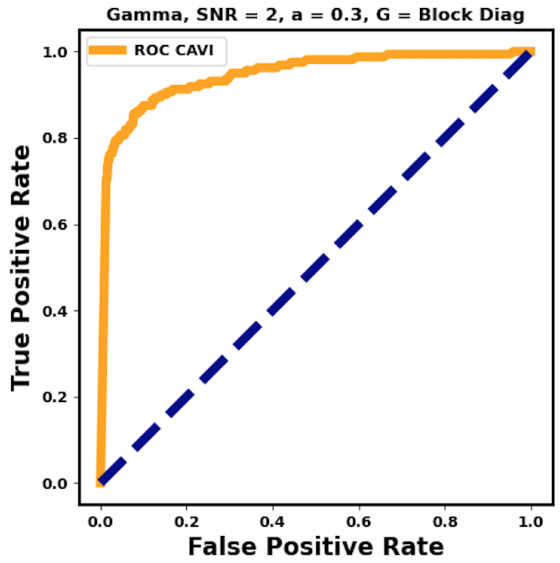
ROC curves averaged over 5 simulated data sets.

A: Sparse variable selection on non-compositional predictors

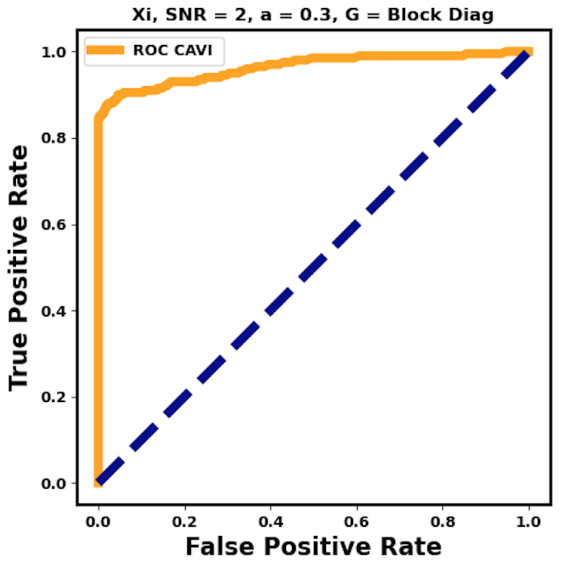
B: Sparse variable selection on compositional predictors

C: Sparse covariance selection

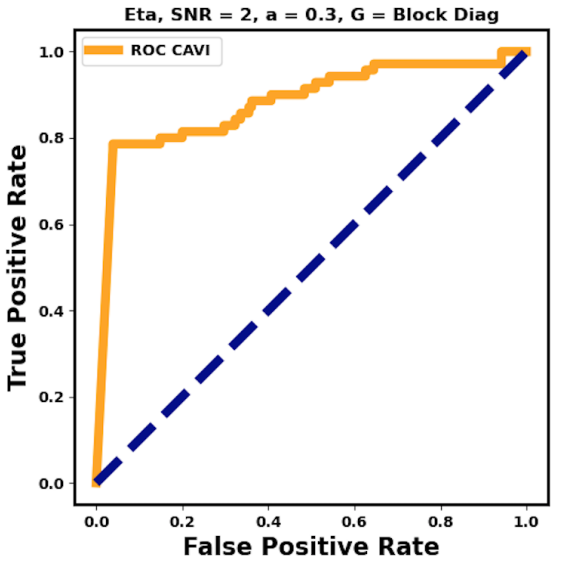
A



B



C



Summary

- Bayesian modelling of multivariate responses: sparse variable selection and covariance selection
- Bayesian model averaging provides flexible modelling of associations
- Variational Bayes to speed up computation for high-dimensional data
- Hybrid VB-MCMC to deal with compositional predictors

Acknowledgements

Thank you!

- Darren Scott
- Leonardo Bottolo
- Marco Banterle
- Sylvia Richardson

Funded by MRC LID Doctoral training grant, MRC Methodology Grant, DynaHealth Healthy Aging EU Horizon 2020 grant.

Variable and covariance selection:
Bottolo, Banterle, et al. doi: 10.1101/467019 on bioRxiv
Microbiome data transformation:
Lin et al. Biometrika 2013.
Hybrid VB-MCMC:
Ye et al. Statistics and Computing 2020.