

BayesSUR: An R package for high-dimensional multivariate Bayesian variable and covariance selection in linear regression

Marco Banterle*
London School of Hygiene
& Tropical Medicine

Zhi Zhao*
University of Oslo

Leonardo Bottolo
University of Cambridge
Alan Turing Institute

Sylvia Richardson
University of Cambridge
Alan Turing Institute

Alex Lewin#
London School of Hygiene
& Tropical Medicine

Manuela Zucknick#
University of Oslo

Abstract

In molecular biology, advances in high-throughput technologies have made it possible to study complex multivariate phenotypes and their simultaneous associations with high-dimensional genomic and other omics data, a problem that can be studied with high-dimensional multi-response regression, where the response variables are potentially highly correlated.

To this purpose, we recently introduced several multivariate Bayesian variable and covariance selection models, e.g., Bayesian estimation methods for sparse seemingly unrelated regression for variable and covariance selection. Several variable selection priors have been implemented in this context, in particular the hotspot detection prior for latent variable inclusion indicators, which results in sparse variable selection for associations between predictors and multiple phenotypes. Here, we also propose an alternative, which uses a Markov random field (MRF) prior for incorporating prior knowledge about the dependence structure of the inclusion indicators. Inference of Bayesian seemingly unrelated regression (SUR) by Markov chain Monte Carlo methods is made computationally feasible by factorisation of the covariance matrix amongst the response variables.

In this paper we present **BayesSUR**, an R package, which allows the user to easily specify and run a range of different Bayesian SUR models, which have been implemented in C++ for computational efficiency. The R package allows the specification of the models in a modular way, where the user chooses the priors for variable selection and for covariance selection separately. We demonstrate the performance of sparse SUR models with the hotspot prior and spike-and-slab MRF prior on synthetic and real data sets representing eQTL or mQTL studies and *in vitro* anti-cancer drug screening studies as examples for typical applications.

Keywords: Seemingly unrelated regression, Bayesian multivariate regression, structured covariance matrix, Markov random field prior, multi-omics data.

1. Introduction

With the development of high-throughput technologies in molecular biology, the large-scale molecular characterisation of biological samples has become common-place, for example by genome-wide measurement of gene expression, single nucleotide polymorphisms (SNP) or CpG methylation status. Other complex phenotypes, for example, pharmacological profiling from large-scale cancer drug screens, are also measured in order to guide personalized cancer therapies (Garnett, Edelman, Heidorn, Greenman, Dastur, Lau, Greninger, Thompson, Luo, Soares, Liu, and *others* 2012; Barretina, Caponigro, Stransky, Venkatesan, Margolin, Kim, Wilson, Lehar, Kryukov, Sonkin, and *others* 2012; Gray and Mills 2015). The analysis of joint associations between multiple correlated phenotypes and high-dimensional molecular features becomes challenging.

When multiple phenotypes and high-dimensional genomic information are jointly analyzed, the Bayesian framework allows to specify in a flexible manner the complex relationships between the highly structured data sets. Much work has been done in this area in recent years. Our software package **BayesSUR** gathers together several models that we have proposed for high-dimensional regression of multiple responses, allowing for different priors for variable selection in the regression models and for different assumptions about the dependence structure between responses.

Bayesian variable selection uses latent indicator variables to explicitly add or remove covariates in each regression during the model search. Here, as we consider simultaneously many predictors and several responses, we have a matrix of variable selection indicators. Different variable selection priors have been proposed in the literature. For example, Jia and Xu (2007) mapped multiple phenotypes to genetic markers (i.e., eQTL) using the spike-and-slab prior and hyper predictor-effect prior. Liquet, Mengersen, Pettitt, and Sutton (2017) incorporated group structures of multiple predictors via a (multivariate) spike-and-slab prior. The corresponding R package **MBSGS** is available on CRAN (<https://cran.r-project.org/packages=MBSGS>). Bottolo, Petretto, Blankenberg, Cambien, Cook, Tiret, and Richardson (2011) and Lewin, Saadi, Peters, Moreno-Moral, Lee, Smith, Petretto, Bottolo, and Richardson (2015) further proposed the hotspot prior for variable selection in multivariate regression, in which the probability of association between the predictor and response is decomposed multiplicatively into predictor and response random effects. This prior is implemented in a multivariate Bayesian hierarchical regression setup in the software **R2HESS**, available from <https://www.mrc-bsu.cam.ac.uk/software/>. Lee, Tadesse, Baccarelli, Schwartz, and Coull (2017) used the Markov random field (MRF) prior to encourage joint selection of the same variable across several correlated response variables. Their C-based R package **mBvs** is available on CRAN (<https://CRAN.R-project.org/package=mBvs>).

For high-dimensional predictors and multivariate responses, the space of models is very large. To overcome the infeasibility of the enumerated model space for the MCMC samplers in the high dimensional situation, Bottolo and Richardson (2010) proposed an Evolutionary Stochastic Search (ESS) algorithm based on Evolutionary Monte Carlo. This sampler has been extended in a number of situations. An efficient implementation of ESS for multivariate Bayesian hierarchical regression is provided with the C++-based R package **R2GUESS** (<https://CRAN.R-project.org/package=R2GUESS>) (Liquet, Bottolo, Campanella, Richardson, and Chadeau-Hyam 2016). Richardson, Bottolo, and Rosenthal (2011) proposed a new model and computationally efficient hierarchical evolutionary stochastic search algorithm (HESS) for multi-response regression which assumes independence between residuals across responses and is implemented in the **R2HESS** package. Petretto, Bottolo, Langley, Heinig,

Mcdermott-Roe, Sarwar, Pravenec, Hubner, Aitman, Cook, and Richardson (2010) used the inverse Wishart prior on the covariance matrix of residuals in order to do simultaneous analysis of multiple response variables allowing for correlations in response residuals, for more moderate sized data sets.

In order to analyse larger numbers of response variables, yet retain the ability to estimate dependence structures between them, sparsity can be introduced into the residual covariances, as well as into the regression model selection. Holmes, Denison, and Mallick (2002) adapted seemingly unrelated regression (SUR) to the Bayesian framework and used a Markov chain Monte Carlo (MCMC) algorithm for the analytically intractable posterior inference. The hyper-inverse Wishart prior has been used to learn a sparser graph structure for the covariance matrix of high-dimensional variables (Carvalho, Massam, and West 2007; Wang 2010; Bhadra and Mallick 2013), thus performing covariance selection. However, these approaches are not computationally feasible if the number of input variables is very large. Banterle, Bottolo, Richardson, Ala-Korpela, Jarvelin, and Lewin (2018) recently developed a Bayesian variable selection model which employs the hotspot prior for variable selection, learns a structured covariance matrix and implements the ESS algorithm in the SUR framework to further improve computational efficiency.

The **BayesSUR** package implements many of these possible choices for high-dimensional multi-response regressions by allowing the user to choose among three different prior structures for the residual covariance matrix and among three priors for the joint distribution of the variable selection indicators. It employs ESS as a basic variable selection algorithm.

2. Models specification

The **BayesSUR** package fits a Bayesian seemingly unrelated regression model with a number of options for variable selection, and where the covariance matrix structure is allowed to be diagonal, dense or sparse. It encompasses three classes of Bayesian multi-response linear regression models: Hierarchical Related Regressions (HRR, Richardson *et al.* (2011)), dense and Sparse Seemingly Unrelated Regressions (dSUR and SSUR, Banterle *et al.* (2018)), and the newly introduced Structured Seemingly Unrelated Regression, which makes use of a Markov random field (MRF) prior.

The regression model is written as

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\mathbf{B} + \mathbf{U}, \\ \text{vec}(\mathbf{U}) &\sim \mathcal{N}(\mathbf{0}, C \otimes \mathbb{I}_n) \end{aligned} \tag{1}$$

where \mathbf{Y} is a $n \times s$ matrix of outcome variables with $s \times s$ covariance matrix C , \mathbf{X} is a $n \times p$ matrix of predictors for all outcomes and \mathbf{B} is a $p \times s$ matrix of regression coefficients.

We use a binary latent indicator matrix $\mathbf{\Gamma} = \{\gamma_{jk}\}$ to perform variable selection. A spike-and-slab prior is used to find a sparse relevant subset of predictors that explain the variability of \mathbf{Y} : conditional on $\gamma_{jk} = 0$ ($j = 1, \dots, p$ and $k = 1, \dots, s$) we set $\beta_{jk} = 0$ and conditional on $\gamma_{jk} = 1$ non-zero regression coefficients follow a diffuse Normal:

$$\beta_{\gamma} | \gamma \sim \mathcal{N}(\mathbf{0}, W_{\gamma}^{-1}). \tag{2}$$

where $\beta = \text{vec}(\mathbf{B})$, $\gamma = \text{vec}(\mathbf{\Gamma})$, β_{γ} consists of the non-zero regression coefficients only, and W_{γ} is the sub-matrix of W formed by the corresponding non-zero coefficients.

The precision matrix W is generally decomposed into a shrinkage coefficient and a matrix that governs the covariance structure of the regression coefficients. Here we use $W = \lambda^{-1}\mathbb{I}_{sp}$, meaning that all the regression coefficients are *a priori* independent, with an inverse gamma hyperprior on the shrinkage coefficient λ , i.e., $\lambda \sim \mathcal{IGamma}(a_\lambda, b_\lambda)$. The binary latent indicator matrix $\mathbf{\Gamma}$ has three possible options for priors: the independent hierarchical Bernoulli prior, the hotspot prior and the MRF prior. The covariance matrix C also has three possible options for priors: the independent inverse gamma prior, the inverse Wishart prior and hyper-inverse Wishart prior. Thus, we consider nine possible models (Table 1) across all combinations of three priors for C and three priors for $\mathbf{\Gamma}$.

| | $\gamma_{jk} \sim \text{Bernoulli}$ | $\gamma_{jk} \sim \text{Hotspot}$ | $\gamma \sim \text{MRF}$ |
|--------------------------|-------------------------------------|-----------------------------------|--------------------------|
| $C \sim \text{indep}$ | HRR-B | HRR-H | HRR-M |
| $C \sim \mathcal{IW}$ | dSUR-B | dSUR-H | dSUR-M |
| $C \sim \mathcal{HIW}_G$ | SSUR-B | SSUR-H | SSUR-M |

Table 1: Nine models across three priors of C by three priors of $\mathbf{\Gamma}$

2.1. Hierarchical Related Regression (HRR)

The Hierarchical Related Regression model assumes that C is a diagonal matrix

$$C = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \sigma_s^2 \end{pmatrix}, \quad (3)$$

which translates into conditional independence between the multiple response variables, so the likelihood factorises across responses. An inverse gamma prior is specified for the residual covariance, i.e., $\sigma_k^2 \sim \mathcal{IGamma}(a_{\sigma_k^2}, b_{\sigma_k^2})$ which, combined with the priors in (2) is conjugate with the model in (1). We can thus sample the variable selection structure $\mathbf{\Gamma}$ marginally with respect to C and \mathbf{B} . For inference for this model, Richardson *et al.* (2011) implemented the hierarchical evolutionary stochastic search algorithm (HESS).

2.1.1. HRR with independent Bernoulli prior

For a simple independent prior on the regression model selection, the binary latent indicators follow a Bernoulli prior

$$\gamma_{jk} | \omega_{jk} \sim \text{Ber}(\omega_j), \quad j = 1, \dots, p, \quad k = 1, \dots, s, \quad (4)$$

with a further hierarchical Beta prior on ω_j , i.e., $\omega_j \sim \text{Beta}(a_\omega, b_\omega)$, which quantifies the probability for each predictor to be associated with any response variable.

2.1.2. HRR with hotspot prior

Richardson *et al.* (2011) and Bottolo *et al.* (2011) proposed decomposing the probability of association parameter ω_{jk} in (4) as $\omega_{jk} = o_k \times \pi_j$, where o_k accounts for the sparsity of each response model and π_j controls the propensity of each predictor to be associated with multiple

responses simultaneously.

$$\begin{aligned} \gamma_{jk}|\omega_{jk} &\sim \text{Ber}(\omega_{jk}), \quad j = 1, \dots, p, \quad k = 1, \dots, s, \\ \omega_{jk} &= o_k \times \pi_j, \\ o_k &\sim \text{Beta}(a_o, b_o), \\ \pi_j &\sim \text{Gamma}(a_\pi, b_\pi). \end{aligned} \tag{5}$$

2.1.3. HRR with MRF prior

To consider the relationship between different predictors and associate highly correlated responses with the same predictors, we set a Markov random field prior on the latent binary vector γ

$$f(\gamma|d, e, G) \propto \exp\{d\mathbf{1}^\top \gamma + e\gamma^\top G\gamma\}, \tag{6}$$

where G is an adjacency matrix containing prior information about similarities amongst the binary model selection indicators $\gamma = \text{vec}(\mathbf{\Gamma})$. The parameters d and e are treated as fixed in the model, but the usual procedure for estimating MRF models is to run with a grid of values for these parameters, in order to detect the phase transition boundary for e (Lee *et al.* 2017) and to identify a sensible combination of d and e that corresponds to prior expectations of overall model sparsity and sparsity for the MRF graph.

2.2. Dense Seemingly Unrelated Regression (dSUR)

The HRR models in Section 2.1 assume residual independence between any two response variables because of the diagonal matrix C in (3). It is possible to estimate a full covariance matrix by specifying an inverse Wishart prior, i.e., $C \sim \mathcal{IW}(v, \tau\mathbb{I}_s)$. To avoid estimating the dense and large covariance matrix directly, Banterle *et al.* (2018) exploited a factorisation of the dense covariance matrix to transform the parameter space (v, τ) of the inverse Wishart distribution to space $\{\sigma_k^2, \rho_{kl}|\sigma_k^2 : k = 1, \dots, s; l < k\}$, with priors

$$\begin{aligned} \sigma_k^2 &\sim \mathcal{IGamma}\left(\frac{v - s + 2k - 1}{2}, \frac{\tau}{2}\right), \\ \rho_{kl}|\sigma_k^2 &\sim \mathcal{N}\left(0, \frac{\sigma_k^2}{\tau}\right). \end{aligned} \tag{7}$$

Here, we assume that $\tau \sim \text{Gamma}(a_\tau, b_\tau)$. Thus, model (1) is rewritten as

$$\begin{aligned} \mathbf{y}_k &= \mathbf{X}\beta_k + \sum_{l < k} \mathbf{u}_l \rho_{kl} + \boldsymbol{\epsilon}_k, \quad k = 1, \dots, s, \\ \boldsymbol{\epsilon}_k &\sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbb{I}_n), \end{aligned} \tag{8}$$

where $\mathbf{u}_l = \mathbf{y}_l - \mathbf{X}\beta_l$ and β_l is the l th column of \mathbf{B} , so again the likelihood is factorised across responses.

Similarly to the HRR model, employing either the simple independence prior (4), the hotspot prior (5) or the MRF prior (6) for the indicator matrix $\mathbf{\Gamma}$ results in different sparsity specifications for the regressions in the dSUR model. The marginal likelihood integrating out \mathbf{B} is

no longer available for this model, so joint sampling of \mathbf{B} , $\mathbf{\Gamma}$ and C is required. However, the reparametrisation of the model (8) enables a fast computation using the MCMC algorithm.

2.3. Sparse Seemingly Unrelated Regression (SSUR)

Another approach to model the covariance matrix C is to specify a hyper-inverse Wishart prior, which means the multiple response variables have an underlying graph \mathcal{G} encoding the conditional dependence structure between responses. In this setup, a sparse graph corresponds to a sparse precision matrix C^{-1} . From a computational point of view, it is infeasible to specify a hyper-inverse Wishart prior directly on C^{-1} in high dimensions (Carvalho *et al.* 2007; Jones, Carvalho, Dobra, Hans, Carter, and West 2005; Uhler, Lenkoski, and Richards 2018; Deshpande, Ročková, and George 2019). However, Banterle *et al.* (2018) used a transformation of C to factorise the likelihood as in equation (8). The hyper-inverse Wishart distribution, i.e., $C \sim \mathcal{HIW}_{\mathcal{G}}(v, \tau \mathbb{I}_s)$, becomes in the transformed variables

$$\begin{aligned} \sigma_{qj}^2 &\sim \mathcal{IGamma}\left(\frac{v-s+t+|S_q|}{2}, \frac{\tau}{2}\right), \quad q=1, \dots, Q, t=1, \dots, |R_q| \\ \boldsymbol{\rho}_{qt} &\sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma_{rt}^2}{\tau}\right) \end{aligned} \quad (10)$$

where Q is the number of prime components in the decomposable graph \mathcal{G} , and S_q and R_q are the separators and residual components of \mathcal{G} , respectively. $|S_q|$ and $|R_q|$ denote the number of variables in these components. For more technical details, please refer to Banterle *et al.* (2018).

As prior for the graph we use an independent Bernoulli prior with probability η on each edge $E_{kk'}$ of the graph as in

$$\begin{aligned} \mathbb{P}(E_{kk'} \in \mathcal{G}) &= \eta, \\ \eta &\sim \mathcal{Beta}(a_\eta, b_\eta). \end{aligned} \quad (11)$$

The three priors on $\boldsymbol{\beta}_\gamma$, i.e., independence (4), hotspot (5) and MRF (6) priors can also be used in the SSUR model.

2.4. MCMC sampler and posterior inference

To sample from the posterior distribution, we use the Evolutionary Stochastic Search algorithm (Bottolo and Richardson 2010; Bottolo *et al.* 2011; Lewin *et al.* 2015), which uses a particular form of Evolutionary Monte Carlo (EMC) introduced by Liang and Wong (2000). Multiple tempered Markov Chains are run in parallel and both exchange and crossover moves are allowed between the chains to improve mixing between potentially different modes in the posterior. Note that we run multiple chains at the same temperature instead of a ladder of different temperatures as was proposed in the original implementations of the (H)ESS sampler in (Bottolo and Richardson 2010; Bottolo *et al.* 2011; Lewin *et al.* 2015).

The main chain samples from the un-tempered posterior distribution, which is used for all inference. For each response variable, we use a Gibbs sampler to update the regression coefficients vector, $\boldsymbol{\beta}_k$ ($k=1, \dots, s$), based on the posterior conditional corresponding to the specific model presented in Section 2.2-2.3. After L MCMC iterations, we obtain $\mathbf{B}^{(1)}, \dots, \mathbf{B}^{(L)}$ and the estimate of the posterior mean is

$$\hat{\mathbf{B}} = \frac{1}{L-b} \sum_{t=b+1}^L \mathbf{B}^{(t)},$$

where b is the number of burn-in iterations. Posterior full-conditionals are also available to update σ_k^2 ($k = 1, \dots, s$) and $\rho_{kl}|\sigma_k^2$ ($k = 1, \dots, s, l < k$) for the dSUR and SSUR models. In the HRR models in Section 2.1, the regression coefficients and residual covariances have been integrated out and therefore the MCMC output cannot be used directly for posterior inference. However, for \mathbf{B} , the posterior distribution conditional on $\mathbf{\Gamma}$ can be derived analytically for the HRR models and is provided in the **BayesSUR** package.

At MCMC iteration t we also update each binary latent vector γ_k ($k = 1, \dots, s$) via a Metropolis-Hastings sampler, jointly proposing an update for the corresponding β_k . After L iterations, using the binary matrices $\mathbf{\Gamma}^{(1)}, \dots, \mathbf{\Gamma}^{(L)}$, the marginal posterior inclusion probabilities (mPIP) of the predictors are estimated by

$$\hat{\mathbf{\Gamma}} = \frac{1}{L-b} \sum_{t=b+1}^L \mathbf{\Gamma}^{(t)}.$$

In the SSUR models, another important parameter is \mathcal{G} in the hyper-inverse Wishart prior for the covariance matrix C . It is updated by the junction tree sampler (Green and Thomas 2013; Banterle *et al.* 2018) jointly with the corresponding proposal for σ_k^2, ρ_k , ($k = 1, \dots, s$). At each MCMC iteration we then extract the adjacency matrix $\mathcal{G}^{(t)}$ ($t = 1, \dots, L$), from which we derive posterior mean estimators of the edge inclusion probabilities as

$$\hat{\mathcal{G}} = \frac{1}{L-b} \sum_{t=b+1}^L \mathcal{G}^{(t)}.$$

The hyper-parameter τ in the inverse Wishart prior or hyper-inverse Wishart prior is updated by a random walk Metropolis-Hastings sampler. The hyper-parameter η and the variance w in the spike-and-slab prior are sampled from their posterior conditional. For details see Banterle *et al.* (2018).

3. The R package BayesSUR

The package **BayesSUR** is available from GitHub (<https://github.com/mbant/BayesSUR>). It can be installed using the **devtools** package as follows.

```
R> library("devtools")
R> install_github("mbant/BayesSUR/BayesSUR")
```

The main function is `BayesSUR()` which has various arguments that can be used to specify the models introduced in Section 2, by setting the priors for the covariance matrix C and the binary latent indicator matrix $\mathbf{\Gamma}$. In addition, MCMC parameters (`nIter`, `burnin`, `nChains`) can also be defined. The following syntax example includes the most important function arguments, which are further explained below. The full list of all arguments in function `BayesSUR()` is given in Table 2.

```
R> BayesSUR(data, Y, X, covariancePrior, gammaPrior, nIter, burnin, nChains,
+           ...)
```

The data can be provided as a large combined matrix $[\mathbf{Y}, \mathbf{X}]$ of dimension $n \times (s + p)$ via the argument `data`; in that case the arguments `Y` and `X` need to contain the dimensions of the individual predictor and response matrices to allow reconstruction of \mathbf{Y} and \mathbf{X} . Alternatively, it is also possible to supply \mathbf{X} and \mathbf{Y} directly via the arguments `X` and `Y`. In that case, argument `data` needs to be `NULL` which is the default.

The arguments `covariancePrior` and `gammaPrior` specify different models introduced in Section 2. When using the Markov random field prior (6) for the latent binary vector $\boldsymbol{\gamma}$, an additional argument `mrfG` is needed to assign the edge potentials; this can either be specified as a matrix or as a file directory path leading to a text file with the corresponding information. For example, the HRR model with independent hierarchical prior in Section 2.1.1 is specified by (`covariancePrior = "IG"`, `gammaPrior = "hierarchical"`), the dSUR model with hotspot prior in Section 2.2 by (`covariancePrior = "IW"`, `gammaPrior = "hotspot"`) and the SSUR model with MRF prior in Section 2.3 for example by (`covariancePrior = "HIW"`, `gammaPrior = "MRF"`, `mrfG = "/mrfGile.txt"`).

The MCMC parameter arguments `nIter`, `burnin` and `nChains` indicate the total number of MCMC iterations, the number of iterations in the burn-in period and the number of parallel chains in the evolutionary stochastic search MCMC algorithm, respectively. See e.g., [Bottolo and Richardson \(2010\)](#) for more details on the ESS algorithm.

The main function `BayesSUR()` returns an object of S3 class `BayesSUR`, in a list format which includes the input parameters and directory paths of output text files, so that other functions can retrieve the MCMC output from the output files, load them into R and further process the output for posterior inference of the model output.

In particular, a `summary()` function has been provided for `BayesSUR` class objects, which is used to summarize the output produced by `BayesSUR()`. For this purpose, a number of covariates are selected into the model by thresholding the posterior means of the latent indicator variables. By default the threshold is 0.5, i.e., variable j is selected into the model for response k if $\hat{\gamma}_{jk} > 0.5$. The `summary()` function also outputs the quantiles of the conditional predictive ordinates (CPO, [Gelfand \(1996\)](#)), top predictors on average mPIP across all response variables and top response variables on average mPIP across all predictors, Expected log pointwise predictive density (i.e., `elpd.L00` and `elpd.WAIC`, [Vehtari et al. \(2017\)](#)), model specification parameters, MCMC running parameters and hyperparameters.

To use a specific estimator, the function `getEstimator()` is convenient to extract the coefficients matrix $\hat{\mathbf{B}}$, latent indicator variable $\hat{\mathbf{\Gamma}}$ or learned structure $\hat{\mathcal{G}}$ from the directory path of the model object. All point estimates are posterior means, thus $\hat{\gamma}_{jk}$ is the marginal posterior inclusion probability for variable j to be selected in the regression for response k , and $\hat{\mathcal{G}}_{kl}$ is the marginal posterior edge inclusion probability between responses k and l , i.e., the marginal posterior probability of conditional dependence between k and l . The regression coefficient estimates $\hat{\mathbf{B}}$ are the posterior means over all models, independently of $\hat{\mathbf{\Gamma}}$. Thus $\hat{\beta}_{jk}$ represents the shrunken estimate of the association effect of variable j in the regression for response k . In addition, the generic S3 methods `coef()`, `predict()`, and `fitted()` can be used to extract regression coefficients, predicted responses, or indices of nonzero coefficients corresponding to the posterior mean estimates of an `BayesSUR` object.

| Argument | Description |
|------------------------------|---|
| <code>data</code> | Data combined matrix $[\mathbf{Y}, \mathbf{X}]$. Default is <code>NULL</code> . |
| <code>Y</code> | Matrix or indexes with respect to the argument <code>data</code> for the responses. |
| <code>X</code> | Matrix or indexes with respect to the argument <code>data</code> for the predictors. |
| <code>X_0</code> | Matrix or indexes with respect to the argument <code>data</code> for predictors forced to be included (i.e., they are not part of variable selection procedure). Default is <code>NULL</code> . |
| <code>outFilePath</code> | Directory path where the output files are written. Default is the current working directory. |
| <code>covariancePrior</code> | Prior for the covariance matrix, "IG": independent inverse gamma prior, "IW": inverse Wishart prior, "HIW": hyper-inverse Wishart prior (default). |
| <code>gammaPrior</code> | Prior for the binary latent variable $\mathbf{\Gamma}$, "hierarchical": independent Bernoulli prior, "hotspot": hotspot prior (default), "MRF": Markov random field prior.. |
| <code>mrfG</code> | A matrix or a path to the file containing the edge list of the G matrix for the MRF prior on $\mathbf{\Gamma}$. Default is <code>NULL</code> . |
| <code>gammaSampler</code> | Local move sampler for the binary latent variable $\mathbf{\Gamma}$, either (default) "bandit" for the Thompson sampling inspired sampler or "MC3" for the usual MC^3 sampler. |
| <code>gammaInit</code> | $\mathbf{\Gamma}$ initialisation to either all-zeros ("0"), all ones ("1"), randomly ("R") or (default) MLE-informed ("MLE"). |
| <code>hyperpar</code> | A list of named hyperparameters to use instead of the default values. Default is an empty list. |
| <code>output_*</code> | Allow (<code>TRUE</code>) or suppress (<code>FALSE</code>) the output for *; possible outputs are $\mathbf{\Gamma}$, \mathcal{G} , \mathbf{B} , $\boldsymbol{\sigma}$, $\boldsymbol{\pi}$, tail (hotspot tail probability, see Bottolo and Richardson (2010)) or <code>model_size</code> . Default is all <code>TRUE</code> . |
| <code>tmpFolder</code> | The path to a temporary folder where intermediate data files are stored (will be erased at the end of the chain) default to local <code>tmpFolder</code> . |

Table 2: Overview of the arguments in the main function `BayesSUR()`

The function `plotEstimator()` visualizes the three estimators. To show the relationship of multiple response variables with each other, the function `plotEstimator()` prints the structure graph based on $\hat{\mathcal{G}}$. Furthermore, the structure relations between multiple response variables and predictors can be shown via function `plotNetwork()`. The marginal posterior probabilities of individual predictors are illustrated via the `plotManhattan()` function, which also shows the number of associated response variables of each predictor. For convenience, a generic S3 method `plot()` is also provided, which creates a selection of the above plots in an interactive manner.

Model fit can be investigated with `elpd()` and `plotCPO()`. `elpd()` estimates the expected log pointwise predictive density ([Vehtari et al. 2017](#)) to assess out-of-sample prediction accuracy. `plotCPO()` plots the conditional predictive ordinate for each individual, i.e., the leave-one-

out cross-validation predictive density. CPOs are useful for identifying potential outliers (Gelfand 1996). To check convergence of the MCMC sampler, function `plotMCMCdiag()` prints traceplots and density plots for moving windows over the MCMC chains.

Table 3 lists all functions. **BayesSUR** uses the **Rcpp** R package (Eddelbuettel and François 2011) to integrate C++ code with R, together with the **RcppArmadillo** (Eddelbuettel and Sanderson 2014) and **BH** (Eddelbuettel, Emerson, and Kane 2019) R packages. The **igraph** package (Csardi and Nepusz 2006) was used for constructing the graph plots.

| Function | Description |
|----------------------------------|---|
| <code>BayesSUR()</code> | Main function of the package. Fits any of the models introduced in Section 2. Returns an object of S3 class <code>BayesSUR</code> , which is a list which includes the input parameters (input) and directory paths of output text files (output), as well as the run status and function call. |
| <code>elpd()</code> | Measure the prediction accuracy by the expected log pointwise predictive density (elpd). The out-of-sample predictive fit can either be estimated by Bayesian leave-one-out cross-validation (LOO) or by widely applicable information criterion (WAIC) (Vehtari <i>et al.</i> 2017). See Appendix for details. |
| <code>summary()</code> | Summarize the fitted model generated by <code>BayesSUR()</code> , which is an object of class <code>BayesSUR</code> . |
| <code>print()</code> | Print a short summary the fitted model generated by <code>BayesSUR()</code> , which is an object of class <code>BayesSUR</code> . |
| <code>getEstimator()</code> | Extract the posterior mean of the parameters of a <code>BayesSUR</code> class object. |
| <code>coef()</code> | Extract the posterior mean of the coefficients of a <code>BayesSUR</code> class object. |
| <code>fitted()</code> | Return the fitted response values that correspond to the posterior mean of the coefficients matrix of a <code>BayesSUR</code> class object. |
| <code>predict()</code> | Predict responses corresponding to the posterior mean of the coefficients, return posterior mean of coefficients or indices of nonzero coefficients of a <code>BayesSUR</code> class object. |
| <code>plotCPO()</code> | Plot the conditional predictive ordinate (CPO) for each individual of a fitted model generated by <code>BayesSUR</code> which is a <code>BayesSUR</code> object. CPO is used to identify potential outliers (Gelfand 1996). |
| <code>plotEstimator()</code> | Plot the estimated relationships between response variables and estimated coefficients of a <code>BayesSUR</code> class object. |
| <code>plotResponseGraph()</code> | Plot the estimated graph for multiple response variables from a <code>BayesSUR</code> class object. |
| <code>plotNetwork()</code> | Plot the network representation of the associations between responses and predictors, based on the estimated $\hat{\mathbf{T}}$ matrix of a <code>BayesSUR</code> class object. |
| <code>plotManhattan()</code> | Plot Manhattan-like plots for marginal posterior inclusion probabilities (mPIP) and numbers of responses of association for predictors of a <code>BayesSUR</code> class object. |
| <code>plotMCMCdiag()</code> | Show trace plots and diagnostic density plots of a fitted model object of class <code>BayesSUR</code> . |
| <code>plot()</code> | Convenience function to create a selection of plots for a <code>BayesSUR</code> class object; individual plots are generated with the specific plot functions above. |

Table 3: Overview of the functions in package `BayesSUR`.

4. Quick start with a simple example

In the following example, we illustrate a simple simulation study where we run two models: the default model choice, which is an SSUR model with the hotspot prior, and in addition an SSUR model with the MRF prior. The purpose of the latter is to illustrate how we can construct an MRF prior graph. We simulate a dataset \mathbf{X} with dimensions $n \times p = 10 \times 15$, i.e. 10 observations and 15 input variables, a sparse coefficients matrix \mathbf{B} with dimension $p \times s = 15 \times 3$, which creates associations between the input variables and $s = 3$ response variables, and random noise \mathbf{E} . The response matrix is generated by the linear model $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$.

```
R> set.seed(6349)
R> n <- 10; s <- 3; p <- 15
R> X <- matrix(rnorm(n * p, 2, 1), nrow = n)
R> B <- matrix(c(0, 0, 0.5,
+              0.5, 0.5, 0,
+              0.5, 0.5, 0,
+              rep(0, s * p - 9)),
+             nrow = p, byrow = TRUE)
R> E <- matrix(rnorm(n * s, 0, 0.2), nrow = n)
R> Y <- X %>% B + E
```

Note that \mathbf{B} is sparse and only the first three input variables have non-zero coefficients:

```
R> print(B)

      [,1] [,2] [,3]
[1,]  0.0  0.0  0.5
[2,]  0.5  0.5  0.0
[3,]  0.5  0.5  0.0
[4,]  0.0  0.0  0.0
[5,]  0.0  0.0  0.0
[6,]  0.0  0.0  0.0
[7,]  0.0  0.0  0.0
[8,]  0.0  0.0  0.0
[9,]  0.0  0.0  0.0
[10,] 0.0  0.0  0.0
[11,] 0.0  0.0  0.0
[12,] 0.0  0.0  0.0
[13,] 0.0  0.0  0.0
[14,] 0.0  0.0  0.0
[15,] 0.0  0.0  0.0
```

First, let's fit the default model, which is the SSUR model with a hotspot prior for $\mathbf{\Gamma}$. The default is to run two MCMC chains with 10000 iterations each, of which the first 5000 iterations are discarded as the burn-in period. The function `print()` returns a short summary of the results from the fitted model object, including the number of selected predictors by thresholding the marginal posterior inclusion probabilities (mPIP) at 0.5, and two measures of the model's prediction accuracy (i.e., `elpd.L00` and `elpd.WAIC`).

```
R> library("BayesSUR")
R> fit <- BayesSUR(Y = Y, X = X, outFilePath = "results/")
```

```
R> print(fit)
```

Call:

```
BayesSUR(Y = Y, X = X, outFilePath = "results/")
```

Number of selected predictors (mPIP > 0.5): 5 of 3x15

Expected log pointwise predictive density (elpd):

```
elpd.L00 = -28.85931, elpd.WAIC = -29.15983
```

The posterior means of the coefficients matrix and latent indicator variable are printed by the function `plotEstimator()` (Figure 1).

```
R> plotEstimator(fit, estimator = c("beta", "gamma"),
+               fig.tex = TRUE, output = "exampleEst",
+               xlab = "Predictors", ylab = "Responses")
```

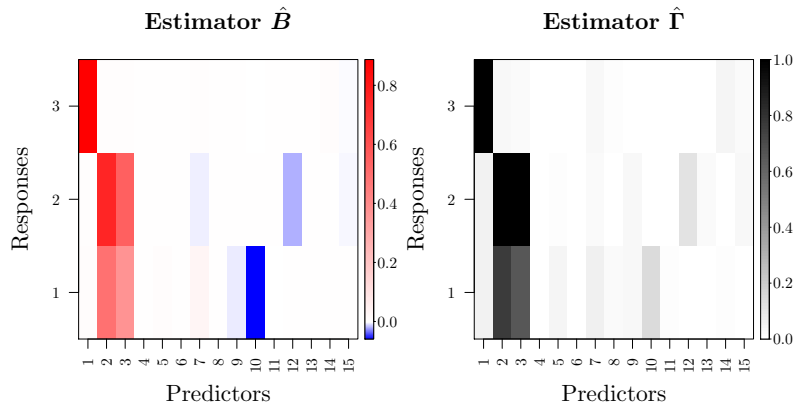


Figure 1: The posterior mean estimates of the coefficients matrix \hat{B} and latent indicator variable $\hat{\Gamma}$ for the SSUR model with the hotspot prior plotted with `plotEstimator()`.

Before running the SSUR model with the MRF prior, we need to construct the edge potentials matrix G . If we assume (in accordance with the true matrix B in this simulation scenario) that the 2nd and 3rd predictors are related to the first two response variables, this implies that γ_{21} , γ_{22} , γ_{31} and γ_{32} are expected to be related and therefore we might want to encourage these variables to be selected together. In addition, we assume that we know that predictor 1 is associated with response 3, and therefore we encourage the selection of γ_{13} as well. Since matrix G represents prior relation of any two predictors corresponding the $\text{vec}\{\Gamma\}$, it can be generated by the following code:

```
R> G <- matrix(0, ncol = s * p, nrow = s * p)
R> combn1 <- combn(rep((1:2 - 1) * p, each = length(2:3))
+               + rep(2:3, times = length(1:2)), 2)
R> combn2 <- rep(rep((3-1) * p, each = length(1))
+               + rep(1, times = length(3)), 2)
R> G[c(combn1[1,], combn2[1]), c(combn1[2,], combn2[2])] <- 1
```

Calling `BayesSUR()` with the argument `gammaPrior = "MRF"` will run the SSUR model with the MRF prior, and the argument `mrfG = G` imports the edge potentials for the MRF prior. The two hyper-parameters d and e for the MRF prior (6) can be specified through the argument `hyperpar`; here we use the default values $d = -3$, $e = 0.001$. See [Lee et al. \(2017\)](#) for an example for how one can select d and e via a grid search strategy which exploits the phase transition of the MRF parametrization. The posterior mean estimates for the coefficients matrix and latent indicator variable are shown in Figure 2.

```
R> fit <- BayesSUR(Y = Y, X = X, outFilePath = "results/",
+               gammaPrior = "MRF", mrfG = G)
R> plotEstimator(fit, estimator = c("beta", "gamma"),
+               fig.tex = TRUE, output = "exampleEst2",
+               xlab = "Predictors", ylab = "Responses")
```

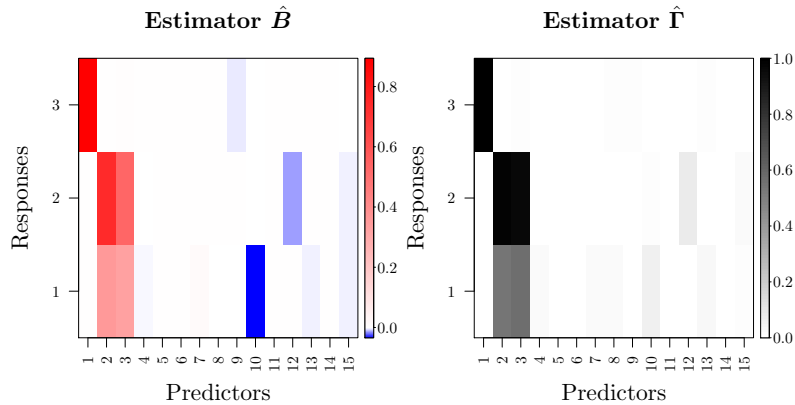


Figure 2: The posterior mean estimates of the coefficients matrix \hat{B} and latent indicator variable $\hat{\Gamma}$ for the SSUR model with MRF prior plotted with `plotEstimator()`.

5. Two extended examples based on real data

In this section, we use a simulated eQTL dataset and a real pharmacogenetic database to illustrate the usage of the **BayesSUR** package. The first example is under the known true model and demonstrates the recovery performance of the models introduced in Section 2. It also demonstrates an example data analysis step by step. The second example illustrates how to use potential relationships between multiple response variables and input predictors as the

prior information in Bayesian SUR models and showcases how the resulting estimated graph structures can be visualized with functions provided in the package.

5.1. Simulated eQTL data

Similarly to Banterle *et al.* (2018), we simulate single nucleotide polymorphism (SNP) data \mathbf{X} by resampling from the **scrim** package (Schwender and Fritsch 2012), with $p = 150$ SNPs and $n = 100$ subjects. To construct multiple response variables \mathbf{Y} (with $s = 10$) with structured correlation - which we imagine to represent gene expression measurements of genes that are potentially affected by the SNPs - we first fix a sparse latent indicator variable $\mathbf{\Gamma}$ and then design a decomposable graph for responses to build association patterns between multi-response variables and predictors. The nonzero coefficients are sampled from the normal distribution independently and the noise term from a multivariate normal distribution with the covariance matrix sampled from a Wishart distribution. Finally, the simulated gene expression data \mathbf{Y} is then generated from the linear model (1). The concrete steps are as follows:

- Simulate SNPs data \mathbf{X} from the **scrim** package, $\dim(\mathbf{X}) = n \times p$.
- Design a decomposable graph \mathcal{G} as the right panel of Figure 3, $\dim(\mathcal{G}) = s \times s$.
- Design a sparse matrix $\mathbf{\Gamma}$ as the left panel of Figure 3, $\dim(\mathbf{\Gamma}) = p \times s$.
- Simulate $\beta_{jk} \sim \mathcal{N}(5, 1)$, $j = 1, \dots, p$ and $k = 1, \dots, s$.
- Simulate $\tilde{u}_{ij} \sim \mathcal{N}(1, 0.5)$, $i = 1, \dots, n$ and $j = 1, \dots, p$.
- Simulate $P \sim \mathcal{W}_G(2, M)$ (G-Wishart distribution, Mohammadi and Wit (2019)) where diagonals of M are 1 and off-diagonals are 0.9, $\dim(P) = s \times s$.
- Use Cholesky decomposition $\text{chol}(P^{-1})$ to get $\mathbf{U} = \tilde{\mathbf{U}} \cdot \text{chol}(P^{-1})$.
- Generate $\mathbf{Y} = (\mathbf{X}\mathbf{B})_{\mathbf{\Gamma}} + \mathbf{U}$.

The signal-to-noise ratio SNR_{β} within 10% of the desired value 35, where

$$\text{SNR}_{\beta} = \frac{1}{(n-1)s} \sum_{k=1}^s \sum_{i=1}^n \frac{(X_{i,\gamma_k} \beta_{\gamma_k})^{\top} (X_{i,\gamma_k} \beta_{\gamma_k})}{\sigma_k^2}.$$

The R code for the simulation can be found through `help("example_eQTL")`.

```
R> data("example_eQTL", package = "BayesSUR")
R> str(example_eQTL)
```

```
List of 2
```

```
$ data      : num [1:100, 1:160] 40.1 63 65.4 50.3 42.2 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : chr [1:100] "1" "2" "3" "4" ...
.. ..$ : chr [1:160] "GEX1" "GEX2" "GEX3" "GEX4" ...
$ blockList:List of 2
..$ : int [1:10] 1 2 3 4 5 6 7 8 9 10
..$ : num [1:150] 11 12 13 14 15 16 17 18 19 20 ...
```

```
R> attach(example_eQTL)
```

In the **BayesSUR** package, the data \mathbf{Y} and \mathbf{X} are saved as one matrix in the first component `data` of the example dataset `example_eQTL`. The second component of `example_eQTL` is `blockList` which specifies the indices of \mathbf{Y} and \mathbf{X} in `data`. Throughout this section we attach the data set for more concise R code.

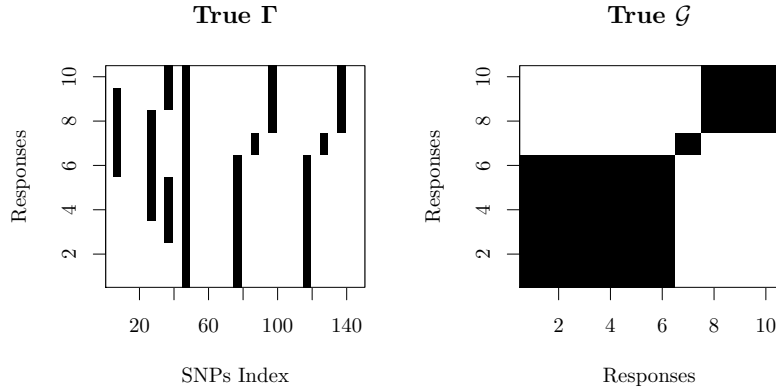


Figure 3: True parameters of the simulated dataset `example_eQTL`. The left panel is the designed sparse $\mathbf{\Gamma}$ and the right panel is the given true structure of responses represented by the decomposable graph \mathcal{G} . Black indicates a value 1 and white indicates 0.

Figure 3 shows the true $\mathbf{\Gamma}$ and decomposable graph \mathcal{G} used in the eQTL simulation scenario. The following code shows how to fit an SSUR model with hotspot prior for the indicator variables $\mathbf{\Gamma}$ and the sparsity-inducing hyper-inverse Wishart prior for the covariance using the main function `BayesSUR()`.

```
R> fit <- BayesSUR(data = data, Y = blockList[[1]], X = blockList[[2]],
+                 outFilePath = "results/", nIter = 210000, nChains = 4,
+                 burnin = 10000, covariancePrior = "HIW",
+                 gammaPrior = "hotspot")
```

Figure 4 summarizes the posterior inference results by plots for $\hat{\mathbf{B}}$, $\hat{\mathbf{\Gamma}}$ and $\hat{\mathcal{G}}$ created with the function `plotEstimator()`. When comparing with Figure 3, we see that this SSUR model has good recovery of the true latent indicator matrix $\mathbf{\Gamma}$ and of the structure of the responses as represented by \mathcal{G} . The function `plotResponseGraph()` visualizes the estimated structure of the ten gene expression variables as shown in the right panel of Figure 5. For comparison, the true structure of them is shown in the left panel. When we threshold the posterior selection probability estimates for \mathcal{G} and for $\mathbf{\Gamma}$ at 0.5, the resulting full network between the ten gene expression variables and 150 SNPs is displayed in Figure 6. Furthermore, the Manhattan-like plots in Figure 7 show both, the marginal posterior inclusion probabilities (mPIP) of the SNP variables (top panel) and the number of gene expression response variables associated with each SNP (bottom panel).

```
R> plotEstimator(fit, fig.tex = TRUE)
```

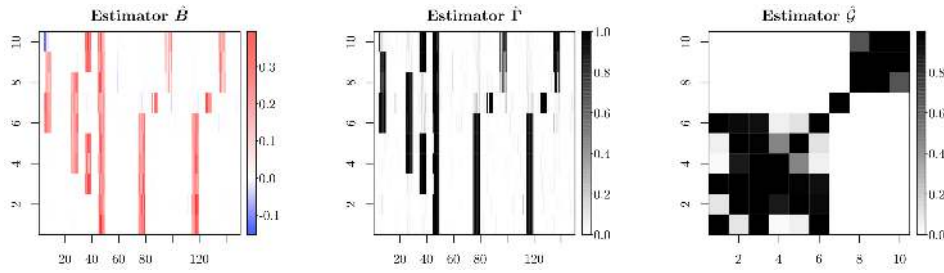



Figure 4: The estimated coefficients matrix \hat{B} , latent indicator variable $\hat{\Gamma}$ and learning structure \hat{G} by the SSUR model with hotspot prior and sparse covariance prior by `plotEstimator()`.

```
R> plotResponseGraph(fit)
```

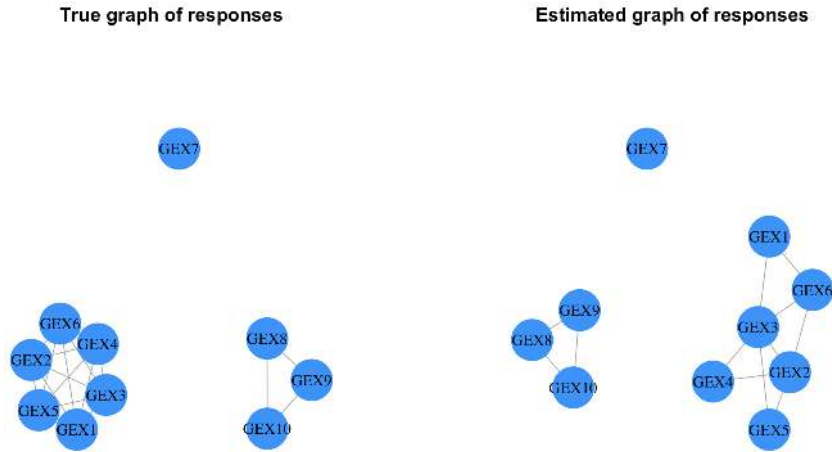


Figure 5: Structure of the ten response variables visualized by `plotResponseGraph()`. Their associations are based on \hat{G} thresholded at 0.5.

```
R> plotNetwork(fit)
```

```
R> plotManhattan(fit)
```

In order to investigate the behaviour of the MCMC sampler, the top two panels of Figure 8 show the trace plots of the loglikelihood and model size, i.e., the total number of selected predictors. We observe that the Markov chain seems to start sampling from the correct distribution after ca. 50,000 iterations. The bottom panels of Figure 8 indicate that the log posterior distribution of the latent indicator variable Γ is stable for the last half of the chains after subtracting the burn-in length.

```
R> plotMCMCdiag(fit)
```

We finish this example analysis by detaching the eQTL example data set.

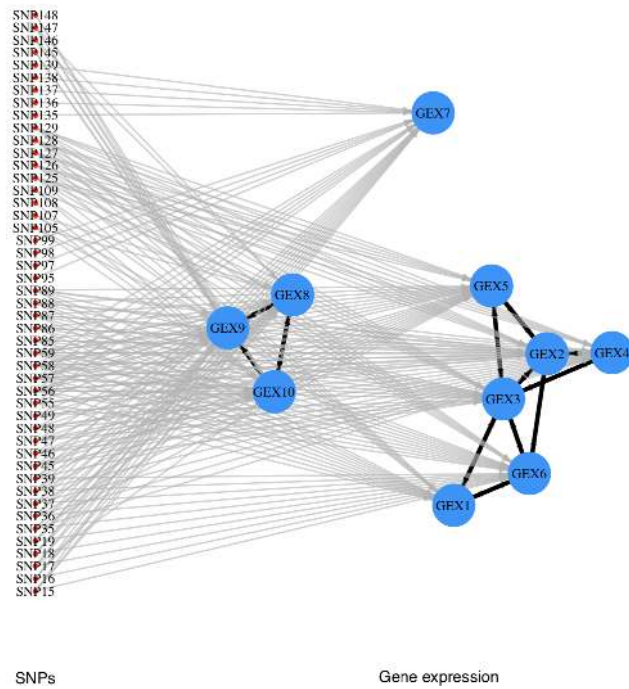


Figure 6: Network representation between the ten expression level genes and 150 SNPs by `plotNetwork()`. The connections between expression level genes are based on \hat{G} thresholded at 0.5, and the connections between the expression level genes and SNPs are based on \hat{I} thresholded at 0.5.

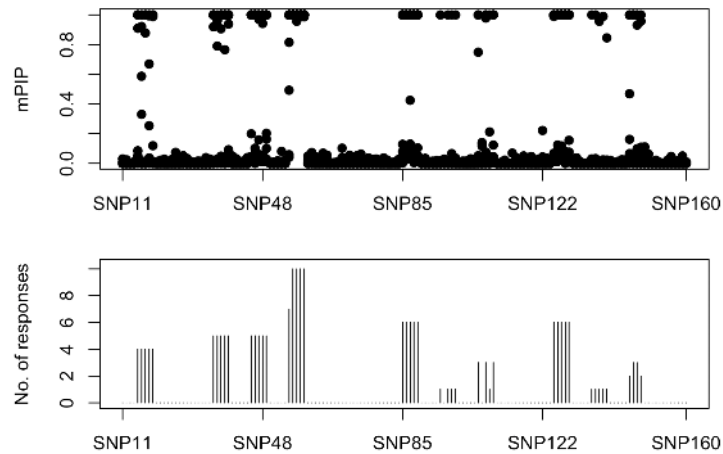
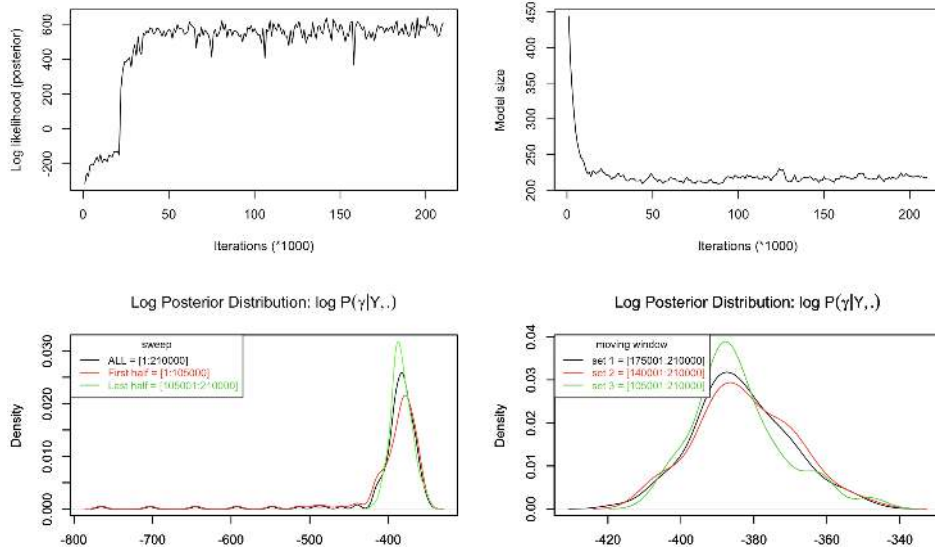


Figure 7: Manhattan-like plots by `plotManhattan()`. The top panel shows the mPIP of each SNP, and the bottom panel shows the number of genes associated with each SNP. The number of responses are based on \hat{I} thresholded at 0.5.

```
R> detach(example_eQTL)
```

Figure 8: Diagnostic plots of the MCMC sampler by `plotMCMCdiag()`.

5.2. The Genomics of Drug Sensitivity in Cancer data

In this section we analyse a subset of the Genomics of Drug Sensitivity in Cancer (GDSC) dataset from a large-scale pharmacogenomic study (Yang, Soares, Greninger, Edelman, Lightfoot, Forbes, Bindal, Beare, Smith, Thompson, Ramaswamy, Futreal, Haber, Stratton, Benes, McDermott, and Garnett 2013; Garnett *et al.* 2012). We analyze $s = 7$ cancer drugs' pharmacological profiling of $n = 498$ cell lines from $p_0 = 13$ cancer tissues. The sensitivity of the cell lines to each of the drugs was summarised by the $\log\text{IC}_{50}$ values estimated from *in vitro* dose response experiments. The cell lines are characterised by $p_1 = 343$ selected gene expression features (GEX), $p_2 = 426$ genes affected by copy number variations (CNV) and $p_3 = 68$ mutated genes (MUT). The data sets were downloaded from <ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/release-5.0/> and processed as described in `help("example_GDSC")`.

Garnett *et al.* (2012) provide the target genes or pathways for all drugs. The aim of this study was to identify molecular characteristics that help predict the response of a cell line to a particular drug. Because many of the drugs share common targets and mechanism of action, the response of cell lines to many of the drugs is expected to be correlated. Therefore, a multivariate model seems appropriate:

$$\mathbf{Y}_{\text{drugs}} = \mathbf{X}_{\text{tissues}}\mathbf{B}_0 + \mathbf{X}_{\text{GEX}}\mathbf{B}_1 + \mathbf{X}_{\text{CNV}}\mathbf{B}_2 + \mathbf{X}_{\text{MUT}}\mathbf{B}_3 + \mathbf{U}_{\text{error}}.$$

We may know the biological relationships within and between drugs and molecular features, so that the MRF prior (6) can be used to learn the above multivariate model well. Drugs RDEA119, PD-0325901, CI-1040 and AZD6244 are MEK inhibitors which affect the MAP-K/ERK pathway. Drugs Nilotinib and Axitinib are Bcr-Abl tyrosine kinase inhibitors which inhibit the mutated BCR-ABL gene. Drug Methotrexate is a chemotherapy agent and general immune system suppressant, which is not associated with a particular target gene or pathway. For the target genes (and genes in target pathways) we consider all characteristics (GEX, CNV, MUT) available in our data set as being potentially associated. Based on this

information, we construct edge potentials for the MRF prior:

- edges between drugs: Group1 ("RDEA119","PD-0325901","CI-1040" and "AZD6244"); Group2 ("Nilotinib","Axitinib")
- edges between genes in MAPK/ERK pathway (target of Group1 drugs)
- edges between genes in the Bcr-Abl fusion gene (target of Group2 drugs)
- edges between genes of MAPK/ERK pathway and Group1 drugs
- edges between genes of the Bcr-Abl fusion gene and Group2 drugs
- edges between the representations of each gene in different data sources (i.e., GEX, CNV and MUT)

By matching the selected genes with the gene set of the MAPK/ERK pathway from the KEGG database, 57 features are considered to be connected to the four MEK inhibitors. The two genes (i.e., BCR and ABL) representing the Bcr-Abl fusion are connected with five features in the data set, which are BCR-ABL mutation, BCR gene expression, BCR copy number variation, ABL gene expression and ABL copy number variation (left panel of Figure 9). In addition, there are 347 small feature groups representing different available data sources for each of the genes in the data set, which are potentially connected to all drugs. Figure 9 illustrates the edges between Drugs Nilotinib, Axitinib and the related genes of the Bcr-Abl fusion gene, and the same gene from different data sources related to each drug. Based on this information, we construct the matrix G for the MRF prior.

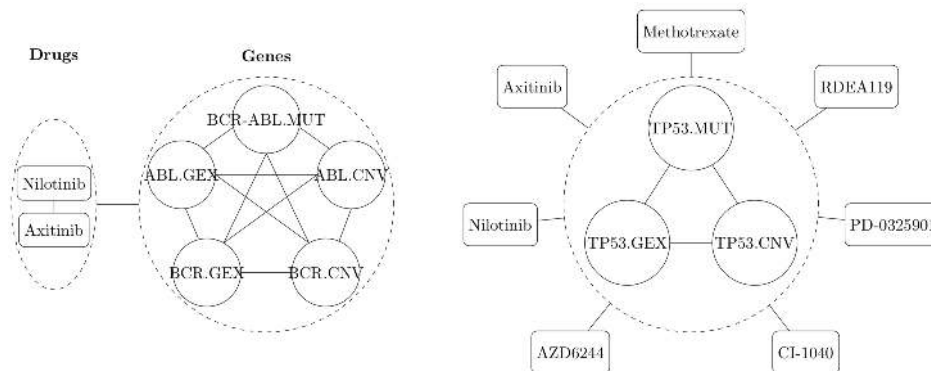


Figure 9: Illustration of the relationship between drugs and a group of related genes. The left panel is for the Bcr-Abl fusion gene and the corresponding related genes. The right panel is for all drugs and gene TP35 as one example with features representing all three data sources. The names with suffix ".GEX", ".CNV" and ".MUT" are features of expression, copy number variation and mutation, respectively.

First, we load and attach the data.

```
R> data("example_GDSC", package = "BayesSUR")
R> attach(example_GDSC)
```

The following code chunk will run the MCMC sampler to fit the model. This represents a full analysis, which might take several hours to run with the chosen MCMC parameter values (`nIter=200000`, `nChains=10`, `burnin=50000`). Approximate results for an initial assessment of the model can be achieved with much shorter MCMC runs.

```
R> hyperpar <- list(mrf_d = -3, mrf_e = 0.2)
R> fit <- BayesSUR(data = data, Y = blockList[[1]], X_0 = blockList[[2]],
+               X = blockList[[3]], outFilePath = "results/",
+               nIter = 200000, nChains = 10, burnin = 50000,
+               covariancePrior = "HIW", gammaPrior = "MRF",
+               mrfG = example_GDSC$mrfG)
```

After fitting an SSUR model with the MRF prior, the structure of the seven drugs, \mathcal{G} , has been learned as illustrated in Figure 10. The learned structure accurately reflects the true drug groups of the MEK inhibitors (group 1) and Bcr-Abl tyrosine kinase inhibitors (group 2) as well as the singleton drug Methotrexate, when thresholding $\hat{\mathcal{G}}$ at 0.5.

```
R> plotEstimator(fit, "Gy", name.responses = c("Methotrexate", "RDEA119",
+               "PD.0325901", "CI.1040", "AZD6244", "Nilotinib",
+               "Axitinib"), fig.tex = TRUE, output = "ResponseGraphGDSC1")
```

```
R> plotResponseGraph(fit)
```

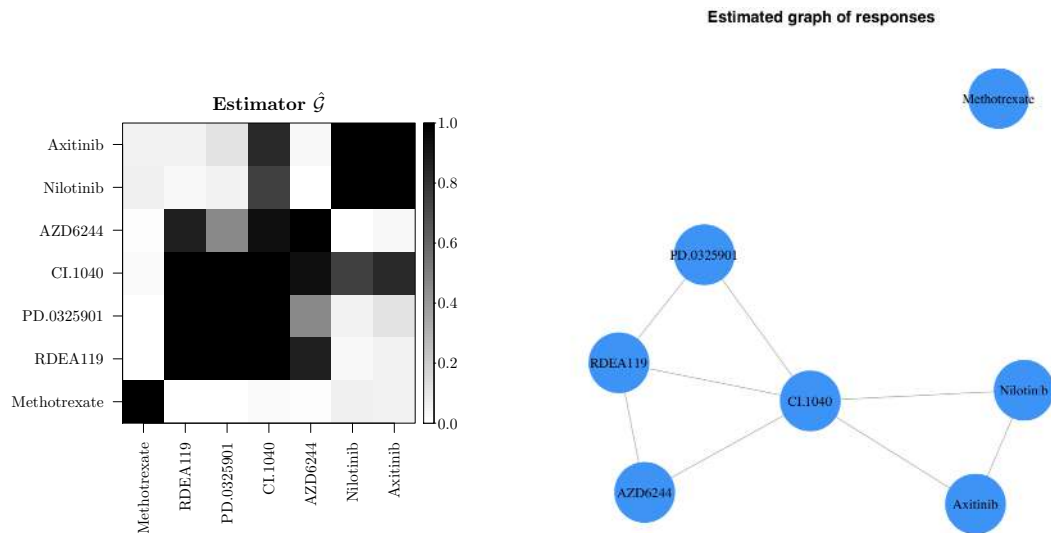


Figure 10: Estimated structure of the seven drugs $\hat{\mathcal{G}}$. Their associations as visualized in the right panel are based on $\hat{\mathcal{G}}$ thresholded at 0.5. Figures created with `plotEstimator()` (left) and `plotResponseGraph()` (right).

The estimated relationships between the drugs and genes are displayed in Figure 11. There are 812 of 5859 molecular features selected in total when thresholding $\hat{\mathbf{\Gamma}}$ at 0.5, 116 for drug Methotrexate, 463 for the MEK inhibitors and 233 for the Bcr-Abl tyrosine kinase inhibitors.

```
R> plotNetwork(fit, label.predictor = "", name.predictors = "Genes",
+             name.responses = "Drugs", nodesizePredictor = 2)
```

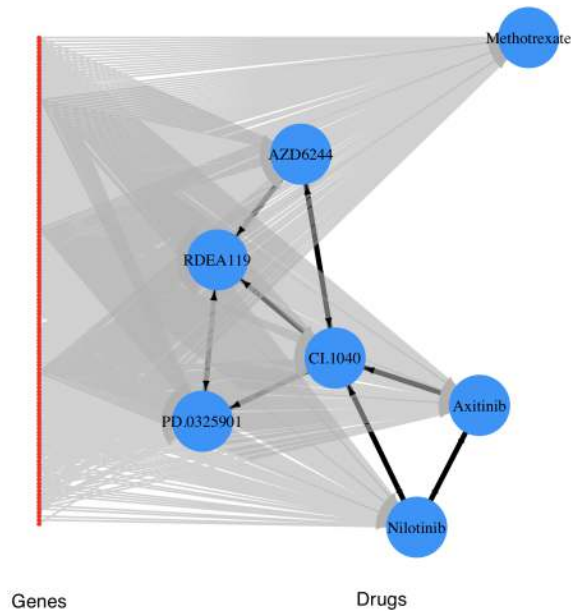


Figure 11: Estimated network between the seven drugs and selected genes based on thresholds 0.5. Figure created with `plotNetwork()`.

Network substructures of interest can also be selected and visualized individually, since the user can specify, which response variables (drugs) and which input variables (molecular features) to include in a figure. For example, Figures 12 and Figure 13 show the estimated network representations of the two groups of drugs, respectively.

```
R> data("targetGene", package = "BayesSUR")
R> plotNetwork(fit, includeResponse = c("RDEA119", "PD.0325901", "CI.1040",
+   "AZD6244"), includePredictor = names(targetGene[[1]]))
```

In addition, Figure 13 illustrates, how one can customize how to display the edges between input and response variables to visualize the strength of the association between nodes. In particular, one can either simply use a threshold, e.g., 0.5, to show all edges with marginal posterior inclusion probabilities larger than the threshold equally (left panel), or the width of edges (greater than the specified threshold) can be weighted by the corresponding inclusion probability (right panel).

```
R> layout(matrix(1:2, ncol = 2))
R> plotNetwork(fit, includeResponse = c("Nilotinib", "Axitinib"),
+             includePredictor = names(targetGene[[2]]), edge.weight = T)
R> plotNetwork(fit, PmaxPredictor = 0.05,
+             includeResponse = c("Nilotinib", "Axitinib"),
+             includePredictor = names(targetGene[[2]]), edge.weight = T)
```

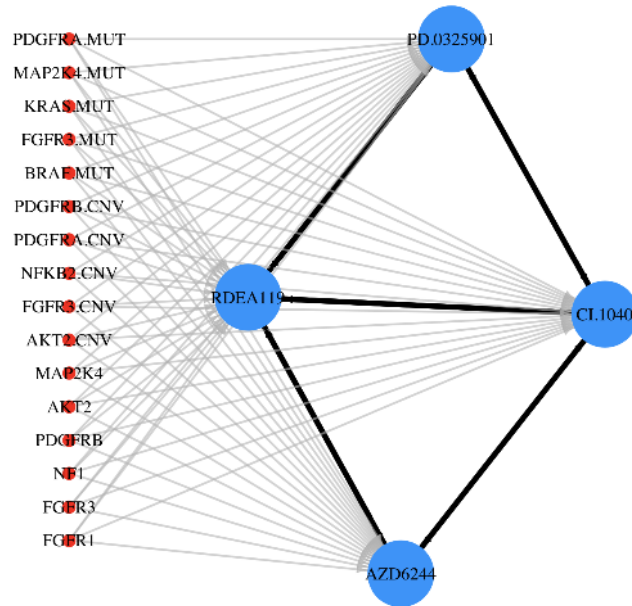


Figure 12: Estimated network between the MEK inhibitors and selected target genes based on thresholds 0.5. Figure created with `plotNetwork()`.

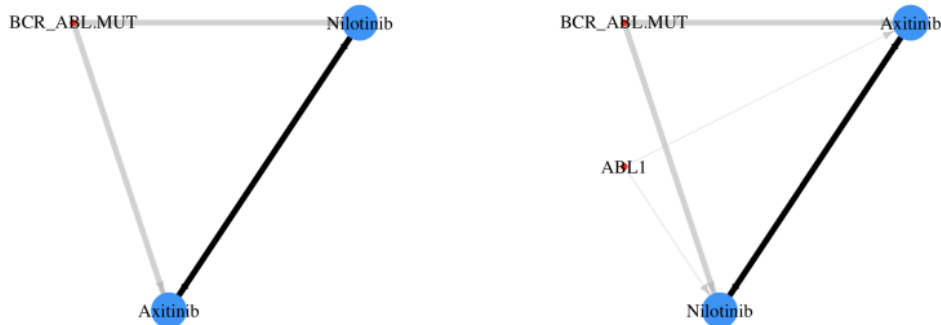


Figure 13: Estimated network between the Bcr-Abl inhibitors and selected target genes. The left plot is based on threshold on $\hat{\Gamma}$ of 0.5 while the right plot is based on threshold 0.05. Both panels use a threshold on $\hat{\mathcal{G}}$ of 0.5. The edges are weighted by the corresponding inclusion probabilities, if they are greater than the specified thresholds. Figures created with `plotNetwork()`.

6. Conclusion

The **BayesSUR** package presents a series of multivariate Bayesian variable selection models which employ the ESS algorithm for posterior inference over the model space. It provides a unified R package and a consistent interface for the C++ implementations of individual models. The package supports all combinations of the covariance priors and variable selection priors from Section 2 in the Bayesian HRR and SUR model frameworks. This includes the MRF prior on the latent indicator variables, which is newly introduced in the context of SUR models, to allow the user to make use of prior knowledge of the relationships between both response variables and predictors. To overcome the computational cost for datasets

with large numbers of input variables, parallel processing is also considered with respect to multiple chains, likelihoods of parameters and samples, although the MCMC algorithm is still challenging to be parallelized. We demonstrated the modelling aspects of variable selection and structure recovery to identify relationships between multivariate responses and between (potentially high-dimensional) responses and high-dimensional predictors, by applying the package to a simulated eQTL dataset and to the GDSC pharmacogenomic data.

Possible extensions of the R package include the implementation of different priors to introduce even more flexibility in the modeling choices. In particular, the g -prior could be considered for the regression coefficients matrix \mathbf{B} (Bottolo and Richardson 2010; Richardson *et al.* 2011; Lewin *et al.* 2015), whereas currently only the independence prior is available. In addition, the spike-and-slab prior on the covariance matrix \mathbf{C} (Wang 2015; Banerjee and Ghosal 2015; Deshpande *et al.* 2019) might be useful, or the horseshoe prior on the latent indicator variable $\mathbf{\Gamma}$, which was recently implemented in the multivariate regression setup by Ruffieux, Davison, Hager, Inshaw, Fairfax, Richardson, and Bottolo (2018).

7. Acknowledgement

*M. Banterle and Z. Zhao contributed equally to the paper. #A. Lewin and M. Zucknick are joint last authors.

References

- Banerjee S, Ghosal S (2015). “Bayesian structure learning in graphical models.” *Journal of Multivariate Analysis*, **136**, 147–162.
- Banterle M, Bottolo L, Richardson S, Ala-Korpela M, Jarvelin MR, Lewin A (2018). “Sparse variable and covariance selection for high-dimensional seemingly unrelated Bayesian regression.” *bioRxiv*, p. 467019. doi:10.1101/467019.
- Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin A, Kim S, Wilson C, Lehár J, Kryukov G, Sonkin D, *others* (2012). “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.” *Nature*, **483**(7391), 603–607. doi:10.1038/nature11003.
- Bhadra A, Mallick BK (2013). “Joint High-Dimensional Bayesian Variable and Covariance Selection with an Application to eQTL Analysis.” *Biometrics*, **69**(2), 447–457. doi:10.1111/biom.12021.
- Bottolo L, Petretto E, Blankenberg S, Cambien F, Cook SA, Tiret L, Richardson S (2011). “Bayesian detection of expression quantitative trait loci hot-spots.” *Genetics*, **189**(4), 1449–1459. doi:10.1534/genetics.111.131425.
- Bottolo L, Richardson S (2010). “Evolutionary Stochastic Search for Bayesian Model Exploration.” *Bayesian Analysis*, **5**(3), 583–618. doi:10.1214/10-BA523.
- Carvalho CM, Massam H, West M (2007). “Simulation of Hyper-Inverse Wishart Distributions in Graphical Models.” *Biometrika*, **94**(3), 647–659. doi:10.1093/biomet/asm056.

- Csardi G, Nepusz T (2006). “The igraph software package for complex network research.” *InterJournal, Complex Systems*, 1695. URL <http://igraph.org>.
- Deshpande SK, Ročková V, George EI (2019). “Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso.” *Journal of Computational and Graphical Statistics*, pp. 1–11.
- Eddelbuettel D, Emerson JW, Kane MJ (2019). *BH: Boost C++ Header Files*. R package version 1.69.0-1, URL <https://CRAN.R-project.org/package=BH>.
- Eddelbuettel D, François R (2011). “Rcpp: Seamless R and C++ Integration.” *Journal of Statistical Software*, **40**(8), 1–18. doi:10.18637/jss.v040.i08. URL <http://www.jstatsoft.org/v40/i08/>.
- Eddelbuettel D, Sanderson C (2014). “RcppArmadillo: Accelerating R with high-performance C++ linear algebra.” *Computational Statistics and Data Analysis*, **71**, 1054–1063. URL <http://dx.doi.org/10.1016/j.csda.2013.02.005>.
- Garnett M, Edelman E, Heidorn S, Greenman C, Dastur A, Lau K, Greninger P, Thompson I, Luo X, Soares J, Liu Q, others (2012). “Systematic identification of genomic markers of drug sensitivity in cancer cells.” *Nature*, **483**(7391), 570–575. doi:10.1038/nature11005.
- Gelfand A (1996). “Model Determination Using Sampling Based Method.” In W Gilks, S Richardson, D Spiegelhalter (eds.), *Markov Chain Monte Carlo in Practice*, chapter 9, pp. 145–161. Chapman Hall, Boca Raton, FL.
- Gray JW, Mills GB (2015). “Large-Scale Drug Screens Support Precision Medicine.” *Cancer Discovery*, **5**(11), 1130–1132. doi:10.1158/2159-8290.CD-15-0945.
- Green PJ, Thomas A (2013). “Sampling decomposable graphs using a Markov chain on junction trees.” *Biometrika*, **100**(1), 91–110.
- Holmes C, Denison D, Mallick B (2002). “Accounting for Model Uncertainty in Seemingly Unrelated Regressionsa.” *Journal of Computational and Graphical Statistics*, **11**(3), 533–551. doi:10.1198/106186002475.
- Jia Z, Xu S (2007). “Mapping Quantitative Trait Loci for Expression Abundance.” *Genetics*, **176**(1), 611–623. doi:10.1534/genetics.106.065599.
- Jones B, Carvalho C, Dobra A, Hans C, Carter C, West M (2005). “Experiments in Stochastic Computation for High-Dimensional Graphical Models.” *Statistical Science*, **20**(4), 388–400.
- Lee KH, Tadesse MG, Baccarelli AA, Schwartz J, Coull BA (2017). “Multivariate Bayesian variable selection exploiting dependence structure among outcomes: Application to air pollution effects on DNA methylation.” *Biometrics*, **73**(1), 232–241. doi:10.1111/biom.12557.
- Lewin A, Saadi H, Peters JE, Moreno-Moral A, Lee JC, Smith KG, Petretto E, Bottolo L, Richardson S (2015). “MT-HESS: an efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues.” *Bioinformatics*, **32**(4), 523–532.

- Liang F, Wong WH (2000). “Evolutionary Monte Carlo: Applications to C_p model sampling and change point problem.” *Statistica Sinica*, **10**, 317–342.
- Liquet B, Bottolo L, Campanella G, Richardson S, Chadeau-Hyam M (2016). “R2GUESS: A Graphics Processing Unit-Based R Package for Bayesian Variable Selection Regression of Multivariate Responses.” *Journal of Statistical Software, Articles*, **69**(2), 1–32. ISSN 1548-7660. doi:10.18637/jss.v069.i02.
- Liquet B, Mengersen K, Pettitt AN, Sutton M (2017). “Bayesian Variable Selection Regression of Multivariate Responses for Group Data.” *Bayesian Analysis*, **12**(4), 1039–1067. doi:10.1214/17-BA1081.
- Mohammadi R, Wit E (2019). “BDgraph: An R package for Bayesian structure learning in graphical models.” *Journal of Statistical Software*, **89**(3), 1–30. ISSN 1548-7660. doi:10.18637/jss.v089.i03.
- Petretto E, Bottolo L, Langley SR, Heinig M, Mcdermott-Roe C, Sarwar R, Pravenec M, Hubner N, Aitman TJ, Cook SA, Richardson S (2010). “New Insights into the Genetic Control of Gene Expression using a Bayesian Multi-tissue Approach.” *PLoS Computational Biology*, **6**(4), e1000737. doi:10.1371/journal.pcbi.1000737.
- Richardson S, Bottolo L, Rosenthal JS (2011). “Bayesian Models for Sparse Regression Analysis of High Dimensional Data.” *Bayesian Statistics 9*, pp. 539–568. doi:10.1093/acprof:oso/9780199694587.001.0001.
- Ruffieux H, Davison AC, Hager J, Inshaw J, Fairfax BP, Richardson S, Bottolo L (2018). “A global-local approach for detecting hotspots in multiple-response regression.” *arXiv preprint arXiv:1811.03334*.
- Schwender H, Fritsch A (2012). “scrim: Analysis of high-dimensional categorical data such as SNP data.” *R package version 1.3.5*, pp. Available from <https://cran.r-project.org/web/packages/scrim/>.
- Uhler C, Lenkoski A, Richards D (2018). “Exact Formulas for the Normalizing Constants of Wishart Distributions for Graphical Models.” *The Annals of Statistics*, **46**(1), 90–118.
- Vehtari A, Gelman A, Gabry J (2017). “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.” *Statistics and Computing*, **27**(5), 1413–1432. doi:10.1007/s11222-016-9696-4.
- Wang H (2010). “Sparse Seemingly Unrelated Regression Modelling: Applications in Finance and Econometrics.” *Comput. Stat. Data Anal.*, **54**(11), 2866–2877. doi:10.1016/j.csda.2010.03.028.
- Wang H (2015). “Scaling it up: Stochastic search structure learning in graphical models.” *Bayesian Analysis*, **10**(2), 351–377.
- Yang W, Soares J, Greninger P, Edelman E, Lightfoot H, Forbes S, Bindal N, Beare D, Smith J, Thompson I, Ramaswamy S, Futreal P, Haber D, Stratton M, Benes C, McDermott U, Garnett M (2013). “Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.” *Nucleic Acids Research*, **41**(Database issue), D955–61. doi:10.1093/nar/gks1111.

Appendix for the elpd

W.L.O.G., here we only consider each response variable \mathbf{y} of the whole response matrix \mathbf{Y} . Then the basic linear model is

$$\begin{aligned}\mathbf{y}|X, \beta, \sigma^2 &\sim \mathcal{N}(X\beta, \sigma^2\mathbb{I}_m), \\ \beta|\sigma^2 &\sim \mathcal{N}(\mu_\beta, \sigma^2V_\beta), \\ \sigma^2 &\sim \mathcal{IG}(a, b).\end{aligned}\tag{A.1}$$

In Bottolo *et al.* (2011) and the HRR model of this article, $\mu_\beta = \mathbf{0}$ and $V_\beta = \mathbb{I}_p$ for nonzero coefficients.

Appendix 1: Posterior predictive for the HRR model

From (A.1), the joint distribution of (β, σ^2) is Normal-Inverse-Gamma, i.e.,

$$f(\beta, \sigma^2) = f(\beta|\sigma^2)f(\sigma^2) = \mathcal{N}(\mu_\beta, V_\beta) \cdot \mathcal{IG}(a, b) = \mathcal{NIG}(\mu_\beta, \sigma^2V_\beta, a, b).$$

Further we can know the posterior distribution of (β, σ^2) is still Normal-Inverse-Gamma $\mathcal{NIG}(\mu_\beta^*, V_\beta^*, a^*, b^*)$, where

$$\begin{aligned}\mu^* &= (V_\beta^{-1} + X^\top X)^{-1}(V_\beta^{-1}\mu_\beta + X^\top \mathbf{y}), \\ V^* &= (V_\beta^{-1} + X^\top X)^{-1}, \\ a^* &= a + \frac{n}{2}, \\ b^* &= b + \frac{1}{2}(\mu_\beta^\top V_\beta^{-1}\mu_\beta + \mathbf{y}^\top \mathbf{y} - \mu^{*\top} V^{*-1} \mu^*).\end{aligned}$$

Now we drive the posterior predictive w.r.t. individual response y_i .

$$\begin{aligned}f(y_i|\mathbf{y}) &= \int f(y_i|\beta, \sigma^2)f(\beta, \sigma^2|\mathbf{y})d\beta d\sigma^2 \\ &= \int \mathcal{N}(X_i\beta, \sigma^2) \cdot \mathcal{NIG}(\mu^*, V^*, a^*, b^*)d\beta d\sigma^2 \\ &= \int \frac{b^{*a^*}}{(2\pi)^{\frac{p+1}{2}}\Gamma(a^*)|V^*|^{1/2}} \left(\frac{1}{\sigma^2}\right)^{a^* + \frac{p+1}{2} + 1} \times \\ &\quad \exp\left\{-\frac{1}{\sigma^2}\left[b^* + \frac{1}{2}\left\{(\beta - \mu^*)^\top V^{*-1}(\beta - \mu^*) + (y_i - X_i\beta)^2\right\}\right]\right\} d\beta d\sigma^2 \\ &= \int \frac{b^{*a^*}}{(2\pi)^{\frac{p+1}{2}}\Gamma(a^*)|V^*|^{1/2}} \left(\frac{1}{\sigma^2}\right)^{a^* + \frac{p+1}{2} + 1} \times \\ &\quad \exp\left\{-\frac{1}{\sigma^2}\left[b^{**} + \frac{1}{2}(\beta - \mu^{**})^\top V^{**,-1}(\beta - \mu^{**})\right]\right\} d\beta d\sigma^2\end{aligned}$$

where

$$\begin{aligned}\mu^{**} &= (V_\beta^{*-1} + X^{*\top} X^*)^{-1}(V_\beta^{*-1}\mu_\beta^* + X^{*\top} \mathbf{y}^*), \\ V^{**} &= (V_\beta^{*-1} + X^{*\top} X^*)^{-1}, \\ b^{**} &= b^* + \frac{1}{2}(\mu_\beta^{*\top} V_\beta^{*-1}\mu_\beta^* + y_i^2 - \mu^{**\top} V^{**,-1}\mu^{**}).\end{aligned}$$

Let $z \triangleq \frac{c}{\sigma^2}$, $c \triangleq b^{**} + \frac{1}{2}(\beta - \mu^{**})^\top V^{**^{-1}}(\beta - \mu^{**})$, and then

$$\begin{aligned}
f(y_i|\mathbf{y}) &= \frac{b^{*a^*}}{(2\pi)^{\frac{p+1}{2}} \Gamma(a^*) |V^*|^{1/2}} \int c^{-(a^* + \frac{p+1}{2} + 1)} z^{a^* + \frac{p+1}{2} + 1} e^{-z} dc z^{-1} d\beta \\
&= \frac{b^{*a^*}}{(2\pi)^{\frac{p+1}{2}} \Gamma(a^*) |V^*|^{1/2}} \int c^{-(a^* + \frac{p+1}{2})} d\beta \\
&= \frac{b^{*a^*}}{(2\pi)^{\frac{p+1}{2}} \Gamma(a^*) |V^*|^{1/2}} \int \left[b^{**} + \frac{1}{2}(\beta - \mu^{**})^\top V^{**^{-1}}(\beta - \mu^{**}) \right] d\beta \\
&= \frac{b^{*a^*}}{(2\pi)^{\frac{p+1}{2}} \Gamma(a^*) |V^*|^{1/2}} b^{** - \frac{2a^* + p + 1}{2}} \times \\
&\quad \int \frac{\Gamma(\frac{2a^* + p + 1}{2})}{\Gamma(\frac{2a^* + 1}{2}) \pi^{p/2} |(2a^* + 1) [\frac{2b^{**}}{2a^* + 1} V^{**}]|^{\frac{1}{2}}} \left[1 + \frac{(\beta - \mu^{**})^\top [\frac{2b^{**}}{2a^* + 1} V^{**}]^{-1} (\beta - \mu^{**})}{2a^* + 1} \right]^{-\frac{2a^* + p + 1}{2}} d\beta.
\end{aligned}$$

The integrable function above is the density of β , which is actually multivariate t -distribution $MVSt_{2a^*+1}(\mu^{**}, \frac{2b^{**}}{2a^*+1} V^{**})$. Since

$$|2b^{**} V^{**}|^{1/2} = 2^{p/2} b^{**\frac{1}{2}} \frac{|V^*|^{1/2}}{|1 + X_i V^* X_i^\top|^{1/2}},$$

then we have

$$f(y_i|\mathbf{y}) = \frac{\Gamma(\frac{2a^*+1}{2})}{\sqrt{2a^*} \pi \Gamma(\frac{2a^*}{2}) [\frac{2b^*}{2a^*} (1 + X_i V^* X_i^\top)]^{1/2}} \left[1 + \frac{(y_i - X_i \mu^*)^2 \{ \frac{2b^*}{2a^*} (1 + X_i V^* X_i^\top) \}}{2a^*} \right]^{-\frac{2a^*+1}{2}}.$$

It is like a univariate t -distribution shifted by $y_i - X_i \mu^*$ and scaled by $\frac{2b^*}{2a^*} (1 + X_i V^* X_i^\top)$.

[Vehtari *et al.* \(2017\)](#) proposed the expected log pointwise predictive density (elpd) to measure the predictive accuracy for the new data \tilde{y}_i ($i = 1, \dots, n$). The elpd is defined as

$$\text{elpd} = \sum_{i=1}^n \int f(\tilde{y}_i) \log f(\tilde{y}_i|\mathbf{y}) d\tilde{y}_i.$$

Therefore, we use the log pointwise predictive density (lpd) to measure the predictive accuracy, i.e., $\text{lpd} = \sum_{i=1}^n \log f(y_i|\mathbf{y})$. The widely applicable information criterion (WAIC) is an alternative approach which is

$$\text{lpd} - \sum_{i=1}^n \text{Var}[\log f(y_i|\mathbf{y})].$$

Appendix 2: Posterior predictive for the dSUR and SSUR models

For the dSUR and SSUR models, the response variables are independent in their reparametrised forms. It is feasible to use the out-of-sample predictive to measure the elpd. The Bayesian leave-one-out estimate is

$$\text{elpd}_{\text{loo}} = \sum_{i=1}^n \log f(y_i|\mathbf{y}_{-i}).$$

As noted by the importance sampling, we get

$$f(y_i|\mathbf{y}_{-1}) \approx \frac{1}{\frac{1}{T} \sum_{t=1}^T \frac{1}{f(y_i|\boldsymbol{\theta}^t)}},$$

where all related parameters $\boldsymbol{\theta}^t$ are drawn from its full posterior. The WAIC is estimated by

$$\widehat{\text{elpd}}_{\text{waic}} = \widehat{\text{elpd}}_{\text{loo}} - \sum_{i=1}^n \text{Var}_{t=1}^T [\log f(y_i|\boldsymbol{\theta}^t)].$$

The posterior predictive $f(y_i|\mathbf{y}_{-1})$ can be used to check outliers, which is also named the Conditional Predictive Ordinate (CPO, Gelfand (1996)).

Affiliation:

Marco Banterle, Alex Lewin
 Department of Medical Statistics
 Faculty of Epidemiology and Population Health
 London School of Hygiene & Tropical Medicine
 Keppel St, Bloomsbury
 London WC1E 7HT, United Kingdom
 E-mail: marco.banterle@gmail.com, alex.lewin@lshtm.ac.uk

Leonardo Bottolo
 Department of Medical Genetics
 University of Cambridge
 J. J. Thomson Avenue
 Cambridge CB2 0QQ, United Kingdom
 E-mail: lb664@cam.ac.uk

Sylvia Richardson
 MRC Biostatistics Unit
 University of Cambridge
 Robinson Way
 Cambridge CB2 0SR, United Kingdom
 E-mail: sylvia.richardson@mrc-bsu.cam.ac.uk

Zhi Zhao, Manuela Zucknick
 Oslo Centre for Biostatistics and Epidemiology
 Department of Biostatistics
 Institute of Basic Medical Sciences
 University of Oslo
 P.O. Box 1122 Blindern
 0317 Oslo, Norway
 E-mail: zhi.zhao@medisin.uio.no, manuela.zucknick@medisin.uio.no