Sparse variable and covariance selection for high-dimensional seemingly unrelated Bayesian regression

M. Banterle[1], L. Bottolo[2,3,4], S. Richardson[4,3],
M. Ala-Korpela[5,6,7,8,9,10], M-R. Järvelin[11,12,13,14,15], A. Lewin[1]

[1] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK
[2] Department of Medical Genetics, Cambridge Biomedical Campus, Cambridge, UK
[3] The Alan Turing Institute, London, UK
[4] MRC Biostatistics Unit, Cambridge Biomedical Campus, Cambridge, UK
[5] Systems Epidemiology, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia
[6] Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu, Oulu, Finland
[7] NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland
[8] Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK
[9] Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol, UK
[10] Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, The Alfred Hospital, Monash University, Melbourne, Victoria, Australia
[11] Center for Life Course Health Research, P.O. Box 5000, 90014 University of Oulu, Oulu, Finland
[12] Biocenter Oulu, University of Oulu, P.O. Box 5000, 90014 Oulu, Finland
[13] Department of Epidemiology and Biostatistics, Imperial College London, St. Mary's Campus, London, UK
[14] MRC-PHE Centre for Environment and Health, Imperial College London, St Mary's Campus, London, UK
[15] Department of Life Sciences, College of Health and Life Sciences, Brunel University London, Uxbridge, UK

Abstract

High-throughput technology for molecular biomarkers is increasingly producing multivariate phenotype data exhibiting strong correlation structures. Existing approaches for combining such data with genetic variants for multivariate Quantitative Trait Loci analysis generally either ignore correlation structure or make other restrictive assumptions about the associations between phenotypes and genetic loci.

We present a Bayesian Variable Selection (BVS) model with sparse variable and covariance selection for high-dimensional seemingly unrelated regressions. The model includes a matrix of binary variable selection indicators for multivariate regression, thus allowing different phenotype responses to be associated with different genetic predictors (a seemingly unrelated regressions framework). A general covariance structure is allowed for the residuals relating to the conditional dependencies between phenotype variables. The covariance structure may be dense (unrestricted) or sparse, with a graphical modelling prior. The graphical structure amongst the multivariate responses can be estimated as part of the model.

To achieve feasible computation of the large and complex model space, we exploit a factorisation of the covariance matrix parameter to enable faster computation using Markov Chain Monte Carlo (MCMC) methods. We are able to infer associations with thousands of candidate predictors multivariately on hundreds of responses.

We illustrate the model using a dataset of 158 NMR spectroscopy measured metabolites and over 9000 Single Nucleotide Polymorphisms on chromosome 16, measured in a cohort of more than 5000 people.

1 Introduction

Integrating multiple high-dimensional molecular biomarker data sets is a fundamental problem in genetic epidemiology and bioinformatics, in the search for molecular mediating mechanisms of the effects of genetic variants on ultimate clinical phenotypes. An important part of the analysis problem is to find associations between a set of genetic variants and downstream molecular phenotypes such as gene expression, proteomics, metabolomics or epigenetic data.

Since molecular phenotypes are assumed to be downstream of genetic variants in the causal pathway, the usual analysis framework is a regression model of phenotype on genotype. In the simplest analysis, multiple univariate regressions are performed for each phenotype-genotype pair, needing post-hoc adjustment for multiple comparisons and ignoring any correlations between genotypes and between phenotypes. Univariate approaches are unlikely to be the best strategy, considering that data where latent structures induce high level of correlation are becoming increasingly available to analyse simultaneously, for example serum metabolomic profiles (Kettunen, Tukiainen, et al., 2012; Soininen, Kangas, Würtz, Tukiainen, et al., 2009; Suhre et al., 2011), 3D-imaging (Hammond and Suttie, 2012; Wang et al., 2013), and gene expressions (Consortium, 2015; Ackermann, Sikora-Wohlfeld, and Beyer, 2013). A recent comparison study showed that utilising multivariate quantitative phenotypes increases the statistical power in the association tests compared to univariate analysis Inouye et al. (2012).

There is an extensive literature in both classical and Bayesian frameworks on univariate multiple regression, where each phenotype, e.g. each metabolite, is regressed separately on all genetic variants. This approach enables the analyst to select a small set of predictive variants, and to find patterns of genetic variants associated with each individual molecular phenotype. Bayesian models may use shrinkage priors (Park and Casella, 2008; Griffin and Brown, 2012) to encourage the majority of regression coefficients to be shrunk to very small values. This approach requires a threshold to be set on the coefficients to select associations. Global-local shrinkage priors have been developed to overcome problems with too heavy shrinkage (Polson and Scott, 2010). An alternative approach (Bayesian variable selection or BVS) is to explicitly select a small set of associations, using auxiliary latent variables (George and McCulloch, 1993; Kuo and Mallick, 1998; Clyde, 1999); this approach produces posterior probabilities of variable inclusion, enabling probabilistic decisions about which sets of variables are associated with the outcome of interest.

Brown, Vannucci, and Fearn (1998) and Brown, Vannucci, and Fearn (2002) developed a general framework for BVS models for multivariate outcomes. For this multivariate model with variable selection on multiple predictors, the posterior space is large and complex. In applications of this model, two alternative simplifying assumptions have been made in order to exploit conjugacy with respect to regression coefficients and residual covariance. One such simplification is used by Petretto et al. (2010), who present a model allowing for residual covariance between response variables, in the context of eQTL (expression Quantitative Trait Loci) analysis. In order to ensure conjugacy in the model and enable feasible computation times in the increasingly high-dimensional *-omics* setting, they assume the same set of predictors is selected in the regression equations for every response variable. Bhadra and Mallick (2013) make this same assumption on the predictors, but go a step further and include sparse residual covariance selection between regressions, using graphical modelling based on decomposable graphs. In both these models, the regression coefficients and residual covariances are integrated out, resulting in a model search over the space of the vector of variable selection indicator variables and hyperparameters.

The alternative simplifying assumption is to assume conditional independence between regressions. In this scenario different predictors can be permitted to be associated with different response variables, thus this model contains a matrix of variable selection indicator variables relating responses and predictors. Jia and Xu (2007), Bottolo, Petretto, et al. (2011), and Scott-Boyer et al. (2012) all take this approach for eQTL analysis, using hierarchical sparsity priors on the probabilities of associations to share information across the gene expression outcomes. The assumption of residual independence of outcome variables ensures conjugacy of the regression coefficients and residual variances and enables exploration of the full posterior space of selection models.

The increasing use of the high-throughput platforms nuclear magnetic resonance (NMR) spectroscopy and liquid chromatography-mass spectrometry (LC-MS), is allowing the simultaneous quantification of a progressively big number of biologically complex phenotypes, such as metabolomics and proteomics (Dunn et al., 2011; Gika et al., 2014). These molecular phenotypes show particularly strong correlation structures in comparison with transcriptomics data used in eQTL analyses, for which most BVS models have been developed. Therefore the need to not only account for correlation to better reveal associations with genetic profiles, but also to infer their dependence structure, is becoming increasingly important to improve our understanding of integrated functioning of living organisms and has raised considerable interest in the scientific literature Dumas et al. (2007), Cichonska et al. (2016), and Rodriguez-Martinez et al. (2017). Suitable statistical tools which can tackle these problems while retaining computational feasibility are thus necessary.

The direction of this work is, accordingly, Quantitative Trait Locus mapping of correlated phenotypes. In particular, we focus on metabolomics Quantitative Trait Locus (mQTL), a powerful approach typically used to identify genes associated with metabolic markers of diseases, where the multivariate response is generally of moderate size (in the order of hundreds of metabolites) and where both the assumption of common predictors and the assumption of independent residuals are likely to be unreasonable. We develop a fully Bayesian variable selection model which allows both for different predictors to be associated with different outcomes and for the outcomes to be residually correlated. We present two versions of the model, one with a dense covariance matrix and one with sparse covariance selection. Each instance of the multivariate regression model has the form of a Seemingly Unrelated Regressions (SUR) model. Since the model loses the conjugacy of the earlier models, we exploit a factorisation of the covariance matrix parameter to enable faster computation using Markov Chain Monte Carlo (MCMC) methods. We are able to infer associations with thousands of candidate predictors multivariately on all the responses. In the case of sparse covariance, we also simultaneously infer an underlying graph detailing the conditional independence relationships between hundreds of response variables.

To illustrate our approach, we analyse a dataset from the Northern Finland Birth Cohort 1966 (NFBC66) 31-year follow-up, where the model developed here is used to uncover associations between 158 NMR spectroscopy measured metabolites and genome data; we will focus on directly genotyped Single Nucleotide Polymorphisms (SNPs) on chromosome 16, which results in over 9000 covariates and a sample size of more than 5000 people. The metabolites set comprises lipoprotein particle concentrations, low molecular weight metabolites such as amino acids, 3-hydroxybutyrate and creatinine, and different serum lipids, including free and esterified cholesterol, sphingomyelin and fatty acid saturation. These data have been observed to exhibit strong residual correlation (Marttinen et al., 2014; Kettunen, Tukiainen, et al., 2012), even after accounting for the variance explained by all reported SNPs.

Section 2 starts by introducing the general matrix Normal regression model and briefly reviews the variable selection prior used in this work. The crucial reparametrisation that allows us to factorise the

likelihood is then presented alongside the relative priors in the transformed space. Section 3 provides the posterior computations for the parameters and details the MCMC used to estimate them. Finally in Section 5 the method is first validated against simulated data and then applied to the NFBC66 dataset.

2 Model

The model can be seen as a set of regressions for multivariate response $Y = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_s)$, $\boldsymbol{y}_k = (y_{1k}, \dots, y_{nk})^t$, for $k = 1, \dots, s$, and corresponding covariate matrices X_k with dimensions $n \times p_k$. The same set of covariates may be used for all regressions, but this is not necessary. Variable selection is performed on the covariates using binary indicators $\boldsymbol{\gamma}_k = (\gamma_{k1}, \dots, \gamma_{kp_k})^t$ where γ_{kj} is 1 if covariate j is included in the regression for response k, and 0 if not. We use the shorthand notation $X_{\boldsymbol{\gamma}_k}$ for the columns of X_k selected by the vector $\boldsymbol{\gamma}_k$, and similar for $\boldsymbol{\beta}_{\boldsymbol{\gamma}_k}$. Thus we can write the set of linked regressions as

$$\boldsymbol{y}_k = X_{\gamma_k}\boldsymbol{\beta}_{\gamma_k} + \boldsymbol{u}_k \quad k = 1, \dots, s \tag{1}$$

but most importantly the residuals will be correlated: $u_i = (u_{i1}, \dots, u_{is}) \sim \mathcal{N}(\mathbf{0}, C)$.

We can also write the likelihood for this model as

$$\mathcal{Y} \mid \mathcal{X}, \{\beta_1, \dots, \beta_s\}, \{\gamma_1, \dots, \gamma_s\}, C \sim \mathcal{N}(\mathcal{X}_\gamma \beta_\gamma, C \otimes \mathbb{I}_n) \tag{2}$$

where $\mathcal{Y} = vec(Y)$, $vec()$ being the vectorisation operator, $\beta_\gamma = vec(\boldsymbol{\beta}_{\gamma_1}, \dots, \boldsymbol{\beta}_{\gamma_s})$ and $\mathcal{X}_\gamma$ is a block diagonal matrix with X_{γ_k} as the kth diagonal element.

2.1 Variable Selection Priors

As mentioned above, we want to select only a subset of the covariates for each outcome; the state-of-the-art prior for this purpose is a Normal spike-and-slab prior (introduced by Brown, Vannucci, and Fearn (1998)) for the regression coefficients where a point-mass on zero is associated with each $\gamma_{kj} = 0$, while β_{kj} follows a diffuse Normal distribution otherwise; calling $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = \{\beta_{kj} | k \in 1, \dots s, j \in 1, \dots p_k, \gamma_{kj} = 1\}$

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \boldsymbol{\gamma} \sim \mathcal{N}\left(0, wV_\gamma\right) \tag{3}$$

Here w can be thought as a shrinking coefficient and $V_{\boldsymbol{\gamma}}$ is a general prior covariance matrix with dimension dictated by the number of selected coefficients in $\boldsymbol{\gamma}$.

Since we assume one common w scaling for all the selected regression coefficients, this prior makes sense only if all the Xs have also a single common scaling, *i.e.* if they represent comparable quantities and/or if they've been standardised. A simple generalisation would be to allow for a vector $\boldsymbol{w}$ of shrinkage coefficients as in Bae and Mallick (2004).

Several possibilities exist for $V_{\boldsymbol{\gamma}}$, the simplest being the one where $V_{\boldsymbol{\gamma}} = \mathbb{I}_{\boldsymbol{\gamma}}$; in this case the values of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ are independent *a priori* and only w controls their variance. Another common choice is the g-prior, first proposed in Zellner (1986); in the case of correlated residuals this would imply $V_{\boldsymbol{\gamma}} = \left(\mathcal{X}_{\boldsymbol{\gamma}}^t \left(C \otimes \mathbb{I}_n\right)^{-1} \mathcal{X}_{\boldsymbol{\gamma}}\right)^{-1}$.

We will in general centre and rescale the covariates X, the automatic scaling feature of the g-prior

(Chipman et al., 2001) is of little interest and so are the computational advantages for the marginal likelihood since this is not analytically available in the Seemingly Unrelated Regression framework. In the following we thus use *a priori* independent regression coefficients, but any of the more complex prior only requires minor modifications to the expressions given.

To retain some flexibility we further specify an Inverse Gamma hyper-prior for w rather than fixing it to some arbitrary value.

$$w \sim IG(a_w, b_w) \tag{4}$$

The values of a_w and b_w can be chosen depending on the application.

Shifting the focus on the prior for $\boldsymbol{\gamma}$, a number of sparsity inducing priors have been used in the literature, the most commons being an independent Bernoulli prior for each binary γ_{kj} coefficient or a Binomial distribution on the model size; the desired prior sparisty level can thus be adjusted by simply tuning one marginal *prior* inclusion probability parameter.

Another successful choice in eQTL analysis is the *hotspot detection* prior used in Bottolo, Petretto, et al. (2011) where each inclusion probability is decomposed into a marginal effect for each response that regulate the sparsity level and a marginal predisposition for each predictor to be associated with a relevant number of outcomes. Because this matches with our motivating example, we will use the *hotspot* prior, details of which are given in Section 5.

2.2 Factorisation of the likelihood

As mentioned in the Introduction, we can only integrate out analytically both β and C by either assuming a diagonal C or the same $\boldsymbol{\gamma}_k$ for all k, and conjugate priors on C and β, giving clear computational advantages (Petretto et al., 2010; Bhadra and Mallick, 2013). In the more general case, the usual priors on these parameters lose conjugacy and can't be integrated out. The full conditionals retain their simple forms, so it is straightforward to write a Gibbs sampler to estimate the posterior distributions (eg Holmes, Denison, and Mallick (2002)); the computational time needed is however prohibitive for most high-dimensional settings.

To overcome this issue we factorise the covariance matrix C iteratively as

$$C_{(k)} = \begin{pmatrix} C_{(k-1)} & \boldsymbol{c}_k \\ \boldsymbol{c}_k^t & c_k \end{pmatrix} \tag{5}$$

for all $k = 2, \cdots, s$, with $C_{(s)} = C$ and $C_{(1)} = c_1$, and $\boldsymbol{c}_1$ is null. Thus each $C_{(k)}$ is the marginal covariance matrix for responses $1, \cdots, k$, c_k is the variance of response k, and $\boldsymbol{c}_k$ is the vector of covariances between response k and responses $\{1, \ldots, k-1\}$.

With this decomposition the likelihood can be factorised as

$$\begin{aligned} p(\mathcal{Y} \mid \mathcal{X}, \beta, \gamma, C) &= \prod_{i=1}^{n} p(\boldsymbol{y}_i \mid \boldsymbol{\mu}_i, C) && (6) \\ &= \prod_{i=1}^{n} \prod_{k=1}^{s} p(y_{ik} \mid \boldsymbol{y}_{i(k-1)}, \boldsymbol{\mu}_{i(k)}, C_{(k)}) && (7) \\ &= \prod_{i=1}^{n} \prod_{k=1}^{s} \varphi(y_{ik} \mid \mu_{ik} + \boldsymbol{c}_k^t C_{(k-1)}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)_{(k-1)}, c_k - \boldsymbol{c}_k^t C_{(k-1)}^{-1} \boldsymbol{c}_k) && (8) \end{aligned}$$

where $\boldsymbol{\mu}_i$ is the vector of means with $\mu_{ik} = \boldsymbol{x}_{i\gamma_k}^t \boldsymbol{\beta}_{\gamma_k}$, the vector $\boldsymbol{y}_{i(k)}$ stands for the vector of responses $1, \cdots, k$ and similar for $\boldsymbol{\mu}_{i(k)}$, and $\varphi(\boldsymbol{y} \mid \boldsymbol{\mu}, \Sigma)$ is the pdf for the Normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Note that the joint distribution $p(\mathcal{Y} \mid \mathcal{X}, \beta, \gamma, C)$ is the same regardless of the order used for the decomposition since we are simply factorising it by chain-conditioning.

Thus defining new parameters

$$\sigma_1^2 \equiv c_1, \tag{9}$$

$$\left.\begin{aligned} \sigma_k^2 &\equiv c_k - \boldsymbol{c}_k^t C_{(k-1)}^{-1} \boldsymbol{c}_k \\ \boldsymbol{\rho}_k &\equiv C_{(k-1)}^{-1} \boldsymbol{c}_k \end{aligned}\right\} k = 2, \cdots, s \tag{10}$$

the likelihood becomes a product of conditionally independent regressions:

$$p(\mathcal{Y} \mid \mathcal{X}, \beta, \gamma, C) = \prod_{k=1}^{s} \varphi(\boldsymbol{y}_k \mid X_{\gamma_k} \boldsymbol{\beta}_{\gamma_k} + U_{(k-1)} \boldsymbol{\rho}_k, \sigma_k^2 \mathbb{I}_n) \tag{11}$$

defining for convenience the matrix $U_{(k-1)} = Y_{(k-1)} - (X_{\gamma_1}\boldsymbol{\beta}_{\gamma_1} \quad X_{\gamma_2}\boldsymbol{\beta}_{\gamma_2} \quad \cdots \quad X_{\gamma_{k-1}}\boldsymbol{\beta}_{\gamma_{k-1}})$ consisting of the first $k-1$ residuals from the original linked regressions.

From (11) is straightforward to see that σ_k^2 is the residual variance of response k conditioned on $U_{(k-1)}$ and $\boldsymbol{\rho}_k$ is a real valued $(k-1)$-vector of regression coefficients on the same $U_{(k-1)}$ residuals. The likelihood is thus decomposed into a product of independent (conditionally on the new parameters) factors over the outcomes and this opens up the possibility of parallelisation, which greatly increases the computational feasibility of this model.

This type of factorisation was first proposed in the graphical modelling literature by Wermuth (1980) for multivariate Normal data with zero means and has been used in these so called *covariance selection* models for computational convenience and for interpretability of the now unconstrained transformed off-diagonal elements of the covariance matrix, see for example Pourahmadi (1999) and Stingo and Marchetti (2014). Modelling these Cholesky factors is also popular in econometrics in both conditional and simultaneous autoregressive models (Datta, Banerjee, and Hodges, 2017). Wang (2012) use a similar idea in the context of sampling the parameters for the Bayesian Lasso, and Zellner and Ando (2010) used this reparametrisation for a Seemingly Unrelated Regressions model with fixed covariates, *i.e.* not resulting from variable selection, specifying a Jeffrey's prior on β, C.

We instead estimate our model completely in the reparametrised space, defining priors on the new parameters $\{\sigma_k^2, \boldsymbol{\rho}_k\}$ starting from a usual Normal-Inverse Wishart distribution.

Note that this factorisation is related to the Cholesky decomposition of C, and not of C^{-1} as for example in Roverato (2000), although as we will see there is a straightforward relation between elements in C^{-1} and the reparametrised $\boldsymbol{\rho}_k$. Although for factorising the likelihood any ordering of the variables can be used for the decomposition, in terms of prior distribution some ordering are better suited to specific sparsity patterns in the precision matrix (Vecchia, 1988; Guinness, 2018).

2.3 Prior for a Dense Covariance Matrix

To allow a general covariance matrix with correlated residuals but no assumed structure, we start from a standard Inverse Wishart prior $C \sim \mathcal{IW}(\nu, M)$. The priors in the transformed space are straightforward to calculate using the block decomposition of C. Since $C_{(k)}$ is a submatrix of C it also has an Inverse Wishart distribution. The new parameters are simply related to the block structure of

the inverse of this matrix, with σ_k^2 being the Schur complements of c_k in $C_{(k)}$, and $\boldsymbol{\rho}_k = C_{(k-1)}^{-1}\boldsymbol{c}_k$, hence their priors can be calculated using standard matrix properties of the Inverse Wishart (Dempster, 1969; Dawid, 1981; Eaton, 1983; Roverato, 2000).

Thus, decomposing M conformally with C into

$$M_{(k)} = \begin{pmatrix} M_{(k-1)} & \boldsymbol{m}_k \\ \boldsymbol{m}_k^t & m_k \end{pmatrix} \tag{12}$$

for $k = 2, \cdots, s$, we see that the priors on the changed variables are

$$\sigma_k^2 \sim IG((\nu - s + k)/2, (m_k - \boldsymbol{m}_k^t M_{(k-1)}^{-1}\boldsymbol{m}_k)/2) \tag{13}$$

$$\boldsymbol{\rho}_k|\sigma_k^2 \sim N(M_{(k-1)}^{-1}\boldsymbol{m}_k, \sigma_k^2 M_{(k-1)}^{-1}) \tag{14}$$

for $k = 2, \ldots, s$, and $\sigma_1^2 \sim \mathcal{IG}((\nu - s + 1)/2, m_1/2)$; see again Dempster (1969) and Dawid (1981) for formal proof of this result.

2.4 Prior for a Sparse Covariance Matrix

To model sparsity structure in the covariances we introduce the graph structure of the precision matrix as a parameter in the model, such that variables are conditionally independent if there is no direct link between them in the graph. In this section we use decomposable graphs to model sparsity, and show that the new parameters $\boldsymbol{\rho}_k$ are simply related to the graph structure and hence retain the useful computational properties for MCMC estimation of the SUR models.

Perfect Elimination Order and Notation

A separator of a graph G is a completely-connected subgraph of G which separates the graph 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 set of subgraphs which cannot be further decomposed form the prime components of the graph. If all the prime components are complete then the graph is said to be decomposable. In this case the prime components P_q can be ordered in such a way that for every $q > 1$ there exists $m < q$ such that

$$P_q \cap H_q \subset P_m \tag{15}$$

where $H_q = \cup_{l=1}^{q-1} P_l$ for $q = 2, \cdots, Q$. This is called the running intersection property Lauritzen (1996). With this notation, we also define the separators $S_q = P_q \cap H_q$ and residuals $R_q = P_q \setminus S_q$ for $q = 2, \cdots, Q$. The separators $\{S_2, \cdots, S_Q\}$ are not necessarily distinct from each other.

This (non-unique) ordering of the prime components is called a perfect ordering and has the property that nodes in residual R_q are independent of nodes in $H_q \setminus S_q$, conditionally on the rest of the graph, and therefore the elements of the precision matrix relating pairs of nodes in those sets are structurally equal to zero.

The so-called perfect elimination ordering (also non-unique, even for a given perfect ordering of prime components) will be denoted ξ. This is defined by taking the nodes in P_1 in any order, followed by the nodes in R_2 in any order, and so on for the rest of the residuals up to R_Q.

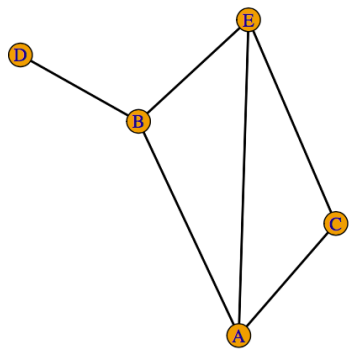

Figure 1: An toy decomposable graph, used to exemplify perfect ordering of its cliques and the perfect elimination ordering.

Figure 1 shows a toy example of a small decomposable graph. One possible perfect ordering of the prime components is $P_1 = \{A, B, E\}$, $P_2 = \{A, C, E\}$, $P_3 = \{B, D\}$, with separators $S_2 = \{A, E\}$ and $S_3 = \{B\}$. Given this ordering of the prime components, a possible perfect elimination ordering of the nodes is $\{A, B, E, C, D\}$.

With the variables ordered accordingly to the perfect elimination ordering we have that $\Lambda_{kl} = 0 \implies \rho_{kl} = 0$, where Λ is the precision matrix C^{-1}; this was proved by Paulsen, Power, and Smith (1989) (a short proof in our notation is also given in the Appendix). This means that the reparametrisation of the covariance matrix is a convenient way to encode the conditional independencies between variables.

In addition we see that conditionally on a given *sparse* graph, we will not need to estimate values of ρ_{kl} for many pairs of variables, allowing for computational savings. More specifically, given a particular decomposable graph structure, ρ_{kl} are only estimated for pairs of variables k, l contained in the same prime component of the graph.

Reparametrisation definition

Here we show further that the new parameters σ_k^2 and $\boldsymbol{\rho}_k$ are defined straightforwardly in terms of the covariance matrices of the prime components of the graph, and hence we easily calculate priors in the reparametrised space. Assuming the perfect elimination ordering, take any prime component P_q, and define variable k to be the final variable in the residual R_q. With this ordering of variables, $\boldsymbol{\rho}_k$ relates variable k to the set of all variables in prime components $1, \cdots, q$, which is the set of nodes in the subgraph H_{q+1}. Thus from the definition of the reparametrisation (shown in the Appendix),

$$\boldsymbol{\rho}_k = -\boldsymbol{\omega}_k \omega_k^{-1} \tag{16}$$

where $\boldsymbol{\omega}_k$ is the final column of $C^{-1}_{H_{q+1}}$ without the last element, and ω_k is the last diagonal entry of $C^{-1}_{H_{q+1}}$.

The marginal precision matrix for the subgraph H_{q+1} consisting of all prime components from 1 to q can be written

$$C^{-1}_{H_{q+1}} = \begin{pmatrix} \Omega_{H_q \setminus S_q} & \Omega_{H_q \setminus S_q, S_q} & 0 \\ \Omega^t_{H_q \setminus S_q, S_q} & \Omega_{S_q} & \Omega_{S_q, R_q} \\ 0 & \Omega^t_{S_q, R_q} & \Omega_{R_q} \end{pmatrix} \tag{17}$$

as shown in the Appendix, which directly show that the elements of $\boldsymbol{\rho}_k$ corresponding to variables in $H_q \setminus S_q$ are zero. A further block matrix calculation shows that

$$C_{P_q}^{-1} = \begin{pmatrix} \Omega_{S_q} - \Omega_{H_q,S_q}^t \Omega_{H_q \setminus S_q}^{-1} \Omega_{H_q,S_q} & \Omega_{S_q,R_q} \\ \Omega_{S_q,R_q}^t & \Omega_{R_q} \end{pmatrix} \tag{18}$$

where we highlight that the off-diagonal blocks in $C_{P_q}^{-1}$ appear unmodified also in $C_{H_{q+1}}^{-1}$, thanks to the zeros in the latter.

Now define $\boldsymbol{\rho}_{k,P_q}$ for the non-zero elements of $\boldsymbol{\rho}_k$. From the above equations we can see that this can be written in terms of the final column of $C_{P_q}^{-1}$ without the last element, and the last diagonal entry of $C_{P_q}^{-1}$. Hence again by block matrix manipulation we see that we can write

$$\boldsymbol{\rho}_{k,P_q} \equiv (C_{P_q}^{(k-1)})^{-1} \boldsymbol{c}_{k,P_q} \tag{19}$$

where $C_{P_q}^{(k-1)}$ is the sub-matrix of C_{P_q} with variable k removed, and $\boldsymbol{c}_{k,P_q}$ is the final column of C_{P_q} without the last element. This is the same as the original definition of $\boldsymbol{\rho}_k$, except that due to the structural zeros in the precision matrix we now only need to use the sub-covariance matrix C_{P_q} corresponding to the prime component P_q containing R_q. Then consider

$$\begin{aligned} \sigma_k^2 &= c_k - \boldsymbol{c}_k^t C_{(k-1)}^{-1} \boldsymbol{c}_k && (20) \\ &= c_k - \boldsymbol{\rho}_k^t C_{(k-1)} \boldsymbol{\rho}_k && (21) \\ &= c_k - \boldsymbol{\rho}_{k,P_q}^t C_{P_q}^{(k-1)} \boldsymbol{\rho}_{k,P_q} && (22) \\ &= c_k - \boldsymbol{c}_{k,P_q}^t (C_{P_q}^{(k-1)})^{-1} \boldsymbol{c}_{k,P_q} && (23) \end{aligned}$$

This shows that σ_k^2 also has the same definition in terms of the sub-covariance matrix C_{P_q}.

Thus we see that with a perfect elimination ordering of variables, the new parameters $\boldsymbol{\rho}_k, \sigma_k^2$ are completely specified by the marginal covariance matrices of the prime components of the decomposable graph. For decomposable graphs a convenient prior distribution for the covariance matrix is the Hyper Inverse Wishart (HIW), which is the conjugate prior for each prime component in a Gaussian graphical model. The Hyper Inverse Wishart $C \sim HIW_G(\nu, M)$ is defined as the distribution such that the covariance matrix for each prime component is marginally Inverse Wishart: $C_{P_q} \sim IW(\nu - (s - |P_q|), M_{P_q})$, where M_{P_q} is the sub-matrix of M corresponding to P_q, and $|P_q|$ is the number of variables in the prime component.

Prior in the reparametrised space

For clarity, we relabel variables conditional on a given decomposable graph structure. To obtain the perfect elimination ordering of variables discussed above, we label variables with index $t = (|S_q|+1), \cdots, |P_q|$ within each residual R_q for components $q = 2, \cdots, Q$. For the first prime component all variables are labelled $t = 1, \cdots, |P_1|$.

With this labelling, the density of the HIW prior on the covariance matrix is transformed to

$$p(\sigma_{11}^2) \prod_{t=2}^{|P_1|} p(\sigma_{1t}^2) p(\boldsymbol{\rho}_{1t} \mid \sigma_{1t}^2) \prod_{q=2}^{Q} \prod_{t=|S_q|+1}^{|P_q|} p(\sigma_{qt}^2) p(\boldsymbol{\rho}_{qt} \mid \sigma_{qt}^2) \tag{24}$$

where the densities are

$$\sigma^2_{qt} \sim IG((\nu - s + t)/2, (m_{qt} - \boldsymbol{m}^t_{qt} M^{-1}_{P_q,(t-1)} \boldsymbol{m}_{qt})/2) \tag{25}$$

$$\boldsymbol{\rho}_{qt} | \sigma^2_{qt} \sim N(M^{-1}_{P_q,(t-1)} \boldsymbol{m}_{qt}, \sigma^2_{qt} M^{-1}_{P_q,(t-1)}) \tag{26}$$

for all q and t except $\sigma^2_{11} \sim \mathcal{IG}((\nu - s + 1)/2, m_{11}/2)$.

A nice feature of this transformation is that we do not need to do the completion operation to fill in the covariances between the separated parts of prime components, so sparsity in the graph can substantially reduce the computational time for this part of the model.

3 Posterior Computations

Recall we started from the model

$$Y = X_{\boldsymbol{\gamma}} \boldsymbol{\beta} + U \tag{27}$$

$$U \sim \mathcal{MN}(\mathbf{0}, \mathbb{I}_n, C) \tag{28}$$

and we reparametrised it into

$$y_{\xi_k} = X_{\boldsymbol{\gamma}_{\xi_k}} \beta_{\xi_k} + \sum_{l<k} u_{\xi_l} \rho_{\xi_k \xi_l} + \varepsilon_{\xi_k} \qquad k = 1, \dots, s \tag{29}$$

$$\varepsilon_j \sim \mathcal{N}\left(0, \sigma^2_j\right) \tag{30}$$

where ξ is the perfect elimination ordering derived from graph G.

By allowing $\boldsymbol{\gamma}$ to be completely unstructured we are not constrained to find the same set of predictors on the right hand side of Equation (27) for every outcome y_k. This leads to a class of linear models called Seemingly Unrelated Regressions (SUR). Unfortunately it is well known, see for example Holmes, Denison, and Mallick (2002), that for the SUR model there is no joint conjugate prior on $\boldsymbol{\beta}, C$.

The major consequence of this is that the strategy followed by Richardson, Bottolo, and Rosenthal (2010), Bottolo, Petretto, et al. (2011), and Lewin et al. (2016) of analytically integrating out the regression coefficients and the residuals' variance to focus on structure learning is not available for our more general model framework.

Posterior full conditional are nonetheless available and are conjugate to the Normal-(Hyper)Inverse Wishart model, however naïve computations of this model require storing and inverting both a $\tilde{p} \times \tilde{p}$ covariance matrix for the regression coefficients (where $\tilde{p} = \sum_{k=1}^{s} p_k$ and $p_k = \sum_{j=1}^{p} \gamma_{kj}$ is the number of selected regressors for outcome k) and a $s \times s$ matrix for the likelihood and the sampling distribution of C. The expressions are given in the appendix; see also for example Holmes, Denison, and Mallick (2002).

The update of $\boldsymbol{\gamma}$ and all other hyper-parameters where the likelihood is involved would require moreover the computation of the non-factorisable log likelihood, with computational complexity of $\mathcal{O}\left(ns^3\right)$.

The reparametrisation (8) and the introduction of a sparse precision matrix via the graph G allows

us much more economical Gibbs perturbation of the parameters, allowing us to scale the MCMC procedure to high dimensional data-sets.

3.1 Posterior full conditional for $\boldsymbol{\sigma}^2$, $\boldsymbol{\rho}$, $\boldsymbol{\beta}$

By using the above priors in the reparametrised space, defined in Eq (3),(25),(26), we can now derive posterior full conditional distribution for $\boldsymbol{\sigma}^2$, $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$.

In particular it will be possible to show that most of the dependency on the previous parameter values will be hidden in the value of the estimated residuals $U = Y - X\boldsymbol{\beta}$ at the previous iteration.

Because of this Zellner and Ando (2010) name this procedure Direct Monte Carlo (DMC) and show how it is possible to derive a tractable expression to update all the parameters of the reparametrised residual covariance matrix C under the Jeffrey's prior; we follow their work but show however that is possible to recover the correct posterior full conditionals, avoiding unnecessary and computationally prohibitive resampling steps, that enables a Gibbs sampler which factorises over the regression outcomes as does the transformed likelihood; we provide below the expressions given our more general priors from previous Sections:

$$\sigma^2_{qt}|\boldsymbol{\beta} \sim IG((n+\nu-s+t+|S_q|)/2, (\tilde{m}_{qt} - \tilde{\boldsymbol{m}}^t_{qt}\tilde{M}^{-1}_{P_q,(t-1)}\tilde{\boldsymbol{m}}_{qt})/2) \tag{31}$$

$$\boldsymbol{\rho}_{qt}|\sigma^2_{qt}, \boldsymbol{\beta} \sim N(\tilde{M}^{-1}_{P_q,(t-1)}\tilde{\boldsymbol{m}}_{qt}, \sigma^2_{qt}\tilde{M}^{-1}_{P_q,(t-1)}) \tag{32}$$

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}_k}|\boldsymbol{\sigma}^2, \boldsymbol{\rho}, \xi \sim \mathcal{N}\left(W_k \times X^t_{\boldsymbol{\gamma}_k}\tilde{y}_k\ ,\ W_k\right) \tag{33}$$

where

$$\tilde{M} = \tau\mathbb{I}_s + U^tU$$

and by defining the following index sets with respect to the perfect elimination order ξ

$$\mathcal{L}(k) = \{l|\xi(l) < \xi(k)\} \quad , \quad \mathcal{M}(k) = \{m|\xi(m) > \xi(k)\} \quad , \quad \mathcal{H}(k,m) = \{h,m|\xi(h) < \xi(m)\,,\, h \neq k\} \tag{34}$$

we can write

$$W_k = \left(X^t_{\boldsymbol{\gamma}_k}X_{\boldsymbol{\gamma}_k}\left(\frac{1}{\sigma^2_k} + \sum_{m\in\mathcal{M}(k)}\frac{\rho^2_{mk}}{\sigma^2_m}\right) + (w\mathbb{I}_{p_k})^{-1}\right)^{-1},$$

$$\tilde{y}_k = \begin{cases} \dfrac{\left(y_k - \sum_{l\in\mathcal{L}(k)}\rho_{kl}u_l\right)}{\sigma^2_k} & k = \xi_s \\ \frac{y_k}{\sigma^2_k} - \sum_{m\in\mathcal{M}(k)}\frac{\rho_{mk}}{\sigma^2_m}(u_m - \sum_{h\in\mathcal{H}(k,m)}\rho_{mh}u_h - \rho_{mk}y_k) & k = \xi_1 \\ \dfrac{\left(y_k - \sum_{l\in\mathcal{L}(k)}\rho_{kl}u_l\right)}{\sigma^2_k} - \sum_{m\in\mathcal{M}(k)}\frac{\rho_{mk}}{\sigma^2_m}(u_m - \sum_{h\in\mathcal{H}(k,m)}\rho_{mh}u_h - \rho_{mk}y_k) & k = \xi_2, \dots, \xi_{s-1} \end{cases}$$

These will be used not only to update $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\rho}$ in a Gibbs sampler fashion inside the MCMC, but also as proposal distributions when we will need to update their value while changing some of their conditioning parameters (namely $\boldsymbol{\gamma}$ and G).

3.2 Sampler details

Here we will give a brief description of the MCMC sampler used, full details can be found in the Appendix (Section 11).

We use the Evolutionary Stochastic Search (Bottolo and Richardson, 2010; Bottolo, Petretto, et al., 2011; Lewin et al., 2016), which uses a particular form of Evolutionary Monte Carlo (EMC) as defined in Liang and Wong (2000); this helps the chain mix in the high dimensional binary space of both covariates selection and graph structure learning. Each chain has a single temperature t_c associated with it and we only temper the likelihood and not the already diffused priors.

The binary Variable Selection coefficients $\boldsymbol{\gamma}$ and the regression coefficients $\boldsymbol{\beta}$ are updated jointly via Metropolis-Hastings.

We propose a new value for $\boldsymbol{\gamma}$ using an adaptive algorithm that learns a good proposal distribution independently for each γ_{kj}, but focuses on coefficients that have been explored less than others. A complete description of the method and implementation is the subject of a manuscript currently under preparation.

New values of $\boldsymbol{\beta}$ are proposed from the posterior full conditional for those elements corresponding to proposed changes in $\boldsymbol{\gamma}$. If connected with a tempered chain the proposal changes slightly to reflect the more diffuse likelihood. Important to note is that by using posterior full conditionals in the joint $\boldsymbol{\gamma}, \boldsymbol{\beta}$ updates we are actually removing the dependence of the acceptance ratios from the sampled values of the non-structural parameters β. This is in fact know as implicit marginalisation Holmes and Held (2006) and has been shown to greatly improve mixing of the chain with respect to the structural parameters which are in our case the main focus of inference.

$\boldsymbol{\beta}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\rho}$ are all also updated by sampling from their posterior full conditionals, while each of the other hyper-parameters is updated via a Metropolis-within-Gibbs update with adaptive proposal distribution (Roberts and Rosenthal, 2009).

4 Structure learning

So far we have discussed reparametrisations of C given a specific graph G but we are instead interested in estimating its structure so that we can infer all the conditional (in)dependency relationships between the outcomes. Even for moderate number of outcomes s, the graph space is exceedingly big and thus posterior sampling can be challenging; to help the Markov Chain explore the whole space we take advantage of the junction tree representation of a decomposable graph as a state variable.

A spanning tree J of a graph G where each node is a subset of the vertices in G has the junction property if it is such that for any two nodes N_q and N_l every node on the path between them in J contains $N_q \cap N_l$. When the nodes of a tree J with the junction property are the prime components of G ($\{P_1, \ldots, P_Q\}$), J is a *junction tree* of G. For a more formal definition see Thomas and Green (2009) and Cowell et al. (2006).

For any decomposable graph it is known that such representation exists (see again for example Cowell et al. (2006)). Note for example that is trivial to construct one such tree from the perfect ordering of the cliques, where the junction tree property is guaranteed by the running intersection property. As with the perfect ordering, a decomposable graph may have many junction trees representation; for a given junction tree, the implied graph is however uniquely determined.

Green and Thomas (2013) introduce an efficient sampler that parametrise the model in terms of the junction tree J which allows local perturbations to multiple edges of G simultaneously, contrarily to the more common single-edge modification Giudici and Green (1999). The junction tree sampler ensures decomposability of the graph at all times and providing as a by-product a perfect elimination ordering ξ; this helps dramatically the mixing of the Markov Chain, especially in high dimensions, and avoids the need for *maximal cardinality search* type of algorithms to check for chordality and to build ξ.

With this in mind, Green and Thomas (2013) define a prior directly on the junction tree itself, proportional to a selected prior $\pi(G)$ on G and uniform over the number of possible junction trees $\mu(G)$ for a specific decomposable G

$$\pi(J) = \frac{\pi(G(J))}{\mu(G(J))}.$$

This allows to correct for the otherwise oversampling bias toward graphs that have more junction tree representations, but at the rather steep cost of computing $\mu(G)$; see Thomas and Green (2009).

Given that graphs with fewer edges will have more junction tree representations, and since we are specifically interested in sparse models, the oversampling bias is not necessarily a problem in our case, as long as is correctly accounted for. We thus specify the prior on G to be proportional to the number of possible junction tree representations; additionally we specify independent Bernoulli(η) prior on each edge of the graph (Binomial on the cardinality edge-set $|\mathcal{E}|$), with a Beta prior on η to help the sampler push toward sparsity when needed.

$$\pi(G) \propto \pi(|\mathcal{E}|)\mu(G(J)) \tag{35}$$

$$\pi(J) = \frac{\pi(G(J))}{\mu(G(J))} \tag{36}$$

$$|\mathcal{E}| \sim Binomial(s(s-1)/2, \eta) \tag{37}$$

$$\eta \sim \mathcal{B}eta\,(a_\eta, b_\eta) \tag{38}$$

where we specify $a_\eta = b_\eta = 1$ so that we don't *over-force* the sampler toward sparsity.

Note that, even though each sample is from a decomposable model, the sampler allows us to discover non-decomposable graph structures via Bayesian Model Averaging of the marginal edge inclusion probabilities.

Finally it is worth mentioning that with changes in the graph, the perfect elimination ordering ξ will change as well and with it the conditioning order in Eq (24); this means that the $\boldsymbol{\rho}$ coefficients lose their interpretation as partial correlations from one iteration to the next. It is important to remark though that since the transformation between C and $\boldsymbol{\sigma}^2$, $\boldsymbol{\rho}$ is bijective, the MCMC will retain its validity. As we are interested mainly in structure learning, both the variable and the covariance selections, we consider the reparametrised covariance as nuisance parameters.

4.1 MCMC sampling of the graphical structure

To sample the graph G we use the Junction Tree Sampler of Green and Thomas (2013), allowing us to connect and disconnect multiple nodes of the graph at the same time.

We have also implemented a randomisation move inspired by Thomas and Green (2009) that uniformly samples an equivalent junction tree given the current graph. This move rebases the junction tree at

different root component and exchanges some of the links in the junction tree between nodes with identical separators. This does not affect the relation between J, the current junction tree, and the current graph but is purely a way to increase mixing as described in Green and Thomas (2013).

The covariance parameters are then jointly proposed with the new graph structure; the reparametrisation from C to $\left(\boldsymbol{\sigma}^2, \boldsymbol{\rho}\right)$ allows us to update only some of the coefficients, connected with the graph update. The easiest way to do this would be to compare with the current ξ and only update elements starting from the first modified index in the perfect elimination order corresponding to the proposed junction tree J'. Their update use again the (possibly tempered) posterior full conditionals described in Section 3.1. The proposed triplet $\left(J', \sigma'^2_{qt}, \rho'_{qt}\right)$ is accepted or rejected it using Metropolis-Hastings.

Finally, η is updated via a Gibbs perturbation, as its full conditional is analytically available.

5 Results

We want to evaluate the performances of the reparametrised multivariate sparse SUR model and of our efficient sampler in both a simulated settings and on real mQTL data. The aim of the study detailed in Section 5.2 is detecting associations between genetic variants and metabolites. Biologically, thanks to the growing evidence in favour of pleiotropy Sabatti et al. (2009) and Sivakumaran et al. (2011), where the same locus is reportedly associated with multiple phenotypes, we expect these associations to be driven by a handful of SNPs that associate with numerous metabolites (Breitling et al., 2008) and thus we specify for each γ_{kj} a prior probability decomposed into an overall sparsity level for outcome k and a propensity parameter for each predictor j to form links; this was introduced by Bottolo, Petretto, et al. (2011) as *hotspot* prior and we notate it as follows:

$$\gamma_{kj} \sim Bernoulli\left(\omega_{kj}\right) \tag{39}$$

$$\omega_{kj} = o_k \times \pi_j \tag{40}$$

$$\pi_j \sim \mathcal{G}amma(a_\pi, b_\pi) \tag{41}$$

$$o_k \sim \mathcal{B}eta(a_o, b_o) \tag{42}$$

$$\text{under the constraint that } o_k \times \pi_j \leq 1 \qquad \forall j, k \tag{43}$$

where o_k defines the sparsity level and π_j describes the prior predisposition for each predictor to be an *hotspot*. In the MCMC we use MH to sample new points for o_k and π_j (see Bottolo and Richardson (2010) for details).

The four hyperparameters are chosen so that the average model size for each outcome is small, as we want to enforce a strong sparsity in the model, while a_π and b_π are chosen so that $\mathbb{E}\left[\pi_j\right] = 2$ and $\mathbb{V}ar\left[\pi_j\right] = 2$.

The prior for $\beta_{kj}|\gamma_{kj}$ is $\mathcal{N}(0, w)$ and since we standardise and centre all responses and covariates, the hyperparameters for τ and w are set that these variances are centred on small values but at the same time allowing the respective prior to be diffuse.

All initialisation are random, except for the graph which is initialised empty (*i.e.* on the independent model).

The sampler with all the above specifications is implemented fully in C++ and freely available at

https://github.com/mbant/Bayesian_SSUR .

5.1 Simulation Study

Here we validate our method against the Hierarchical Evolutionary Stochastic Search (HESS) of Bottolo, Petretto, et al. (2011) in a synthetic setting.

Similarly to Richardson, Bottolo, and Rosenthal (2010) and Bhadra and Mallick (2013) we set up our simulation study by randomly subsampling $p = 300$ SNPs from our real *-omics* dataset (see Section 5.2). This will form our covariate set X and allows us to mimic real correlation effects and linkage disequilibrium between genes that would be otherwise difficult to simulate artificially. The observed correlations between covariates ranges from very small to over 0.8 in absolute value.

We then set $n = 200$ and $s = 30$ and proceed by selecting the correlation structure for the outcomes in form of a graph. We explore either a block diagonal structure with 4 components, a decomposable graph with 4 components or a non decomposable (cyclic) model with 5 components.

In order to present a range of possible association patterns between outcomes and predictors, we fix (conditionally on the selected graph structure) the $\boldsymbol{\gamma}$ matrix as shown in Figure 2. This is built so that a small group of different covariates is connected to each of

- all outcomes, *i.e.* a true *hotspot*;
- all outcomes within each Prime component;
- all outcomes within each Residual component, so that each set of predictors is associated with different outcomes;
- a set of selected outcomes that spans multiple components, so that the covariates are linked to both correlated and (conditionally) independent outcomes.

Now that the structure is fixed, we sample the non-zero $\beta_{\boldsymbol{\gamma}}$ coefficients as random normals $\beta_{\boldsymbol{\gamma}} \sim \mathcal{N}(5, 1)$ (so that most of the real βs are distinct from zero) and the residuals' *precision* matrix C^{-1} from a $\mathcal{W}_{\mathcal{G}}(s+2, M)$ (using the *BDgraph* package of Mohammadi and Wit (2015)) with $M = f\mathbb{C}$, $f \in \mathbb{R}^+$ and $\mathbb{C}$ a correlation matrix with 0.85 on the off-diagonal elements; C is obtained by inverting the precision matrix and the residuals are finally sampled from $u \sim \mathcal{MN}(0, \mathbb{I}_n, C)$.

We choose f at each repetition so that we can obtain a data-set with the desired empirical signal to noise ratio (within 10% of the desired value), measured as

$$SNR = \frac{\frac{1}{s}\sum_{k\in\xi} v_k}{\frac{1}{s}\sum_{k\in\xi} \sigma_k^2}, \qquad v_k = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{i,\boldsymbol{\gamma}_k}\beta_{\boldsymbol{\gamma}_k} + \sum_{l\in\mathcal{L}(k)} u_{i,l}\rho_{kl}\right)^t \left(X_{i,\boldsymbol{\gamma}_k}\beta_{\boldsymbol{\gamma}_k} + \sum_{l\in\mathcal{L}k} u_{i,l}\rho_{kl}\right) \tag{44}$$

where $\left(\sigma^2, \rho\right)$ are the reparametrised values of the covariance matrix C and $\mathcal{L}$ is as defined in Eq (34).

We repeat the procedure 50 times for each combination of correlation structure and f/SNR to produce different datasets and average the result obtained from each run of 250000 iterations, for both our Sparse SUR model and the HESS model. We expect HESS to perform well even in the presence of correlated

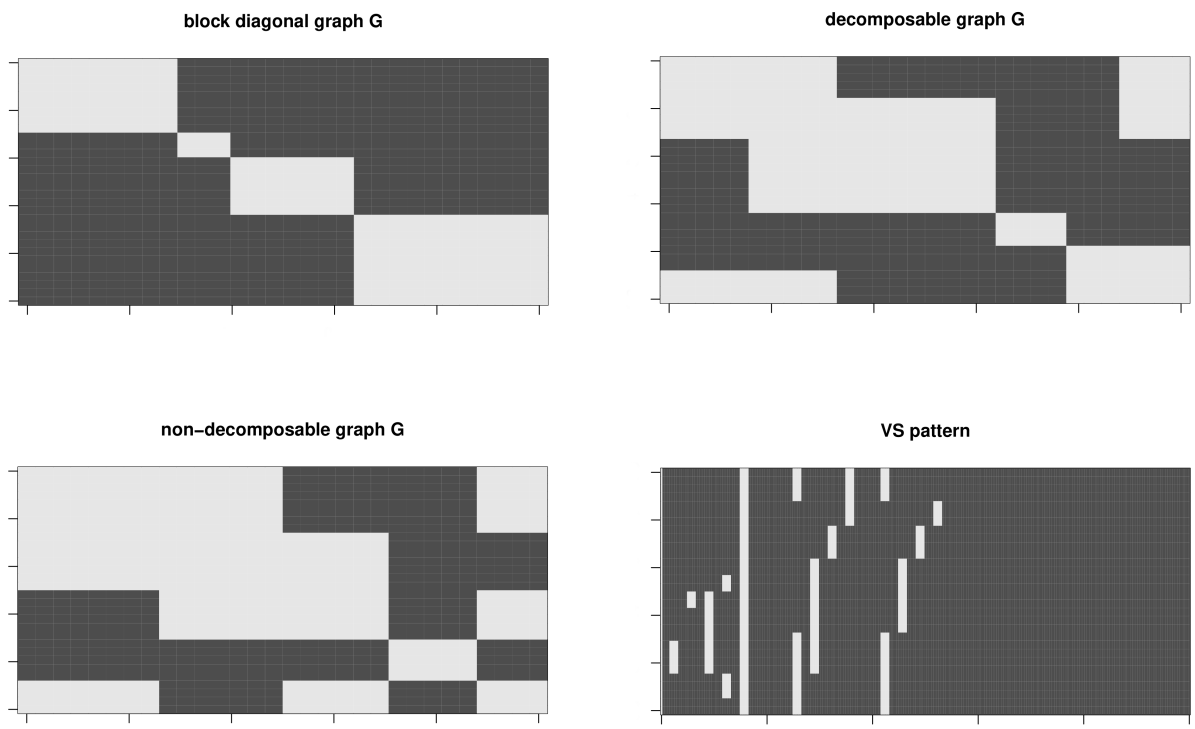

Figure 2: *Top-left, top-right and bottom-left*: Graph structures, represented as adjacency matrices, used in the simulations (white tassels represent edges); *bottom-right* example $\boldsymbol{\gamma}$ pattern for a decomposable model, outcomes on the rows and predictors on the columns (white tassels represent associations).

residuals, as reported in previous publications (Bottolo, Petretto, et al., 2011; Lewin et al., 2016), but to return noisier estimate due to this confounding effect.

To report on the performances of both algorithms, we show the averaged results (exemplified below for a $SNR = 1.5$ and a decomposable graph) and focus our attention on posterior marginal inclusion probabilities, *i.e.* the average over the whole Markov Chain of every single element γ_{kj} or $G_{kk'}$. The results are further summarised across the different dataset with the same conditions into averaged ROC curves, constructed thresholding the posterior marginal inclusion probabilities at different levels.

Note that both HESS and SSUR outputs are far more complex, having access technically to the posterior distribution of the whole matrices $\boldsymbol{\gamma}$ and G, as well as the (convoluted to report) junction tree sample path and, in the case of SSUR, the posterior distribution of $\boldsymbol{\beta}$ and possibly of C as well. Given one chosen structure ($\boldsymbol{\gamma}$ and G) for the model, chosen either by Bayesian model averaging or as the MAP estimator, conditional samples for both $\boldsymbol{\beta}$ and C are also straightforward to obtain. The work of Clyde, Ghosh, and Littman (2011) provide additional insights on reporting from similar outputs.

When looking at Table 1, that summarise the results for the graph reconstructions averaged over the simulations conditions, is important to note that the graph structure has much more implicit noise than the variable selection. While the regression coefficients $\boldsymbol{\beta}$ are fixed to a significantly greater than zero value, we are not only *sampling* from the G-Wishart to obtain the precision matrix but also, because of the positive definite constraint, some of the observed correlation values decrease down to ~ 0.25 in the non-decomposable case and down to ~ 0.55 for the decomposable models, from the nominal value of 0.85 imposed by $\mathbb{C}$ to the prior of C^{-1}. It is thus not so surprising that the estimated graph, even for very low values of specificity and high signal-to-noise ratios, doesn't always reconstruct the nominal

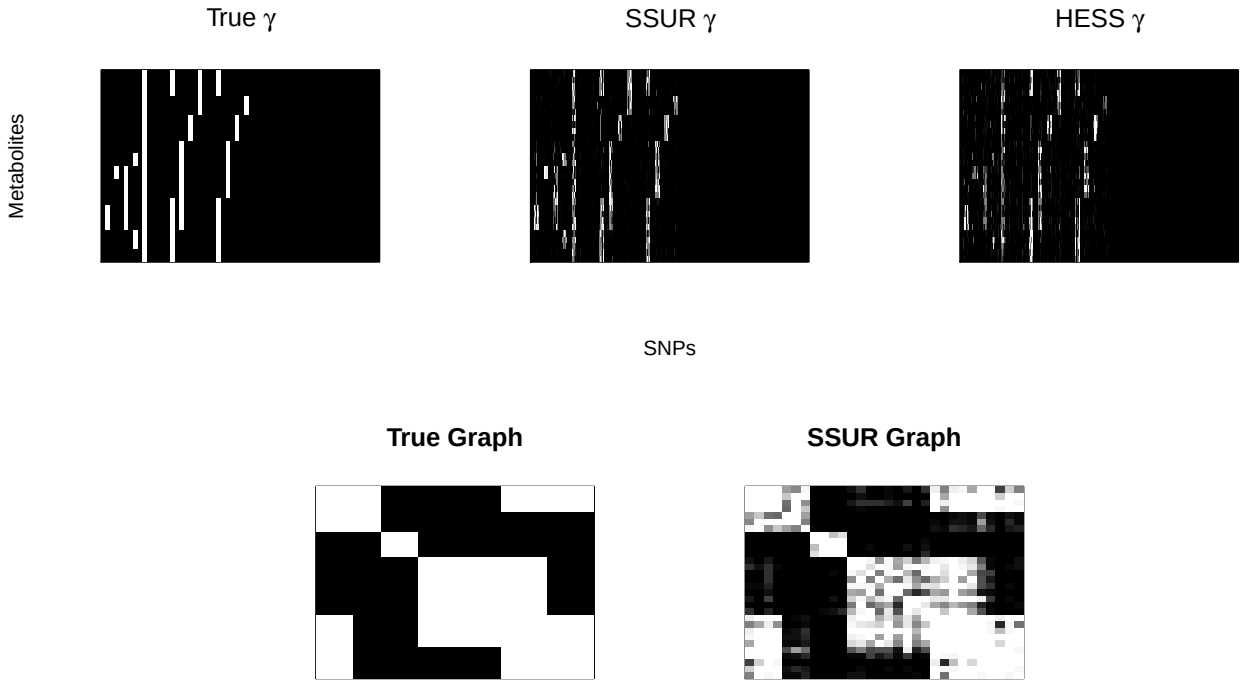

Figure 3: An example result ($SNR = 1.5$): plotted are *top-row* true and estimated marginal inclusion probabilities for the variable selection coefficients for both SSUR and HESS. HESS reconstruction appears very good even for this relatively low SNR, but markedly noisier than for SSUR. *Bottom-row* true and estimate posterior marginal edge inclusion probabilities, used to reconstruct the graph, from SSUR

Model	SNR	True Positives	False Positives
Block-diagonal	1.5	0.943	0.140
Block-diagonal	3	0.938	0
Decomposable	1.5	0.915	0.119
Decomposable	3	0.948	0.070
Non-decomposable	1.5	0.927	0.183
Non-decomposable	3	0.977	0.167

Table 1: Averaged true positive rates and false positive rates relative to graph reconstruction when the resulting posterior marginal edge inclusion probabilities are thresholded at the 0.5 level.

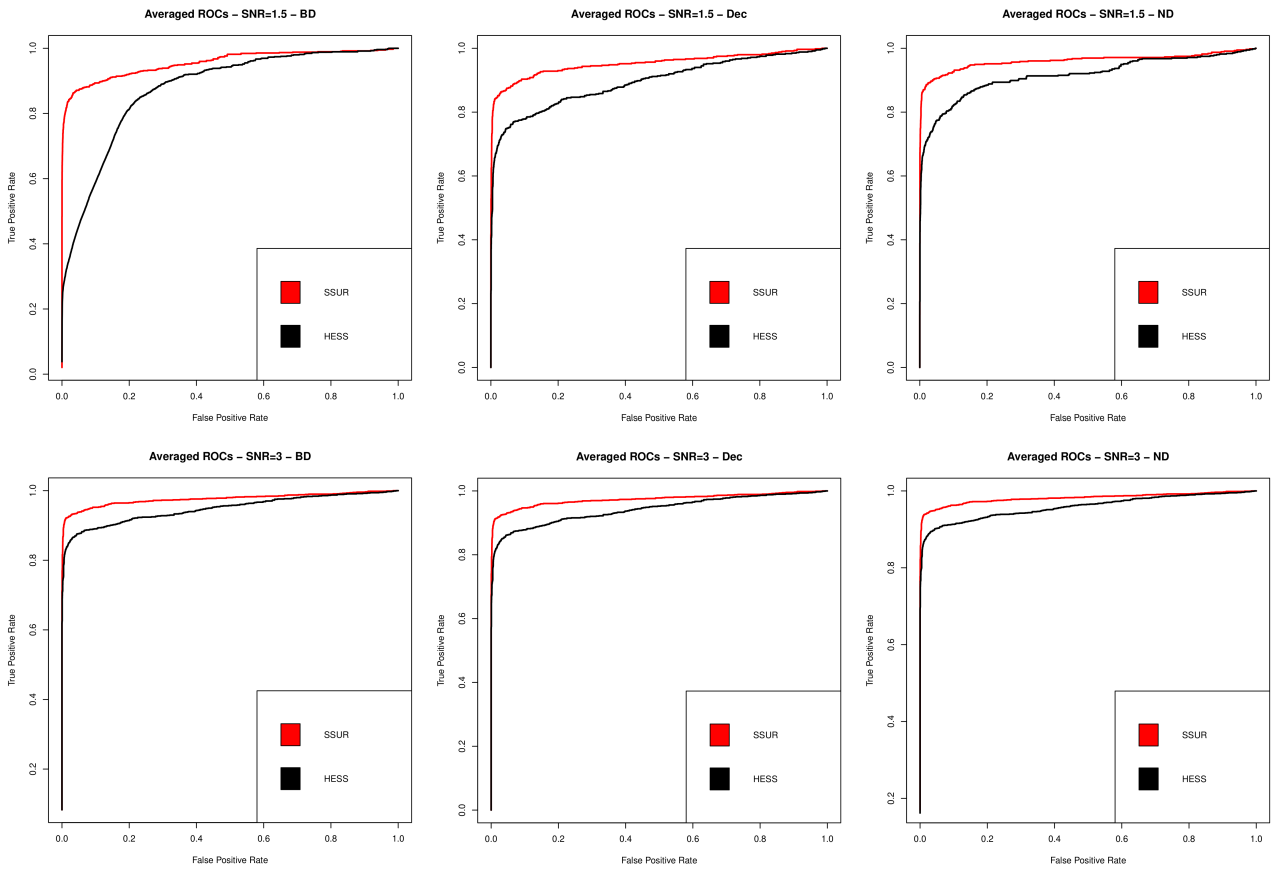

Figure 4: Averaged ROC curves for SSUR (red) and HESS (black) to compare the variable selection reconstruction of the two methods for all the combinations of SNR and graph structure (**B**lock-**D**iagonal , **Dec**omposable and **N**on-**D**ecomposable). HESS performs uniformly worse on average, albeit slightly, for the same number of model evaluations; taking into account the residual correlations in SSUR leads to improved reconstruction.

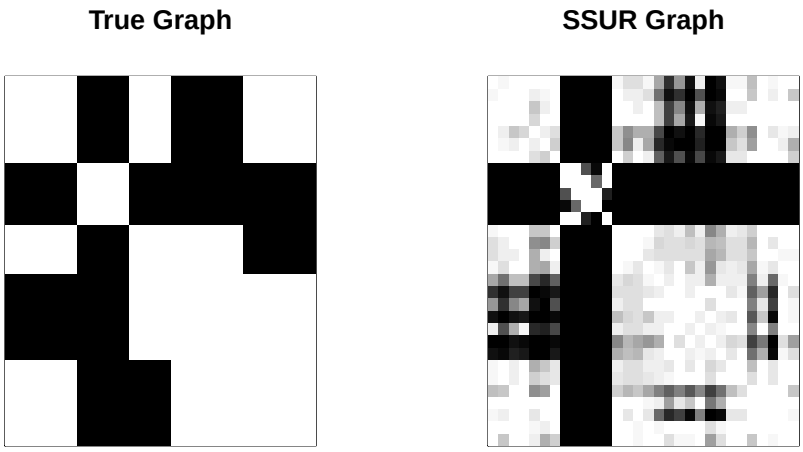

Figure 5: An example *non-decomposable* graph reconstruction, where the sampler seems to converge to a denser triangulations of the true graph.

graph perfectly.

The results correspond to our expectations; HESS is known to perform well even in cases where residuals are correlated (Bottolo, Petretto, et al., 2011), as long as SNR is not too low and we indeed observe that the simpler model detect with high probability the true effects. In most cases though, especially at low levels of 'signal', HESS estimates are more noisy and more false positives are picked up due to the confounding effect of the unaccounted for correlations. SSUR has a more marked separations between positive (very probable) and negative (very improbable) signals and overall returns less noisy estimates, especially at lower level of signal. SSUR is also able at the same time to recover appropriately the conditional (in)dependence structure of the residuals especially when the decomposability assumption is not violated; in the non-decomposable cases graph reconstruction seems to converge to a close (in the graph space), more dense chordal graph alternative (see Figure 5), which is what one would expect based on Fitch, Jones, and Massam (2014), see also the discussion in Section 6.1.

5.2 High dimensional -*omics* data analysis

Finally, we performed an analysis of human e-QTL data; the data come from the 31-year follow-up study of the Northern Finland Birth Cohort NFBC66 collection, previously described in details in Rantakallio (1969) and more recently by Soininen, Kangas, Würtz, Tukiainen, et al. (2009) and Soininen, Kangas, Würtz, Suna, et al. (2015), and consists, after quality control and data cleaning, of $n = 5154$ directly genotyped individuals on a total of $p = 9310$ SNPs on chromosome 16, which were also measured on $s = 158$ metabolites. The metabolites have been normalised and standardised via the inverse rank-Normal transformation (Blom, 1958), following Kettunen, Demirkan, et al. (2016).

Figures 6 and 7 show some of the results from this analysis.

The posterior expectations of $\boldsymbol{\pi}$ uncover the presence of some *hotspots* (i.e. genetic variants that are associated with multiple outcomes), in particular we report `rs4985124`, `rs931406`, and most importantly `rs12102766` and `rs3764261` as shown in Figure 6 and Table 2.

Analysing the whole $\boldsymbol{\gamma}$ gives us a lot more information though, as shown in Figure 7, where we plotted

posterior marginal inclusion probabilities for each predictor against their position in chromosome 16 with different outcomes being superimposed; we can see how some SNPs are associated with only one or a few outcomes and would thus be missed by only looking at hotspot detections; this results are presented in Table 3 as well.

By thresholding the posterior inclusion probabilities at 0.5 we discover a total of 38 associations, and the average Bayes FDR (bFDR, see e.g. Lewin et al. (2016)) is ~ 0.0578. By increasing the threshold to 0.9 we would keep 32 coefficients, with a bFDR < 0.01.

SNPs	Associated Metabolite				
rs4985124		FAw3.FA	FAw6.FA		
rs7191766	L.LDL.P	L.LDL.L	L.LDL.PL	L.LDL.C	L.LDL.CE
	L.LDL.FC	M.LDL.P	M.LDL.L	M.LDL.PL	M.LDL.C
	M.LDL.CE	M.LDL.FC	S.LDL.P	S.LDL.L	S.LDL.PL
	S.LDL.C	S.LDL.CE	S.LDL.FC	S.HDL.C	S.HDL.CE
	Serum.C	LDL.C	EstC	FreeC	LA
		FAw3	FAw6	PUFA	
rs3764261	XXL.VLDL.C	XXL.VLDL.CE	XL.VLDL.CE	L.VLDL.CE	M.VLDL.CE
	S.VLDL.PL	S.VLDL.C	S.VLDL.CE	XS.VLDL.C	XS.VLDL.CE
	XL.HDL.P	XL.HDL.L	XL.HDL.PL	XL.HDL.C	XL.HDL.CE
	XL.HDL.FC	L.HDL.P	L.HDL.L	L.HDL.PL	L.HDL.C
	L.HDL.CE	L.HDL.FC	M.HDL.P	M.HDL.L	M.HDL.PL
	M.HDL.C	M.HDL.CE	M.HDL.FC	M.HDL.TG	S.HDL.TG
	LDL.D	HDL.D	VLDL.C	HDL.C	HDL2.C
	HDL3.C	TotPG	TG.PG	PC	SM
	TotCho	ApoA1	ApoB.ApoA1	LA	LA.FA
rs12102766	XXL.VLDL.PL	XXL.VLDL.C	XXL.VLDL.CE	XXL.VLDL.FC	XL.VLDL.P
	XL.VLDL.L	XL.VLDL.PL	XL.VLDL.C	XL.VLDL.CE	XL.VLDL.FC
	L.VLDL.P	L.VLDL.L	L.VLDL.PL	L.VLDL.C	L.VLDL.CE
	L.VLDL.FC	M.VLDL.L	M.VLDL.PL	M.VLDL.C	M.VLDL.CE
		M.VLDL.FC	VLDL.C		
rs931406	L.LDL.PL	M.LDL.PL	S.LDL.P	S.LDL.L	S.LDL.PL
	Serum.C	EstC	FreeC	TotPG	SFA

Table 2: Uncovered hotspots and relative associated metabolites with $BF > 10$.

These results confirm the known association (Sabatti et al., 2009; Kettunen, Tukiainen, et al., 2012) between rs3764261 and, predominantly, HDL but also highlight the the relevance of the CEPT locus on different lipoproteins; also rs4985124 is situated only $\sim$ 23Kb and on the same PDXDC1 locus of rs11075253 and rs11644601, which were reported previously (Kettunen, Tukiainen, et al., 2012; Kettunen, Demirkan, et al., 2016) as associated with fatty acids metabolism. Finally rs12102766, that we report as associated with multiples metabolites connected to VLDL, is situated in the proximity of rs74249229, previously published by Kettunen, Demirkan, et al. (2016).

Our novel method that allows the residuals to be correlated was also able to uncover other potential

SNPs	Associated Metabolite
rs17683096	LDL.D
rs4782717	XL.HDL.CE
rs4843914	Cit
rs2544630	Gln

Table 3: Relevant associations for non-hotspot SNPs.

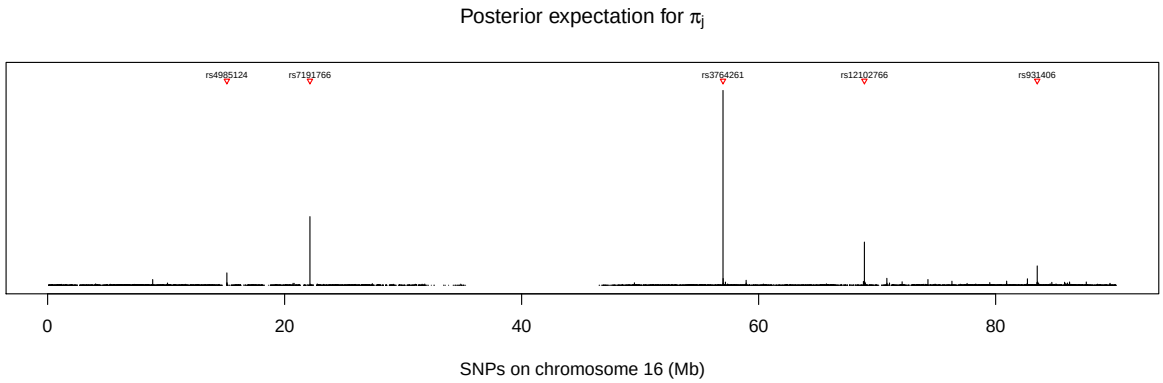

Figure 6: Manhattan plot of the posterior expectation for π_j, ordered along chromosome 16, for hotspots detection

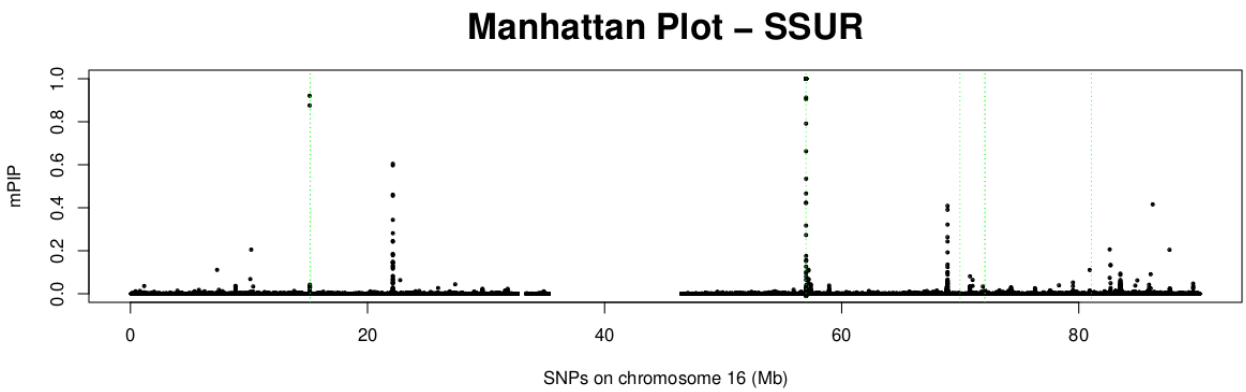

Figure 7: Manhattan plots for marginal posterior inclusion probabilities of association for genetic variant on chromosome 16, all metabolites superimposed. Only a few SNPs seems to be relevant and for a restricted number of metabolites. Some SNPs appear to be associated with more than one outcome. In green dotted lines we show other previous known loci from Kettunen, Demirkan, et al. (2016).

important associations that warrant further investigations, in particular `rs7191766`, associated with multiple cholesterol-related phenotypes.

6 Discussion

6.1 The decomposability assumption

One might think that the restriction to decomposable graphs would be too stringent for real applications; various attempt have in fact been made to relax such an assumption, using the G-Wishart distribution first introduced by Roverato (2002) (see Wang (2012), Mohammadi and Wit (2015), and Mitsakakis, Massam, and Escobar (2011) and references therein for some recent examples). The computational disadvantages connected with a general graph are usually exceedingly big though, as shown extensively in Jones et al. (2005); the posterior computations require in fact expensive approximations, usually Monte Carlo, to the normalising constant of the G-Wishart. Even where more recent results (Uhler, Lenkoski, and Richards, 2018) shows that exact computations are possible, they still relies on expensive iterative methods that do not scale well with the dimension of the problem.

Another disadvantage of course is that for a non-decomposable graph the correspondence between the reparametrised $\boldsymbol{\rho}$ and the zeros in the precision matrix C^{-1} is harder to derive, which makes unrestricted graphs less interesting in the context of this work. Lastly, as we will see later (see Section 4), assuming decomposability allows for the Junction Tree representation of G (Lauritzen, 1996), which gives us the ability to sample multi-edge modification of the graph.

The work of Fitch, Jones, and Massam (2014) concludes that, under model assumptions similar to ours, inference on G will asymptotically converge toward minimal triangulations of the true graph, *i.e.* the graph with the least amount of extra edges so that chordality is obtained, and that inference on the covariance matrix is competitive in terms of prediction errors against penalised likelihood methods that estimate unrestricted graphs like the graphical lasso. In practice, assuming decomposability seems to be sensible and inference on the covariance matrix under such an assumption sound; the only obvious word of caution is against over-interpretation of the estimated graph.

6.2 Conclusion

This work presents a novel method to perform Bayesian variable selection in a matrix-variate regression settings jointly for multivariate outcomes that takes into account residual correlation while maintaining a flexible association pattern between responses and predictors. Although conjugacy is lost for this very general model, by virtue of a crucial reparametrisation of the covariance matrix the computations can be greatly sped up and thus our method is able to analyse a large number of both outcomes and covariates, thanks as well to the efficient C++ implementation, as demonstrated in the real data example. It is moreover possible to introduce further computational gains by assuming that the conditional independence structure between the residuals is sparse and inference on the resulting graph is straightforward to obtain; we also show that is possible to obtain simple expressions for the prior distribution of the reparametrised parameters starting from the usual Normal-(Hyper)Inverse Wishart prior for both the dense and the sparse cases.

We have shown in the simulation study how, in the presence of non-negligible residual correlations, our method exhibit better performances in selecting relevant predictors and is able at the same time to perform *covariance selection*. Nevertheless, this method is inherently slower than models that assume independent outcomes or other simplifying assumptions which allow to retain conjugacy, that can comparatively analyse a greater number of outcomes.

Considering approximate estimation techniques would alleviate this problem at the cost of inferring an approximation of the exact posterior distribution; in spite of that some of these approaches has been shown to be effective in settings similar to ours (Carbonetto and Stephens, 2012; Ruffieux et al., 2017) and the posterior full conditional presented in this manuscript should help in the derivation of the updates for at least the Variational Inference family of methods. Another possible extension is the generalisation from the linear model to a wider family of regressions, relaxing the normality assumption on the residuals.

The method is nonetheless able to scale to scale well in the regime of hundreds of outcomes and thousands of predictors, as demonstrated by our application to the NFBC66 mQTL dataset; we were able to recover already published and known associations, as well as possibly uncover some more previously unknown associations that might offer new insights in the relationships between chromosome 16 and lipid metabolism.

Thanks to the very general formulation of the SUR models we expect the present work to find appli-

cations beyond the mQTL-type data presented here, for example in finance, econometrics and other biological settings where linked regression models are already widespread.

Acknowledgements

This work was supported by the UK Medical Research Council grant MR/M013138/1 "Methods and tools for structural models integrating multiple high-throughput omics data sets in genetic epidemiology" (MB, AL, LB, SR, MRJ), the European Union Horizon 2020 grant "DynaHealth: Understanding the dynamic determinants of glucose homeostasis and social capability to promote healthy and active aging" grant agreement No 633595 (MRJ and AL), a Medical Research Council grant MC_UP_0801/1 (SR), and the Alan Turing Institute under the Engineering and Physical Sciences Research Council grant EP/N510129/1 (LB and SR). MAK works in a unit that is supported by the University of Bristol and UK Medical Research Council (MC_UU_12013/1). The Baker Institute is supported in part by the Victorian Government's Operational Infrastructure Support Program.

References

[1] Marit Ackermann, Weronika Sikora-Wohlfeld, and Andreas Beyer. "Impact of natural genetic variation on gene expression dynamics". In: *PLoS genetics* 9.6 (2013), e1003514.

[2] Kyounghwa Bae and Bani K Mallick. "Gene selection using a two-level hierarchical Bayesian model". In: *Bioinformatics* 20.18 (2004), pp. 3423–3430.

[3] Anindya Bhadra and Bani K Mallick. "Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis". In: *Biometrics* 69.2 (2013), pp. 447–457.

[4] Gunnar Blom. *Statistical estimates and transformed beta-variables.* John Wiley & Sons, New York, 1958.

[5] Leonardo Bottolo, Enrico Petretto, et al. "Bayesian detection of expression quantitative trait loci hot-spots". In: *Genetics* (2011), genetics–111.

[6] Leonardo Bottolo, Sylvia Richardson, et al. "Evolutionary stochastic search for Bayesian model exploration". In: *Bayesian Analysis* 5.3 (2010), pp. 583–618.

[7] Rainer Breitling et al. "Genetical Genomics: Spotlight on QTL Hotspots". In: *PLOS Genetics* 4.10 (Oct. 2008), pp. 1–4. DOI: `10.1371/journal.pgen.1000232`. URL: `https://doi.org/10.1371/journal.pgen.1000232`.

[8] Philip J Brown, Marina Vannucci, and Tom Fearn. "Bayes model averaging with selection of regressors". In: *Journal of the Royal Statistical Society, Series B* 64 (2002), pp. 519–536.

[9] Philip J Brown, Marina Vannucci, and Tom Fearn. "Multivariate Bayesian variable selection and prediction". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.3 (1998), pp. 627–641.

[10] Peter Carbonetto, Matthew Stephens, et al. "Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies". In: *Bayesian Analysis* 7.1 (2012), pp. 73–108.

[11] Carlos M Carvalho, Helene Massam, and Mike West. "Simulation of hyper-inverse Wishart distributions in graphical models., Biometrika 94 647–659". In: *Mathematical Reviews (MathSciNet): MR2410014 Zentralblatt MATH* 1135 (2007).

[12] Carlos M Carvalho, Nicholas G Polson, and James G Scott. "The horseshoe estimator for sparse signals". In: *Biometrika* 97.2 (2010), pp. 465–480.

[13] Hugh Chipman et al. "The practical implementation of Bayesian model selection". In: *Lecture Notes-Monograph Series* (2001), pp. 65–134.

[14] Anna Cichonska et al. "metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis". In: *Bioinformatics* 32.13 (2016), pp. 1981–1989.

[15] Merlise Clyde. "Bayesian model averaging and model search strategies". In: Proceedings of the Sixth Valencia International Meeting (1999). Ed. by J.M. Bernardo et al., pp. 157–185.

[16] Merlise A Clyde, Joyee Ghosh, and Michael L Littman. "Bayesian adaptive sampling for variable selection and model averaging". In: *Journal of Computational and Graphical Statistics* 20.1 (2011), pp. 80–101.

[17] GTEx Consortium et al. "The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans". In: *Science* 348.6235 (2015), pp. 648–660.

[18] Robert G Cowell et al. *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks.* Springer Science & Business Media, 2006.

[19] Abhirup Datta, Sudipto Banerjee, and James S Hodges. "Spatial disease mapping using Directed Acyclic Graph Auto-Regressive (DAGAR) models". In: *arXiv preprint arXiv:1704.07848* (2017).

[20] A Philip Dawid. "Some matrix-variate distribution theory: notational considerations and a Bayesian application". In: *Biometrika* 68.1 (1981), pp. 265–274.

[21] Arthur Pentland Dempster. *Elements of continuous multivariate analysis.* Tech. rep. 1969.

[22] Marc-Emmanuel Dumas et al. "Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models". In: *Nature genetics* 39.5 (2007), p. 666.

[23] Warwick B Dunn et al. "Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy". In: *Chemical Society Reviews* 40.1 (2011), pp. 387–426.

[24] Morris L Eaton. *Multivariate statistics: a vector space approach.* John Wiley & Sons, New York, 1983.

[25] A Marie Fitch, M Beatrix Jones, Hélene Massam, et al. "The performance of covariance selection methods that consider decomposable models only". In: *Bayesian Analysis* 9.3 (2014), pp. 659–684.

[26] Edward George and Robert E McCulloch. "Variable selection via Gibbs sampling". In: *Journal of the American Statistical Association* 88 (1993), pp. 881–889.

[27] Helen G Gika et al. "Current practice of liquid chromatography–mass spectrometry in metabolomics and metabonomics". In: *Journal of pharmaceutical and biomedical analysis* 87 (2014), pp. 12–25.

[28] Paolo Giudici and PJ Green. "Decomposable graphical Gaussian model determination". In: *Biometrika* 86.4 (1999), pp. 785–801.

[29] Peter Green and Alun Thomas. "Sampling decomposable graphs using a Markov chain on junction trees". In: *Biometrika* 100.1 (2013), pp. 91–110.

[30] Jim E Griffin and Philip J Brown. "Structuring shrinkage: some correlated priors for regression". In: *Biometrika* 99.2 (2012), pp. 481–487.

[31] Joseph Guinness. "Permutation and Grouping Methods for Sharpening Gaussian Process Approximations". In: *Technometrics* just-accepted (2018).

[32] Heikki Haario, Eero Saksman, Johanna Tamminen, et al. "An adaptive Metropolis algorithm". In: *Bernoulli* 7.2 (2001), pp. 223–242.

[33] Peter Hammond and Michael Suttie. "Large-scale objective phenotyping of 3D facial morphology". In: *Human mutation* 33.5 (2012), pp. 817–825.

[34] Chris C Holmes, David G T Denison, and Bani K Mallick. "Accounting for model uncertainty in seemingly unrelated regressions". In: *Journal of Computational and Graphical Statistics* 11.3 (2002), pp. 533–551.

[35] Chris C Holmes, Leonhard Held, et al. "Bayesian auxiliary variable models for binary and multinomial regression". In: *Bayesian analysis* 1.1 (2006), pp. 145–168.

[36] Michael Inouye et al. "Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis". In: *PLoS genetics* 8.8 (2012), e1002907.

[37] Zhenyu Jia and Shizhong Xu. "Mapping quantitative trait loci for expression abundance". In: *Genetics* 176.1 (2007), pp. 611–623.

[38] Beatrix Jones et al. "Experiments in stochastic computation for high-dimensional graphical models". In: *Statistical Science* (2005), pp. 388–400.

[39] Johannes Kettunen, Ayşe Demirkan, et al. "Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA". In: *Nature communications* 7 (2016), p. 11122.

[40] Johannes Kettunen, Taru Tukiainen, et al. "Genome-wide association study identifies multiple loci influencing human serum metabolite levels". In: *Nature genetics* 44.3 (2012), p. 269.

[41] Lynn Kuo and Bani Mallick. "Variable selection for regression models". In: *Sankhyā: The Indian Journal of Statistics, Series B* (1998), pp. 65–81.

[42] Steffen L Lauritzen. *Graphical models, volume 17 of Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996.

[43] Alex Lewin et al. "MT-HESS: an efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues". In: *Bioinformatics* 32.4 (2016), pp. 523–532. DOI: `10.1093/bioinformatics/btv568`. URL: `http://dx.doi.org/10.1093/bioinformatics/btv568`.

[44] Faming Liang and Wing Hung Wong. "Evolutionary Monte Carlo: Applications to C_p model sampling and change point problem". In: *Statistica Sinica* 10 (2000), pp. 317–342.

[45] Pekka Marttinen et al. "Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression". In: *Bioinformatics* 30.14 (2014), pp. 2026–2034.

[46] Nicholas Mitsakakis, Hélene Massam, Michael D Escobar, et al. "A Metropolis-Hastings based method for sampling from the G-Wishart distribution in Gaussian graphical models". In: *Electronic Journal of Statistics* 5 (2011), pp. 18–30.

[47] Abdolreza Mohammadi, Ernst C Wit, et al. "Bayesian structure learning in sparse Gaussian graphical models". In: *Bayesian Analysis* 10.1 (2015), pp. 109–138.

[48] Trevor Park and George Casella. "The bayesian lasso". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 681–686.

[49] Vern I Paulsen, Stephen C Power, and Roger R Smith. "Schur products and matrix completions". In: *Journal of functional analysis* 85.1 (1989), pp. 151–178.

[50] Enrico Petretto et al. "New insights into the genetic control of gene expression using a Bayesian multi-tissue approach". In: *PLoS Computational Biology* 6.4 (2010), e1000737.

[51] Nicholas G Polson and James G Scott. "Shrink globally, act locally: Sparse Bayesian regularization and prediction". In: *Bayesian statistics* 9 (2010), pp. 501–538.

[52] Mohsen Pourahmadi. "Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation". In: *Biometrika* 86.3 (1999), pp. 677–690.

[53] Paula Rantakallio. "Groups at risk in low birth weight infants and perinatal mortality." In: *Acta Paediatrica Scandinavica* 193 (1969), Suppl–193.

[54] Sylvia Richardson, Leonardo Bottolo, and Jeffrey S Rosenthal. "Bayesian models for sparse regression analysis of high dimensional data". In: *Bayesian Statistics* 9 (2010), pp. 539–569.

[55] Gareth O Roberts and Jeffrey S Rosenthal. "Examples of adaptive MCMC". In: *Journal of Computational and Graphical Statistics* 18.2 (2009), pp. 349–367.

[56] Andrea Rodriguez-Martinez et al. "MWASTools: an R/Bioconductor package for metabolome-wide association studies". In: *Bioinformatics* 34.5 (2017), pp. 890–892.

[57] Alberto Roverato. "Cholesky decomposition of a hyper inverse Wishart matrix". In: *Biometrika* 87.1 (2000), pp. 99–112.

[58] Alberto Roverato. "Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models". In: *Scandinavian Journal of Statistics* 29.3 (2002), pp. 391–411.

[59] Helene Ruffieux et al. "Efficient inference for genetic association studies with multiple outcomes". In: *Biostatistics* 18.4 (2017), pp. 618–636.

[60] Chiara Sabatti et al. "Genome-wide association analysis of metabolic traits in a birth cohort from a founder population". In: *Nature genetics* 41.1 (2009), p. 35.

[61] Marie Pier Scott-Boyer et al. "An integrated hierarchical Bayesian model for multivariate eQTL mapping". In: *Statistical Applications in Genetics and Molecular Biology* 11.4 (2012).

[62] Shanya Sivakumaran et al. "Abundant pleiotropy in human complex diseases and traits". In: *The American Journal of Human Genetics* 89.5 (2011), pp. 607–618.

[63] Pasi Soininen, Antti J Kangas, Peter Würtz, Teemu Suna, et al. "Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics". In: *Circulation: Genomic and Precision Medicine* 8.1 (2015), pp. 192–206.

[64] Pasi Soininen, Antti J Kangas, Peter Würtz, Taru Tukiainen, et al. "High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism". In: *Analyst* 134.9 (2009), pp. 1781–1785.

[65] Francesco Stingo and Giovanni Marchetti. "Efficient local updates for undirected graphical models". In: 25 (Dec. 2014).

[66] Karsten Suhre et al. "Human metabolic individuality in biomedical and pharmaceutical research". In: *Nature* 477.7362 (2011), p. 54.

[67] Alun Thomas and Peter Green. "Enumerating the junction trees of a decomposable graph". In: *Journal of Computational and Graphical Statistics* 18.4 (2009), pp. 930–940.

[68] Caroline Uhler, Alex Lenkoski, Donald Richards, et al. "Exact formulas for the normalizing constants of Wishart distributions for graphical models". In: *The Annals of Statistics* 46.1 (2018), pp. 90–118.

[69] Aldo V Vecchia. "Estimation and model identification for continuous spatial processes". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1988), pp. 297–312.

[70] Hao Wang et al. "Bayesian graphical lasso models and efficient posterior computation". In: *Bayesian Analysis* 7.4 (2012), pp. 867–886.

[71] Yue Wang et al. "Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes". In: *BMC bioinformatics* 14.16 (2013), S6.

[72] Nanny Wermuth. "Linear recursive equations, covariance selection, and path analysis". In: *Journal of the American Statistical Association* 75.372 (1980), pp. 963–972.

[73] Arnold Zellner. "On assessing prior distributions and Bayesian regression analysis with g-prior distributions". In: *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti* 6 (1986), pp. 233–243.

[74] Arnold Zellner and Tomohiro Ando. "A direct Monte Carlo approach for Bayesian analysis of the seemingly unrelated regression model". In: *Journal of Econometrics* 159.1 (2010), pp. 33–45.

7 Appendix: Reparametrisation in terms of the precision matrix

We can decompose the precision matrix Λ in the same way as we decompose C:

$$\Lambda_{(k)} = \begin{pmatrix} \Lambda_{(k-1)} & \boldsymbol{\lambda}_k \\ \boldsymbol{\lambda}_k^t & \lambda_k \end{pmatrix} \tag{45}$$

for all $k = 2, \cdots, s$, with $\Lambda_{(s)} = \Lambda$ and $\Lambda_{(1)} = \lambda_1$ and $\boldsymbol{\lambda}_1$ null. Note that $\Lambda_{(k)}$ is not the marginal precision matrix for responses $1, \cdots, k$. For this we use a different decomposition:

$$C_{(k)}^{-1} = \begin{pmatrix} \Omega_{(k-1)} & \boldsymbol{\omega}_k \\ \boldsymbol{\omega}_k^t & \omega_k \end{pmatrix} \tag{46}$$

for all $k = 2, \cdots, s$, with $\Omega_{(s)} = \Lambda$, $\Omega_{(s-1)} = \Lambda_{(s-1)}$, $\Omega_{(1)} = \omega_1$ and $\boldsymbol{\omega}_1$ is null.

Recall the original decomposition

$$C_{(k)} = \begin{pmatrix} C_{(k-1)} & \boldsymbol{c}_k \\ \boldsymbol{c}_k^t & c_k \end{pmatrix} \tag{47}$$

Using standard block matrix inverse identities, we see that the decompositions in Equations 46 and 47 are related by

$$C_{(k)}^{-1}\boldsymbol{c}_k = -\boldsymbol{\omega}_k\omega_k^{-1} \tag{48}$$

$$c_k - \boldsymbol{c}_k^t C_{(k-1)}^{-1}\boldsymbol{c}_k = \omega_k^{-1} \tag{49}$$

Hence we can write down the new parameters in the transformed space as

$$\boldsymbol{\rho}_k = -\boldsymbol{\omega}_k\omega_k^{-1} \qquad k = 2, \cdots, s \tag{50}$$

$$\sigma_k^2 = \omega_k^{-1} \qquad k = 1, \cdots, s \tag{51}$$

with $\boldsymbol{\omega}_s = \boldsymbol{\lambda}_s$ and $\omega_s = \lambda_s$.

So we notice that the new parameters are directly related to the marginal precision matrices of the subsets of response variables, and hence the $\boldsymbol{\rho}$ directly relate to the partial correlations amongst the subsets of variables, with the remaining variables marginalised out.

We can use the Cholesky decomposition of C to derive a relation between the new parameters and the complete precision matrix Λ. Let P be a lower triangular matrix with $P_{kl} = \rho_{kl}$ $(k > l)$ and zeros on the diagonal. The reparametrisation of the covariance matrix converts

$$\boldsymbol{y}_k = X_{\gamma_k}\boldsymbol{\beta}_{\gamma_k} + \boldsymbol{u}_k, \quad \boldsymbol{u}_k \sim \mathcal{N}(\boldsymbol{0}, C) \tag{52}$$

into

$$\boldsymbol{y}_k = X_{\gamma_k}\boldsymbol{\beta}_{\gamma_k} + P\boldsymbol{u}_k + \boldsymbol{\varepsilon}_k, \quad \boldsymbol{\varepsilon}_k \sim \mathcal{N}(\boldsymbol{0}, S) \tag{53}$$

where S is a diagonal matrix such that $S_{kk} = \sigma_k^2$. Thus we see that

$$\boldsymbol{u}_k = (\mathbb{I} - P)^{-1}\boldsymbol{\varepsilon}_k \sim \mathcal{N}\left(\boldsymbol{0}, (\mathbb{I} - P)^{-1} S((\mathbb{I} - P)^{-1})^t\right) \tag{54}$$

Given that C is symmetric and positive definite, there exists a unique modified Cholesky decomposition such that $C = LDL^T$. We thus see that the reparametrisation is related to the Cholesky decomposition

with $L = (\mathbb{I} - P)^{-1}$ and $D = S$.

From this we can easily write the precision matrix in terms of the new parameters:

$$\Lambda = C^{-1} = (\mathbb{I} - P)^T S^{-1} (\mathbb{I} - P) \tag{55}$$

from which we can see that

$$\boldsymbol{\lambda}_k = -\sigma_k^{-2} \boldsymbol{\rho}_k + \sum_{m=k+1}^{s} \sigma_m^{-2} \rho_{m,k} \boldsymbol{\rho}_{m[1:k-1]} \tag{56}$$

for $k = 2, \cdots, s$.

8 Appendix: Proof that zeros in $\boldsymbol{\rho}_k$ correspond to zeros in Λ

The decomposable graph structure means that the precision matrix Λ for the entire graph can be written as

$$\Lambda = \begin{pmatrix} \Lambda_{H_Q \setminus S_Q} & \Lambda_{H_Q \setminus S_Q, S_Q} & 0 \\ \Lambda^t_{H_Q \setminus S_Q, S_Q} & \Lambda_{S_Q} & \Lambda_{S_Q, R_Q} \\ 0 & \Lambda^t_{S_Q, R_Q} & \Lambda_{R_Q} \end{pmatrix}, \tag{57}$$

where Q is the final prime component in the ordering above. The blocks of zeros are due to the conditional independence defined by the graph structure, since by definition R_Q is conditionally independent of $H_Q \setminus S_Q$ given the separator S_Q.

With the perfect elimination ordering, variable s is the final variable in the residual of the final prime component R_Q, corresponding to the final row/column in Λ as ordered here. For the variable s, we know from standard matrix operations that

$$\boldsymbol{\rho}_s \equiv -\boldsymbol{\lambda}_s \lambda_s^{-1} \tag{58}$$

so we can see here that all elements of the vector $\boldsymbol{\rho}_s$ corresponding to the partial correlations between variable s and variables in $H_Q \setminus S_Q$ are zero. Next consider the remaining variables in R_Q, labelling these $s-1, s-2, \ldots$. For any variable k, we have the recursion formula relating k to variables $m > k$:

$$\sigma_k^{-2} \boldsymbol{\rho}_k = \sum_{m>k} \sigma_m^{-2} \rho_{mk} \boldsymbol{\rho}_m - \boldsymbol{\lambda}_k \tag{59}$$

From this it is easy to see that for any variable k in the final residual R_Q, the zeros in $\boldsymbol{\rho}_k$ are directly inherited from the $\boldsymbol{\rho}_k, \boldsymbol{\rho}_{k+1}, \ldots, \boldsymbol{\rho}_s$. All these variables have the same zeros in the $\boldsymbol{\lambda}_k, \boldsymbol{\lambda}_{k+1}, \ldots, \boldsymbol{\lambda}_s$ sub-vectors of the precision matrix, so for all variables in R_Q, the vector $\boldsymbol{\rho}_k$ has zeros where $\boldsymbol{\lambda}_k$ has zeros, corresponding to partial correlations with variables in $H_Q \setminus S_Q$.

Now consider the residual of the next to final prime component, R_{Q-1}. By definition of the decomposable graph structure, the variables in R_{Q-1} are separated from variables in $H_{Q-1} \setminus S_{Q-1}$, and so $\Lambda_{H_{Q-1} \setminus S_{Q-1}, R_{Q-1}} = 0$. The running intersection property of a decomposable graph implies that R_Q must be separated from at least one of R_{Q-1} and $H_{Q-1} \setminus S_{Q-1}$. If R_Q is separated from R_{Q-1} then $\rho_{mk} = 0$ for all $m \in R_Q$ and $k \in R_{Q-1}$. If R_Q is separated from $H_{Q-1} \setminus S_{Q-1}$ then $\lambda_{kl} = 0 \implies \lambda_{ml} = 0 \implies \rho_{ml} = 0$ for all $m \in R_Q$ and $k \in R_{Q-1}$. In either case this leads to the

conclusion that $\lambda_{kl} = 0 \implies \rho_{kl} = 0$. Hence again, $\boldsymbol{\rho}_k$ has zeros where $\boldsymbol{\lambda}_k$ has zeros, corresponding to partial correlations between variables in R_{Q-1} and variables in $H_{Q-1} \setminus S_{Q-1}$.

The same argument applies to all successive residuals of the graph, so we see that for all variables, $\rho_{kl} = 0$ where $\Lambda_{kl} = 0$.

9 Appendix: Marginal precision matrix for subgraph

Here we consider decomposable graphs. We show that the marginal precision matrix $C_{H_q}^{-1}$ for each subgraph H_q has the same zero structure as the corresponding block of the precision matrix for the entire graph Λ_{H_q}. First see that we can partition the precision matrix into the final residual and the rest of the graph:

$$\Lambda = \begin{pmatrix} \Lambda_{H_Q} & \Lambda_{H_Q,R_Q} \\ \Lambda^t_{H_Q,R_Q} & \Lambda_{R_Q} \end{pmatrix} \tag{60}$$

conformably with the covariance matrix:

$$C = \begin{pmatrix} C_{H_Q} & C_{H_Q,R_Q} \\ C^t_{H_Q,R_Q} & C_{R_Q} \end{pmatrix} \tag{61}$$

and from these we have

$$C_{H_Q}^{-1} = \Lambda_{H_Q} - \Lambda_{H_Q,R_Q}\Lambda_{R_Q}^{-1}\Lambda^t_{H_Q,R_Q} \tag{62}$$

Next we partition Λ_{H_Q} into the next residual R_{Q-1}, its separator S_{Q-1} and the remainder:

$$\Lambda_{H_Q} = \begin{pmatrix} \Lambda_{H_{Q-1}\setminus S_{Q-1}} & \Lambda_{H_{Q-1}\setminus S_{Q-1},S_{Q-1}} & 0 \\ \Lambda^t_{H_{Q-1}\setminus S_{Q-1},S_{Q-1}} & \Lambda_{S_{Q-1}} & \Lambda_{S_{Q-1},R_{Q-1}} \\ 0 & \Lambda^t_{S_{Q-1},R_{Q-1}} & \Lambda_{R_{Q-1}} \end{pmatrix} \tag{63}$$

where the zero blocks appear by definition of the decomposability of the graph.

Now calculate the 2nd term in Equation 62:

$$\Lambda_{H_Q,R_Q}\Lambda_{R_Q}^{-1}\Lambda^t_{H_Q,R_Q} = \begin{pmatrix} \Lambda_{H_{Q-1}\setminus S_{Q-1},R_Q} \\ \Lambda_{S_{Q-1},R_Q} \\ \Lambda_{R_{Q-1},R_Q} \end{pmatrix} \Lambda_{R_Q}^{-1} \begin{pmatrix} \Lambda_{H_{Q-1}\setminus S_{Q-1},R_Q} & \Lambda_{S_{Q-1},R_Q} & \Lambda_{R_{Q-1},R_Q} \end{pmatrix} \tag{64}$$

The running intersection property of a decomposable graph implies that either (a) R_Q is separated from R_{Q-1} or (b) R_Q is separated from $H_{Q-1} \setminus S_{Q-1}$. If (a), then $\Lambda_{R_{Q-1},R_Q} = 0$. If (b), then $\Lambda_{H_{Q-1}\setminus S_{Q-1},R_Q} = 0$.

In either case this leads to the block decomposition for the marginal precision matrix as

$$C_{H_Q}^{-1} = \begin{pmatrix} \Omega_{H_{Q-1}\setminus S_{Q-1}} & \Omega_{H_{Q-1}\setminus S_{Q-1},S_{Q-1}} & 0 \\ \Omega^t_{H_{Q-1}\setminus S_{Q-1},S_{Q-1}} & \Omega_{S_{Q-1}} & \Omega_{S_{Q-1},R_{Q-1}} \\ 0 & \Omega^t_{S_{Q-1},R_{Q-1}} & \Omega_{R_{Q-1}} \end{pmatrix} \tag{65}$$

that is the zero-structure of the overall graph is maintained for a subgraph where the nodes in R_Q

have been removed.

The same argument can be applied when removing successive residuals R_{Q-1}, R_{Q-2} etc, since each time the removal results in another decomposable graph.

10 Appendix : Computational Advantages of the reparametrised sparse model

The reparametrisation from (6) to (8) is computationally beneficial in multiple ways. The most obvious one is that the likelihood is now computed as a product of independent factors, and thus its computation can be trivially parallelised once we know $X\beta$ and $U\boldsymbol{\rho}$ and $\boldsymbol{\sigma}^2$.

Sparsity in both the variable selection coefficients, contrarily to shrinkage priors like the horseshoe (Carvalho, Polson, and Scott, 2010) for example, and in the graph structure, as described by (Carvalho, Massam, and West, 2007), allow us to work at each iteration with only a subset of the coefficients, reducing thus the computational cost of inverting $p \times p$ and $s \times s$ matrices to operations with cubic scaling only with respect to the maximum number of γ coefficients selected for a single outcome and the maximum number of vertices in a Prime component; *i.e* with respectively $\max_k \sum_j \gamma_{kj}$ and $\max_q |P_q|$.

Another, perhaps less obvious, advantage is the fact that the availability of all the full conditionals for a given index j (in Eq (68)) allows us to update only a subset of the coefficients when we update both Γ or G, proposing thus less *daring* move in the parameter space that have a higher probability of being accepted.

11 Appendix: Detailed sampler specification

Here we will details how the sampling of each parameter is performed. We start by re-stating the model and all the priors and then describe each Gibbs or Metropolis-within-Gibbs update for each parameter conditioned on all the others.

$$Y = X_\Gamma \beta + U \tag{66}$$

$$U \sim \mathcal{MN}\left(\mathbf{0}, \mathbb{I}_n, C\right) \tag{67}$$

where C is an $s \times s$ general covariance matrix, which we reparametrise into

$$y_{\xi_k} = X_{\Gamma_{\xi_k}.} \beta_{\xi_k} + \sum_{l<k} u_{\xi_l} \rho_{\xi_k \xi_l} + \varepsilon_{\xi_k} \qquad k = 1, \dots s \tag{68}$$

$$\varepsilon_j \sim \mathcal{N}\left(0, \sigma_j^2\right) \tag{69}$$

$$C \sim \mathcal{HIW}_G\left(\nu, \tau \mathbb{I}_s\right) \tag{70}$$

$$\tag{71}$$

with ξ being a (non-unique) global *perfect elimination order*, determined by G.
In the reparametrised space, this $\mathcal{HIW}$ prior implies

$$\sigma^2_{qt} \sim IG\left((\nu - s + t + |S_q|)/2, \tau/2\right) \qquad q = 1, \dots Q\,,\, t = 1, \dots |R_q| \tag{72}$$

$$\boldsymbol{\rho}_{qt} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{rt}/\tau\right) \tag{73}$$

where Q is the number of Prime components P_q in the decomposable graph G, S_q is the separator associated with component P_q, with the convention that $S_1 = \emptyset$, and is R_q the corresponding residual component. It is also assumed that inside each component the nodes are ordered in accordance with ξ.

$$\beta_{kj}|\gamma_{kj} = 1 \sim \mathcal{N}\left(0, w\right) \qquad j = 1, \dots p \tag{74}$$

$$\beta_{kj}|\gamma_{kj} = 0 \sim \delta_0\left(\beta_{kj}\right) \tag{75}$$

$$\gamma_{kj} \sim Bernoulli\left(\omega_{kj}\right) \tag{76}$$

$$\omega_{kj} = o_k \times \pi_j \tag{77}$$

$$\pi_j \sim \mathcal{G}amma(a_\pi, b_\pi) \tag{78}$$

$$o_k \sim \mathcal{B}eta(a_o, b_o) \tag{79}$$

$$\text{under the constraint that } o_k \times \pi_j \leq 1 \qquad \forall j, k \tag{80}$$

$$\mathbb{P}r(E_{kl} \in G) = \eta \qquad \text{(prior for } G\text{, indep } \forall \text{ edges)} \tag{81}$$

$$\eta \sim \mathcal{B}eta\left(a_\eta, b_\eta\right) \tag{82}$$

$$\tag{83}$$

$$\tau \sim \mathcal{G}amma(a_\tau, b_\tau) \tag{84}$$

$$w \sim IG(a_w, b_w) \tag{85}$$

In the following we will use $p(\cdot)$ to indicate the prior distributions and $\ell(\cdot)$ for the likelihood function; we will avoid including explicitly the Jacobian of the transformation (68) as it will usually be included in the posterior full conditional for $\left(\boldsymbol{\sigma}^2, \boldsymbol{\rho}\right)$. $q(\cdot)$ will usually be reserved for proposal probability distributions.

Following Bottolo and Richardson (2010) we will also define multiple chains $c = 1, \dots N_c$, but as our likelihood (6) is not factorisable, we will use a slightly less complex tempering scheme. In particular each chain will only have one temperature t_c associated (rather than one for each outcome) and we will only temper the likelihood and not the already diffused priors.

11.1 τ sampling

We use a simple random walk Metropolis update for τ; given that the parameter is constrained to be positive, we propose the new point τ' using a symmetric Normal jump on the log-space, centred in zero and with variance σ^2_τ:

$$\tau' = \exp\{\log(\tau) + Z_\tau\} \quad , \quad Z_\tau \sim \mathcal{N}\left(0, \sigma_\tau^2\right) \tag{86}$$

The MH ratio is then simply

$$\min\left(1, \frac{p_\tau(\tau'|a_\tau, b_\tau) \prod_{k=1}^{s} \left[p_{\sigma_k^2}(\sigma_k^2|G, \nu, \tau') p_{\boldsymbol{\rho}_k}(\boldsymbol{\rho}_k|\sigma_k^2, G, \tau')\right]}{p_\tau(\tau|a_\tau, b_\tau) \prod_{k=1}^{s} \left[p_{\sigma_k^2}(\sigma_k^2|G, \nu, \tau') p_{\boldsymbol{\rho}_k}(\boldsymbol{\rho}_k|\sigma_k^2, G, \tau')\right]}\right) \tag{87}$$

where the factorisation over k is a direct consequence of the diagonal scale matrix for the prior over C.

11.2 η sampling

Since the prior over the graph G is a product of independent Bernoulli on the edge-set $\mathcal{E}$, the Beta prior on η is conjugate and we can analytically derive its posterior full-conditional distribution.

The update follows then

$$\eta|a_\eta, b_\eta, G \sim \mathcal{B}eta\left(a_\eta + \sum_{k,k'} \mathbf{1}_{[E_{k,k'} \in \mathcal{E}]}, b_\eta + \left\{s(s-1)/2 - \sum_{k,k'} \mathbf{1}_{[E_{k,k'} \in \mathcal{E}]}\right\}\right) \tag{88}$$

where $\sum_{k,k'} \mathbf{1}_{[E_{k,k'} \in \mathcal{E}]}$ essentially counts how many edges are present in the current graph.

11.3 G sampling

Once a new junction tree jt' has been proposed from the aforementioned Junction Tree sampler, we compute deterministically its set of prime components P', separators S' and residual components R' as well as its modified *perfect elimination ordering* ξ'. Since we are not able to analytically integrate out the elements of the residual covariance matrix we also need to sample a proposed point $\left({\sigma'}_{qt}^2, \rho'_{qt}\right)$ for them (see section 11.8 for details) that is in accordance with the newly proposed graph / junction tree.

When associated with a tempered chain, these posterior full conditionals changes slightly to

$$\begin{aligned} {\sigma'}_{qt}^2 &\sim IG((n + \nu - s + t + |S_q|)/2, (\bar{m}_{qt} - \bar{\boldsymbol{m}}_{qt}^t \bar{M}_{P_q,(t-1)}^{-1} \bar{\boldsymbol{m}}_{qt})/2) \\ \boldsymbol{\rho}'_{qt}|{\sigma'}_{qt}^2 &\sim N(\bar{M}_{P_q,(t-1)}^{-1} \bar{\boldsymbol{m}}_{qt}, \sigma_{qt}^2 \bar{M}_{P_q,(t-1)}^{-1}) \end{aligned}$$

where

$$\bar{M} = \tau \mathbb{I}_s + \frac{1}{t_c}\left(U^t U\right)$$

Once the proposed point is complete, we accept it with probability

$$\min\left(1, \frac{p_G(G'|\eta) \prod_{k=1}^{s} \left[p_{\sigma_k^2}({\sigma'}_k^2|G', \nu, \tau) p_{\boldsymbol{\rho}_k}(\boldsymbol{\rho}'_k|{\sigma'}_k^2, G', \tau)\right] \left[\ell(Y, X|\beta, \Gamma, \boldsymbol{\sigma'}^2, \boldsymbol{\rho}')\right]^{1/t_c}}{p_G(G|\eta) \prod_{k=1}^{s} \left[p_{\sigma_k^2}(\sigma_k^2|G, \nu, \tau) p_{\boldsymbol{\rho}_k}(\boldsymbol{\rho}_k|\sigma_k^2, G, \tau)\right] \left[\ell(Y, X|\beta, \Gamma, \boldsymbol{\sigma}^2, \boldsymbol{\rho})\right]^{1/t_c}} \times r\right) \tag{89}$$

where $r = \frac{q_G(G|G') \prod_{k=1}^{s} \left[q_{\sigma_k^2, t_c}(\sigma_k^2 | \xi(G), \nu, \tau, Y, X, \boldsymbol{\beta}) q_{\boldsymbol{\rho}_k, t_c}(\boldsymbol{\rho}_k | \sigma_k^2, \xi(G), \tau, Y, X, \boldsymbol{\beta}) \right]}{q_G(G'|G) \prod_{k=1}^{s} \left[q_{\sigma_k^2, t_c}(\sigma'^2_k | \xi'(G'), \nu, \tau, Y, X, \boldsymbol{\beta}) q_{\boldsymbol{\rho}_k, t_c}(\boldsymbol{\rho}'_k | \sigma'^2_k, \xi(G'), \tau, Y, X, \boldsymbol{\beta}) \right]}$ is the proposal probabilities ratio.

Note that now that the likelihood is involved we also need to take into account t_c, the temperature associated with the current chain c.

We need to consider as well that given different *perfect elimination order*, we are effectively changing the conditioning order in our reparametrisation. This makes it hard to track the actual values for $\left(\boldsymbol{\sigma}^2, \boldsymbol{\rho}\right)$ but, as the transformation between these and the original covariance matrix C is bijective, structure learning doesn't pose any problem.

By considering that our proposal for $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}^2$ (when we sample all the coefficients in particular) is actually equivalent to the standard Gibbs proposal for C given in Equation (**??**), essentially its posterior full conditional, in the MH ratio r will simplify with the relevant terms in the posterior distributions ratio, thus eliminating the dependence of the acceptance probability on this coefficients.

11.4 π sampling

To update each π_j we perturb the current value using a Normal random walk in the log-space, similarly to Sec 11.1.

$$\pi'_j = \exp\left\{\log\left(\pi_j\right) + Z_\pi\right\} \quad , \quad Z_\pi \sim \mathcal{N}\left(0, \sigma_\pi^2\right) \tag{90}$$

Assuming Eq (80) is satisfied (the proposal is otherwise immediately rejected), the acceptance probability is

$$\min\left(1, \frac{p_{\pi_j}(\pi'_j | a_\pi, b_\pi) \prod_{k=1}^{s} p_{\gamma_{kj}}(\gamma_{kj} | o_k, \pi'_j)}{p_{\pi_j}(\pi_j | a_\pi, b_\pi) \prod_{k=1}^{s} p_{\gamma_{kj}}(\gamma_{kj} | o_k, \pi_j)}\right). \tag{91}$$

We repeat this procedure for a random subset of all j at each iteration.

11.5 o sampling

Once again we use a random walk Metropolis-Hastings perturbation of o_k that proceeds in a similarly way as the update for π_j; the only difference being that the proposal is from a truncated Normal $\mathcal{N}(0, \sigma_o^2, -\infty, -\log\left(o_k\right))$ in order to ensure that $o'_k \in (0, 1)$.

The update is therefore

$$o'_k = \exp\left\{\log\left(o_k\right) + Z_o\right\} \quad , \quad Z_o \sim \mathcal{N}\left(0, \sigma_o^2, -\infty, -\log\left(o_k\right)\right) \tag{92}$$

The acceptance probability is

$$\min\left(1, \frac{p_{o_k}(o'_k|a_o,b_o)\left[\prod_{j=1}^{p} p_{\gamma_{kj}}(\gamma_{kj}|o'_k,\pi_j)\right] q_{o_k}(o_k|o'_k)}{p_{o_k}(o_k|a_o,b_o)\left[\prod_{j=1}^{p} p_{\gamma_{kj}}(\gamma_{kj}|o_k,\pi_j)\right] q_{o_k}(o'_k|o_k)}\right). \tag{93}$$

still assuming Eq (80) is satisfied; note moreover that we needed to include an extra ratio this time as the truncation makes the proposal asymmetric.

As for π_j, we randomly select a subset of all possible indexes k to update at each iteration.

11.6 γ sampling

We propose a new value for the matrix of binary variable selection coefficients Γ using Banterle & Lewin (2018b) ? How much detail should we provide here?

For similar reasons as in Sec 11.3, once the new Γ' matrix is proposed we need to update the corresponding β coefficients accordingly. In order to avoid updating the whole matrix though, we will only update the columns that correspond to outcomes that had their γ_k vector modified. This is possible because the posterior full conditional for β is Normal (see Sec 11.8) and thus recovering its conditional distribution is trivial.

Again if connected with a tempered chain the proposal changes slightly to

$$\boldsymbol{\beta}_k|\boldsymbol{\sigma}^2,\boldsymbol{\rho},\xi \sim \mathcal{N}\left(W_k \times \frac{X^t_{\Gamma_{k,\cdot}}\tilde{y}_k}{t_c}\ ,\ W_k\right)$$

where

$$W_k = \left(\frac{1}{t_c}X^t_{\Gamma_{k,\cdot}}X_{\Gamma_{k,\cdot}}\left(\frac{1}{\sigma_k^2}+\sum_{l=\xi(k)+1}^{s}\frac{\rho^2_{\xi_l\,k}}{\sigma^2_{\xi_l}}\right)+(w\mathbb{I}_{p_k})^{-1}\right)^{-1}.$$

Once a new β' has been sampled, we compute its acceptance probability as follows

$$\min\left(1, \frac{\prod_{k\in\mathcal{K}}\left[p_{\gamma_k}(\gamma'_k|\eta,\boldsymbol{o},\boldsymbol{\pi})p_{\beta_k}\left(\beta'_k|\gamma'_k,w\right)\right]\left[\ell(Y,X|\boldsymbol{\beta}',\Gamma',\boldsymbol{\sigma}^2,\boldsymbol{\rho})\right]^{1/t_c}}{\prod_{k\in\mathcal{K}}\left[p_{\gamma_k}(\gamma_k|\eta,\boldsymbol{o},\boldsymbol{\pi})p_{\beta_k}\left(\beta_k|\gamma_k,w\right)\right]\left[\ell(Y,X|\boldsymbol{\beta},\Gamma,\boldsymbol{\sigma}^2,\boldsymbol{\rho})\right]^{1/t_c}}\times r\right) \tag{94}$$

where $\mathcal{K}$ is the set of indexes where Γ and $\boldsymbol{\beta}$ were modified to obtain Γ' and $\boldsymbol{\beta}'$ and

$$r = \frac{\prod_{k\in\mathcal{K}}\left[q_{\gamma_k}(\gamma_k|\gamma'_k)q_{\beta_k}\left(\beta_k|\gamma_k,w,Y,X,\boldsymbol{\sigma}^2,\boldsymbol{\rho},\xi\right)\right]}{\prod_{k\in\mathcal{K}}\left[q_{\gamma_k}(\gamma'_k|\gamma_k)q_{\beta_k}\left(\beta'_k|\gamma'_k,w,Y,X,\boldsymbol{\sigma}^2,\boldsymbol{\rho},\xi\right)\right]}.$$

Note again that because we use the posterior full conditional as a proposal $q_{\beta_k}\left(\beta_k|\gamma_k,w,Y,X,\boldsymbol{\sigma}^2,\boldsymbol{\rho},\xi\right)$, the terms relative to β_k will simplify from the posterior distribution and the acceptance ratio will depend only on $\boldsymbol{\gamma}$. The resulting expression is a special case of the marginal distribution for $\boldsymbol{\gamma}$ given for example in Holmes, Denison, and Mallick, 2002.

11.7 w sampling

Finally, the update for the prior variance of the slab component in the β prior is again a Gibbs sampling update. w has an Inverse-Gamma prior distribution and thus is conjugate to the prior over the non-zero $\boldsymbol{\beta}$ elements.

The update is thus

$$w' \sim IG\left(a_w + 0.5 \times \sum_{j,k} \gamma_{kj} \ , \ b_w + 0.5 \times \sum_{j,k} \beta_{kj}^2\right) \tag{95}$$

where we update the prior parameters using the number of non-zero regression coefficients and their second moment.

11.8 $\boldsymbol{\beta}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\rho}$ sampling

$\boldsymbol{\beta}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\rho}$ are all updated by sampling from their (possibly tempered) posterior full-conditionals.

See Sec 3.1 for additional details.

11.9 Proposal variances adaptation

For all the Normal Metropolis-Hastings updates (τ , o_k and π_j) we use a simple adaptive scheme (see Haario, Saksman, and Tamminen (2001), Roberts and Rosenthal (2009)) that targets the asymptotically optimal proposal distribution. We start by setting a reasonably small value $\sigma_*^2 = s_*^{(0)}$ and then every once in a while we update it using

$$\sigma_*^2 = 0.05\left(s_*^{(0)}\right) + 0.95\left(2.38^2 \times s_*^{(i)}\right) \tag{96}$$

where $* = (\tau, o, \pi)$ and $s_*^{(i)}$ is the posterior sample variance computed using the MCMC chain up to the current iteration i. To avoid wasting computations we keep track of the first two moments of each parameters as the Markov chain progresses. Finally for o and π a vector of these is obtained for each index k and j and we average them to get the actual proposal variance.

12 Appendix: Naïve Gibbs sampler for the SUR model

As mentioned in the main text, a simple expression for the posterior full-conditionals on $\boldsymbol{\beta}, C$ is available but potentially computationally expensive is little care is given in optimising the matrix operations within.

Moreover the likelihood in the original parameters' space is $\mathcal{O}\left(ns^3\right)$ and not trivially parallelisable.

We give nonetheless the full conditionals below:

$$C|\boldsymbol{\beta}, G \sim \mathcal{HIW}_G\left(\nu + n, M + \tau \mathbb{I}_s\right) \tag{97}$$

$$vec(\beta)|\boldsymbol{\gamma}, C \sim \mathcal{N}\left(\Sigma_\beta \times \mathcal{X}^t \left(C^{-1} \otimes \mathbb{I}_n\right) vec(Y), \Sigma_\beta\right) \tag{98}$$

$$[S]_{kl} = \left(y_k - X_{\boldsymbol{\gamma}_k} \boldsymbol{\beta}_{\boldsymbol{\gamma}_k}\right)^t \left(y_l - X_{\boldsymbol{\gamma}_l} \boldsymbol{\beta}_{\boldsymbol{\gamma}_l}\right)$$

$$\Sigma_\beta = \left(\mathcal{X}^t \left(C^{-1} \otimes \mathbb{I}_n\right) \mathcal{X} + w \times \mathbb{I}_{\tilde{p}}\right)^{-1}.$$