# Sparse variable selection for high-dimensional Seemingly Unrelated Regression and Structural Equation Models

Alex Lewin and Marco Banterle

London School of Hygiene and Tropical Medicine

ISCB40 Leuven July 2019

# Northern Finland 1966 Birth Cohort (NFBC1966)

Population-based birth cohort

Recruitment:

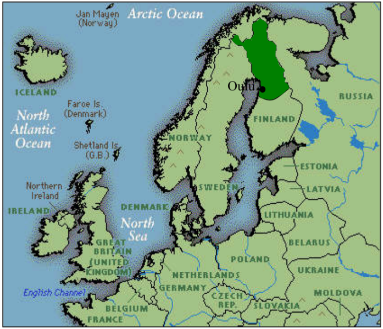
Pregnant mothers living in the provinces of Oulu and Lapland

Study population:

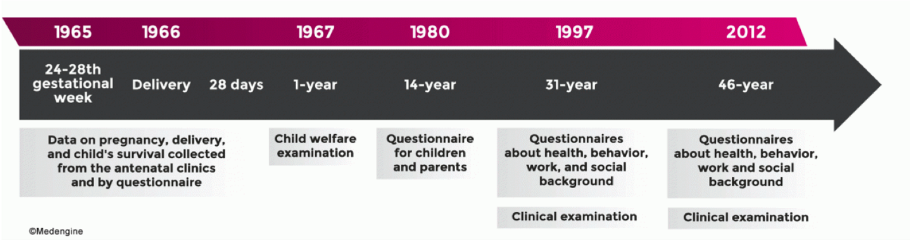
12,055 mothers with expected dates of delivery for year 1966;

12,058 alive born offsprings

96% of all births in the area



# Northern Finland 1966 Birth Cohort (NFBC1966)



1965 — 24-28th gestational week

1966 — Delivery — 28 days

Data on pregnancy, delivery, and child's survival collected from the antenatal clinics and by questionnaire

1967 — 1-year — Child welfare examination

1980 — 14-year — Questionnaire for children and parents

1997 — 31-year — Questionnaires about health, behavior, work, and social background — Clinical examination

2012 — 46-year — Questionnaires about health, behavior, work and social background — Clinical examination

©Medengine

Our focus:

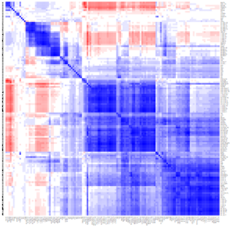
- 31 year collection: blood samples ⟶ metabolites, DNA

# Seemingly Unrelated Regressions: mQTL discovery in the NFBC66 study

- Question of interest is the discovery of genetic markers associated with metabolite regulation of lipids

- After quality control,
  n = 5154 people
  q = 158 metabolites
  p = 9310 SNPs on chromosome 16

- These responses are highly structured, with strong correlations



# Bayesian Seemingly Unrelated Regressions Model

Frame the problem as a multivariate linear regression model:

$$\underset{n\times q}{Y} = \underset{n\times p}{X}\,\underset{p\times q}{B} + \underset{n\times p}{E}$$

or equivalently:

$$Y \sim \mathcal{MN}(XB, \mathbb{I}_n, C)$$

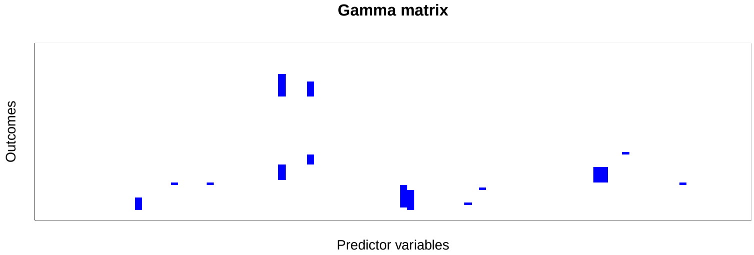
- Sparse variable selection on associations ($B$)
- Sparse covariance selection ($C$)

- Estimate using MCMC
- Provides the posterior probability of association for each predictor and each response (model averaging).

Variable selection performed through binary matrix $\Gamma$ ($p \times q$)

$$\gamma_{jk} = \begin{cases} 1 & \implies B_{jk} \neq 0 \\ 0 & \implies B_{jk} = 0 \end{cases}$$

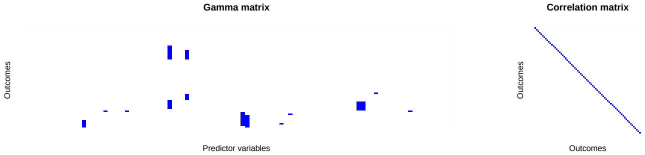
Sparsity prior $\gamma_{jk} \sim Bern(\omega_{jk}), \qquad \omega_{jk} \sim Beta()$



Predictor $X_j$ only appears in a regression if $\gamma_{jk}$ is 1.

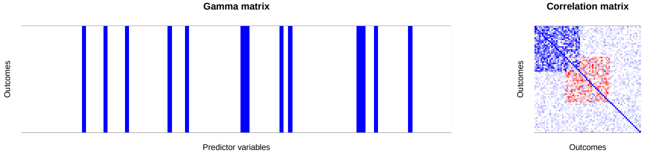
# Previous work in Bayesian multivariate regression

- Either assume diagonal covariance matrix



*Bottolo L, Chadeau-Hyam, M et al. (2013)*
*Lewin A et al. (2015)*

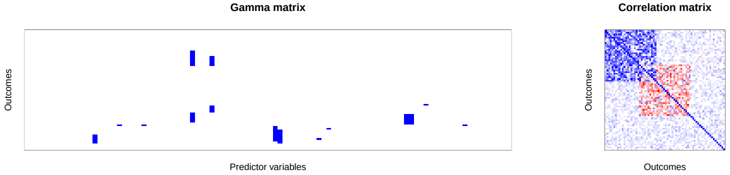
- Or assume all responses related to the same set of predictors



*Bottolo L, Petretto, E et al. (2011)*
*Bhadra A and Mallick BK (2013)*

# Our work on SUR model

- Full selection matrix $\Gamma$; Full covariance matrix $R$



Formulate as a Seemingly Unrelated Regressions (SUR) model:

$$\underset{n\times 1}{\boldsymbol{y}_k} = \underset{n\times d_k}{X_{\gamma_k}} \; \underset{d_k\times 1}{\boldsymbol{\beta}_{\gamma_k}} + \underset{n\times 1}{\boldsymbol{\epsilon}_k} \qquad \text{for } k = 1, \cdots, q$$

$Cov[\epsilon_k \epsilon_l] = C_{kl} \neq 0 \implies$ Outcomes do not naturally separate as in previous hierarchical model.

In both “previous” cases, models are conjugate in $B$ and $C$
$\longrightarrow$ only $\Gamma$ (variable selection) are updated.

- In the SUR model, Standard priors (Normal, Inverse Wishart) $\longrightarrow$ Not Conjugate in $B$ or $C$
- Can calculate posterior full conditionals for $\boldsymbol{\beta}_k$ and $C \rightarrow$ Gibbs sampler for $\gamma_k, \boldsymbol{\beta}_k$ and $C$.
- However, computationally intensive if use naive updates.

- Transform $C \longrightarrow \{\boldsymbol{\rho}_k, \sigma_k^2 : k = 1, \cdots, q\}$
- Factorise priors across the $q$ response variables: $C \sim \mathcal{IW}(\nu, M)$ becomes $\prod_{k=1}^{q} \mathcal{N}(\boldsymbol{\rho}_k | \sigma_k^2, M) \times \mathcal{IG}(\sigma_k^2 | \nu, M)$
- So posterior conditionals factorise also:

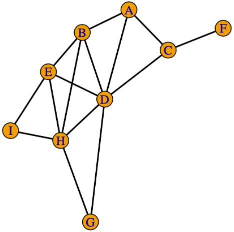
$$\prod_{k=1}^{q} \mathcal{N}(\boldsymbol{\rho}_k | \sigma_k^2, M, X, Y, B, \Gamma) \times \mathcal{IG}(\sigma_k^2 | \nu, M, X, Y, B, \Gamma)$$

So MCMC updates for $C$ parameters factorise over responses.
→ feasible computation for omics data

# Sparse covariance selection

Replace IW prior by Hyper-IW prior conditional on a sparse graph.

Decomposable (chordal or triangulated) graph: $C \sim \mathcal{HIW}_G(\nu, M)$



- Sparse prior on graph $G$ (Binomial on number of edges)
- Retain simple Normal and Inverse Wishart priors on $\boldsymbol{\rho}_k$ and $\sigma_k^2$.
- Sparsity leads to another computational gain (only non-zero $\rho_{kl}$).

# Bayesian Model Averaging

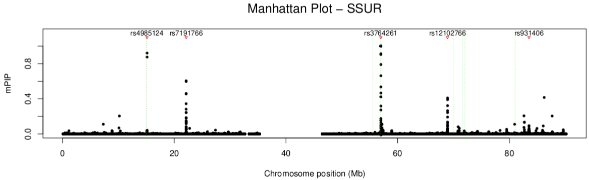
Marginal inclusion probabilities for covariate selection:

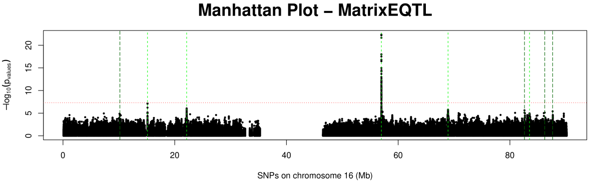
$$P(\gamma_{jk} = 1 \mid \textsf{data}) = \frac{1}{N_{iter}} \sum_{t=1}^{N_{iter}} \gamma_{jk}^{(t)}$$

Marginal edge inclusion probabilities for graph estimation:

$$P(\varepsilon_{kl} = 1 \mid \textsf{data}) = \frac{1}{N_{iter}} \sum_{t=1}^{N_{iter}} \varepsilon_{kl}^{(t)}$$

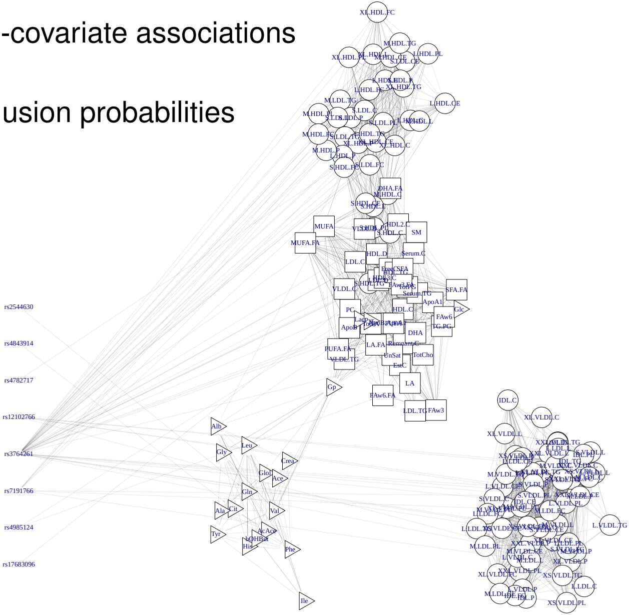
# $\Gamma$ response-covariate associations





- Only 1 SNP detected using standard GWAS univariate analysis
- 2 SNPs near to other SNPs that have been previously reported
- 1 SNP not previously reported, but univariate analysis shows “suggestive” evidence

# $\Gamma$ response-covariate associations and $G$ edge inclusion probabilities



# Extension to Structural Equation Models

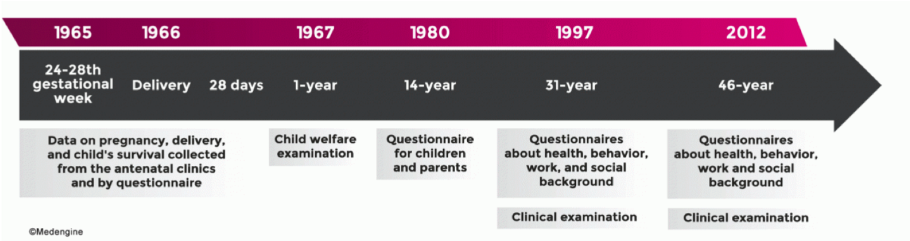
SUR model is $X \longrightarrow Y$ (link two blocks of variables)

SEM model: multiple blocks (Directed Acyclic Graph)

$$\begin{aligned} X \longrightarrow Y_1 &\longrightarrow Y_2 \\ X &\longrightarrow Y_2 \\ &\ldots \end{aligned}$$

- Multivariate regression model linking pairs of blocks
- Variable selection for each set of input variables

# Northern Finland 1966 Birth Cohort (NFBC1966)



1965 — 24-28th gestational week
1966 — Delivery; 28 days

Data on pregnancy, delivery, and child's survival collected from the antenatal clinics and by questionnaire

1967 — 1-year

Child welfare examination

1980 — 14-year

Questionnaire for children and parents

1997 — 31-year

Questionnaires about health, behavior, work, and social background

Clinical examination

2012 — 46-year

Questionnaires about health, behavior, work and social background

Clinical examination

©Medengine

Our focus:

- Maternal background and pregnancy data at 24-28 weeks
- Genetic variants for BMI
- Early growth parameters from follow-ups during childhood

# Blocks of variables: small data set example

For 3-stage model we use 4 blocks (29 variables).

$X$ = exogenous, $Y$ = endogenous.

$\underline{X_{prenatal} / X_{birth}}$

socio-economic variables (7)

maternal variables (12)

polygenic risk score

$\underline{X_{common}}$

sex

$\underline{Y_{infancy}}$

growth parameters (4)

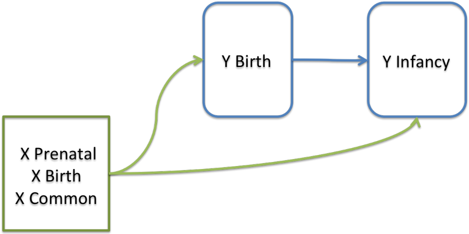
$\underline{Y_{birth}}$

gestational age

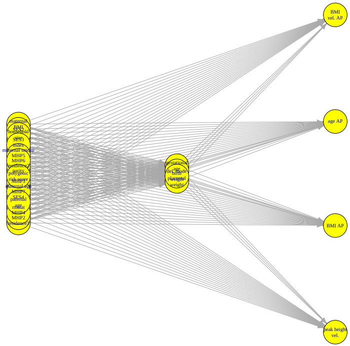
placental weight

delivery mode

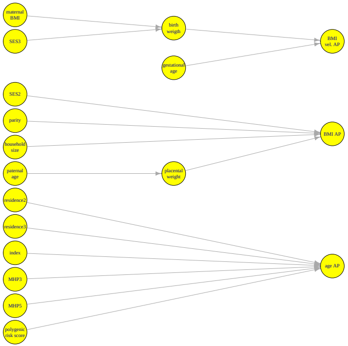
birth weight



## Input graph (variables)

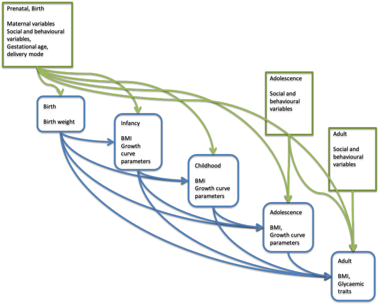


## Output graph between (variables)



Edge included if Marginal Posterior Inclusion Probability $>$ 0.5.

# Six life stages model



Prenatal, Birth

Maternal variables
Social and behavioural variables,
Gestational age,
delivery mode

Birth

Birth weight

Infancy

BMI
Growth curve parameters

Childhood

BMI
Growth curve parameters

Adolescence

Social and behavioural variables

Adolescence

BMI,
Growth curve parameters

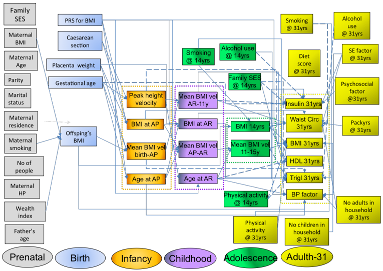
Adult

Social and behavioural variables

Adult

BMI,
Glycaemic traits

Input model: each arrow represents multiple associations ($\sim 500$ total).



Family SES

Maternal BMI

Maternal Age

Parity

Marital status

Maternal residence

Maternal smoking

No of people

Maternal HP

Wealth index

Father's age

PRS for BMI

Caesarean section

Placenta weight

Gestational age

Offsping's BMI

Peak height velocity

BMI at AP

Mean BMI vel birth-AP

Age at AP

Mean BMI vel AR-11y

BMI at AR

Mean BMI vel AP-AR

Age at AR

Smoking @ 14yrs

Alcohol use @ 14yrs

Family SES @ 14yrs

BMI 14yrs

Mean BMI vel 11-15y

Physical activity @ 14yrs

Smoking @ 31yrs

Alcohol use @ 31yrs

Diet score @ 31yrs

SE factor @ 31yrs

Psychosocial factor @31yrs

Insulin 31yrs

Waist Circ 31yrs

Packyrs @ 31yrs

BMI 31yrs

HDL 31yrs

Trigl 31yrs

BP factor

No adults in household @ 31yrs

Physical activity @ 31yrs

No children in household @ 31yrs

Prenatal

Birth

Infancy

Childhood

Adolescence

Adulth-31

Output model includes $\sim 40$ associations.

# Summary

- Bayesian SUR and SEM models with sparsity priors to perform variable selection for multiple responses.
- Modelling sparsity in the residual covariance matrix aids computations and increases the accuracy of the variable selection
- Bayesian modelling averaging framework gives robust results; can go further with joint modelling

## Thank you:

- Marco Banterle
- Sylvia Richardson
- Leonardo Bottolo
- Zhi (George) Zhao
- Manuela Zucknick
- Lia Tzala
- Marjo-Riitta Jarvelin

## R package:

`https://github.com/mbant/BayesSUR`

## Papers:

*Banterle M, Bottolo L, Richardson S, Ala-Korpela M, Jarvelin M-R and Lewin A (2018)*
Sparse variable and covariance selection for high-dimensional seemingly unrelated Bayesian regression, **BioRxiv preprint**
*Banterle, Zhao, Bottolo, Richardson, Zucknick and Lewin (2019)*
BayesSUR: An R package for high-dimensional multivariate Bayesian variable and covariance selection in regression, **submitted to Journal of Statistical Software Special Issue on Software for Bayesian Statistics**

Bayes SEM available soon!