Bayesian Variable Selection Regression Of Multivariate Responses For Group Data

B. Liquet[1,2] and K. Mengersen[2] and A. N. Pettitt[2] and M. Sutton[2]

[1] LMAP, Université de Pau et des Pays de L'Adour'

[2] ACEMS, QUT, Australia

Contents

- **Motivation:**
  - Integrative Analysis for group data
  - Application to pleiotropy investigation
- **Sparse Model:** Lasso penalty, Group penalty, Sparse Group Lasso
- **Bayesian framework**
- **Simulation Studies**
- **Illustration: genomics data**
- **R package: MSGBSS**

Integrative Analysis

Goal of Integrative Analysis:

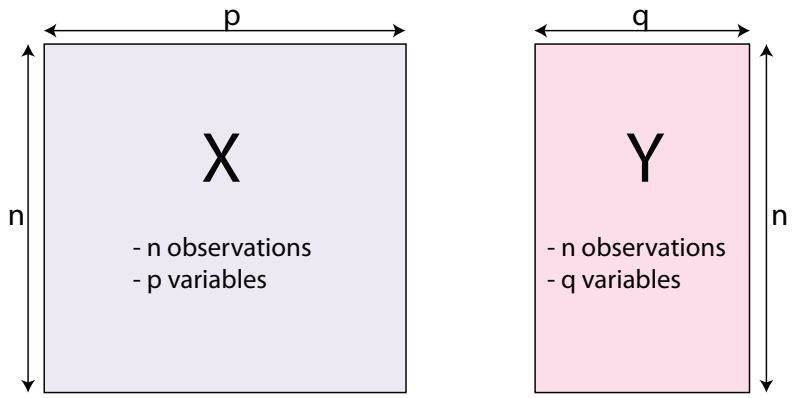
Wikipedia. **Data integration** "involves **combining data** residing in different sources and providing users with a unified view of these data. This process becomes significant in a variety of situations, which include both commercial and **scientific**".

System Biology. **Integrative Analysis:** Analysis of heterogeneous types of data from inter-platform technologies.

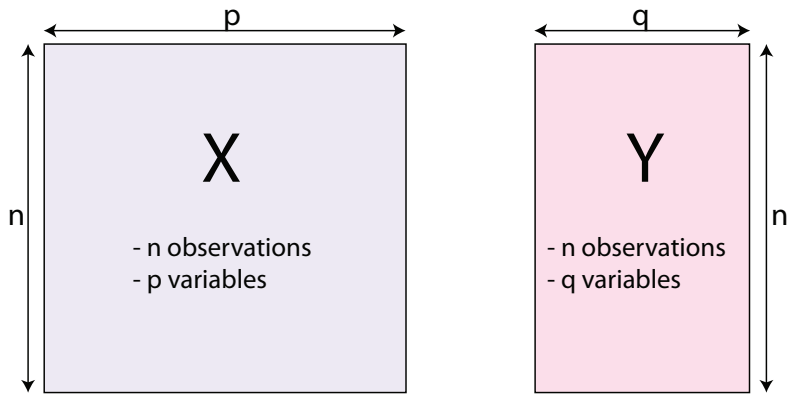
Goal. Combine multiple types of data:

- Contribute to a better understanding of biological mechanism.
- Have the potential to improve the diagnosis and treatments of complex diseases.

Example: Data definition



Example: Data definition



- "Omics." **Y** matrix: gene expression, **X** matrix: SNP (single nucleotide polymorphism). Many others such as proteomic, metabolomic data.
- "neuroimaging". **Y** matrix: behavioral variables, **X** matrix: brain activity (e.g., EEG, fMRI, NIRS)
- "neuroimaging genetics." **Y** matrix: fMRI (Fusion of functional magnetic resonance imaging), **X** matrix: SNP
- "Ecology/Environment." **Y** matrix: Water quality variables , **X** matrix: Landscape variables

Data: Constraints and Aims

- Main constraint: situation with $p > n$

Group structures within the data

- Natural example: Categorical variables which is a group of dummies variables in a regression setting.

Group structures within the data

- Natural example: Categorical variables which is a group of dummies variables in a regression setting.
- Genomics: genes within the same pathway have similar functions and act together in regulating a biological system.

$\hookrightarrow$ These genes can add up to have a larger effect

$\hookrightarrow$ can be detected as a group (i.e., at a pathway or gene set/module level).

Group structures within the data

- Natural example: Categorical variables which is a group of dummies variables in a regression setting.
- Genomics: genes within the same pathway have similar functions and act together in regulating a biological system.

$\hookrightarrow$ These genes can add up to have a larger effect

$\hookrightarrow$ can be detected as a group (i.e., at a pathway or gene set/module level).

We consider variables are divided into groups:

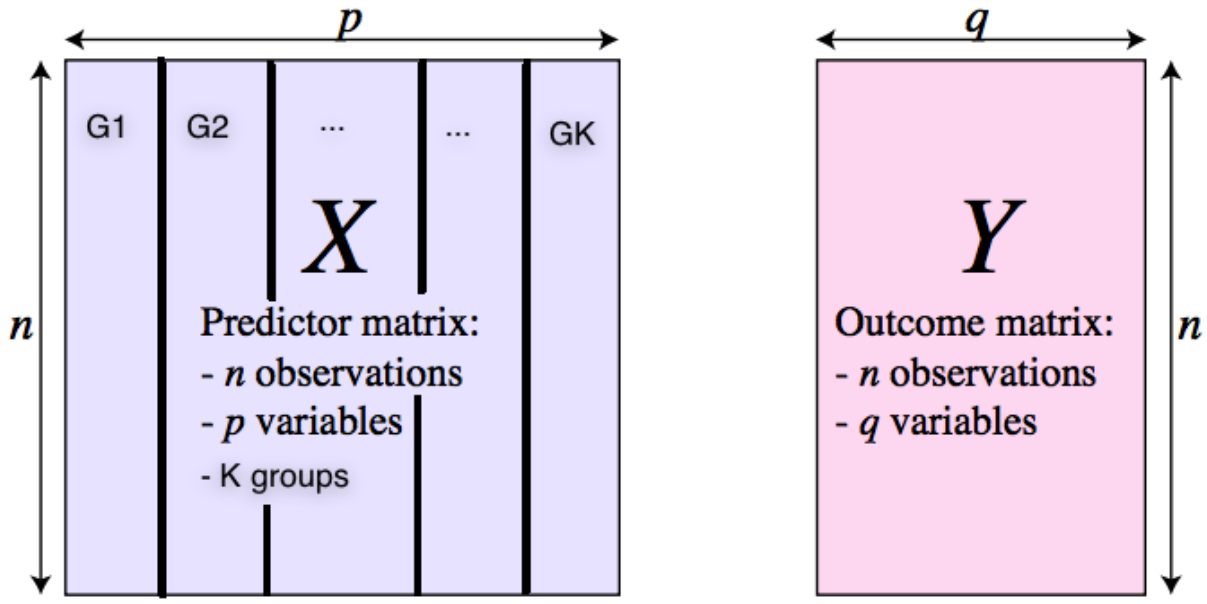
- Example p: SNPs grouped into K genes

$$\mathbf{X} = [\underbrace{SNP_1, \ldots + SNP_k}_{gene_1} \mid \underbrace{SNP_{k+1}, SNP_{k+2}, \ldots, SNP_h}_{gene_2} \mid \ldots \mid \underbrace{SNP_{l+1}, \ldots, SNP_p}_{gene_K}]$$

- Example p: genes grouped into K pathways/modules ($X_j = \text{gene}_j$)

$$\mathbf{X} = [\underbrace{X_1, X_2, \ldots, X_k}_{M_1} \mid \underbrace{X_{k+1}, X_{k+2}, \ldots, X_h}_{M_2} \mid \ldots \mid \underbrace{X_{l+1}, X_{l+2}, \ldots, X_p}_{M_K}]$$

Aims in regression setting:



- Select group variables taking into account the data structures; all the variables within a group are selected otherwise none of them are selected
- Combine both sparsity of groups and within each group; only relevant variables within a group are selected

Some frequentist Appoaches

- Lasso models: regression model
  - use L_1 and L_2 penalties for performing variable selection

Lasso models

Univariate model : $Y = \mathbb{X}\boldsymbol{\beta} + \varepsilon,\ Y \in \mathfrak{R}^n, \boldsymbol{\beta} \in \mathfrak{R}^p$

- **Lasso regression:** sparse model with L_1 penalty

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\text{lasso}} &= \underset{\boldsymbol{\beta}\in\mathfrak{R}^p}{\text{argmin}} \sum_{i=1}^{n}(Y_i - X_i^T\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p}|\beta_j| \\ &= \underset{\boldsymbol{\beta}\in\mathfrak{R}^p}{\text{argmin}} \underbrace{||Y - \mathbb{X}\boldsymbol{\beta}||_2^2}_{\text{Loss}} + \lambda \underbrace{||\boldsymbol{\beta}||_1}_{\text{Penalty}}\end{aligned}$$

- **Group Lasso regression:** sparse model using group structure with L_2 penalty on group ($\mathbb{X}_g$ with $\boldsymbol{\beta}_g \in \mathbb{R}^{p_g}$)

$$\underset{\beta\in\mathbb{R}^p}{min}\left(||Y - \sum_{g=1}^{G}\mathbb{X}_g\boldsymbol{\beta}_g||_2^2 + \lambda \sum_{g=1}^{G}\sqrt{p_g}||\boldsymbol{\beta}_g||_2\right)$$

-If $p_g = 1 \Longrightarrow$ Lasso method ($||\boldsymbol{\beta}_g||_2 = \sqrt{\boldsymbol{\beta}_g^2} = |\boldsymbol{\beta}_g|$)

Lasso models

Univariate model : $Y = \sum_{g=1}^{G} \mathbb{X}_g \boldsymbol{\beta}_g + \varepsilon,\ Y \in \mathfrak{R}^n, \boldsymbol{\beta} \in \mathfrak{R}^p$

- **Sparse Group Lasso regression:** sparse group model using group structure combining L_1 and L_2 penalties

$$\min_{\beta \in \mathbb{R}^p} \left(\| Y - \sum_{g=1}^{G} \mathbb{X}_g \boldsymbol{\beta}_g \|_2^2 + \lambda_1 \sum_{g=1}^{G} \sqrt{p_g} \| \boldsymbol{\beta}_g \|_2 + \lambda_2 \| \boldsymbol{\beta} \|_1 \right)$$

Lasso model for Multivariate $\mathbb{Y}$

$$\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{E}$$

where $\mathbb{Y}$ $(n \times q)$, $\mathbb{B}$ $(p \times q)$

- Sparse model: select predictors associated to the multiple phenotype Y $(n \times q)$:

$$\min_{\boldsymbol{B}} \left(\|\mathbb{Y} - \mathbb{X}\mathbb{B}\|_F^2 + \lambda \sum_{l=1}^{p} \|\boldsymbol{\beta}_l\|_2 \right)$$

where $\mathbb{B} = \begin{pmatrix} \beta_1^1 & \beta_1^2 & \dots & \beta_1^q \\ \beta_2^1 & \beta_2^2 & \dots & \beta_2^q \\ \vdots & \vdots & \vdots & \vdots \\ \beta_p^1 & \beta_p^2 & \dots & \beta_p^q \end{pmatrix}$, $\boldsymbol{\beta}_l = (\beta_l^1, \beta_l^2, \dots, \beta_l^q)$

- Implemented in `glmnet()`
- Sparse Group model: see e.g., Li et al. (2015), Wand et al. (2012).

Bayesian Framework

- Univariate regression model: $Y - \sum_{g=1}^{G} \mathbb{X}_g \beta_g \sim N_n\left(0, \sigma^2 \mathbb{I}_n\right)$
  - Xu and Ghosh (2015): Bayesian group lasso using spike and slab priors for group variable selection.
  - Rockova and Lesafre (2014): (EM) algorithm for a hierarchical model incorporating grouping information.
  - Stingo et al. (2011): PLS approach for pathway and gene selection using variable selection priors and Markov chain Monte Carlo (MCMC)

Bayesian Framework

- Multivariate regression model:

$$\mathbb{Y} - \sum_{g=1}^{G} \mathbb{X}_g \mathbb{B}_g \sim MN_{n\times q}\left(\mathbf{0}_{p\times q}, \mathbb{I}_n, \Sigma\right)$$

  - Zhu et al. (2014): Bayesian generalized low rank regression model.
  - Greenlaw et al. (2016): Bayesian hierarchical modeling for imaging genomics data. Limited to $\Sigma = \sigma^2 \mathbb{I}_q$.

Bayesian group lasso model with spike and slab priors

Xu and Ghosh (2015) approaches:

- **spike and slab** priors providing variable selection at the group level.
- hierarchical spike and slab prior structure to select variables both at the group level and within each group.
- Limited to **univariate case**.
- Doesn't take into account group size in the penalisation part.

Multivariate Bayesian Group Lasso with Spike and Slab prior

$$\mathbb{Y}|\mathbb{X}, \mathbb{B}, \Sigma \sim MN_{n\times q}(\mathbb{X}\mathbb{B}, \Sigma, \mathbb{I}_n), \tag{1}$$

$$Vec(\mathbb{B}_g^T|\Sigma, \tau_g, \pi_0) \overset{ind}{\sim} (1-\pi_0)N_{m_g q}(0, \mathbb{I}_{m_g} \otimes \tau_g^2\Sigma) + \pi_0\delta_0(Vec(\mathbb{B}_g^T)), \tag{2}$$

$$\tau_g^2 \overset{ind}{\sim} \text{Gamma}\left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right), \quad g = 1, \ldots, G, \tag{3}$$

$$\Sigma \sim \text{IW}(d, Q) \tag{4}$$

$$\pi_0 \sim \text{Beta}(a, b) \tag{5}$$

where $\delta_0(Vec(\mathbb{B}_g^T))$ denotes a point mass at $\mathbf{0} \in \mathbb{R}^{m_g q}$, $\mathbb{B}_g$ is the $m_g \times q$ regression coefficient matrix for the group g

Calibration of λ

- λ is related to the coefficient of shrinkage.
- tuned using an empirical Bayes approach.
- a Monte Carlo EM algorithm is used.
- the kth EM update for λ is:

$$\lambda^{(k)} = \sqrt{\frac{p+G}{\sum_{g=1}^{G} \mathrm{E}_{\lambda^{(k-1)}}\left[\tau_g^2 | \mathbb{Y}\right]}},$$

in which the posterior expectation of τ_g^2 is replaced by the Monte Carlo sample average of τ_g^2 generated in the Gibbs sample based on $\lambda^{(k-1)}$.

Median Thresholding Estimator

- In the case of a block orthogonal design matrix $\mathbb{X}$ (i.e., $\mathbb{X}_i^T \mathbb{X}_j = 0$ for $i \neq j$), we have for $1 \leqslant g \leqslant G$

$$Vec(\widehat{\mathbb{B}}_g^T) = Vec\left(((\mathbb{X}_g^T \mathbb{X}_g)^{-1} \mathbb{X}_g^T \mathbb{Y})^T\right) \sim N_{m_g q}\left(Vec(\mathbb{B}_g^T), (\mathbb{X}_g^T \mathbb{X}_g)^{-1} \otimes \Sigma\right).$$

- We showed assuming $\pi_0 > \frac{c}{1+c}$, then there exists $t(\pi_0) > 0$, such that the marginal posterior median of β_{ij}^g satisfies

$$\text{Med}(\beta_{ij}^g | \widehat{\mathbb{B}}_g) = 0 \quad \text{for any } 1 \leqslant i \leqslant m_g \text{ and } 1 \leqslant j \leqslant q$$

  when $\|Vec(\widehat{\mathbb{B}}_g^T)\|_2 < t$.
- The marginal posterior median estimator of the gth group of regression coefficients is zero when the norm of the corresponding block least square estimator is less than a certain threshold.

Posterior median as a soft thresholding estimator

- Assuming an orthogonal design matrix $\mathbb{X}$, i.e., $\mathbb{X}^T\mathbb{X} = n\mathbb{I}_p$ and consider our model defined with fixed τ_g^2 $(1 \leqslant g \leqslant G)$.
- Posterior distribution of $\mathbb{B}_g$ is a spike and slab distribution,

$$Vec(\mathbb{B}_g^T)|\mathbb{X}, \mathbb{Y} \sim (1 - I_g)N_{m_g q}\left((1 - D_g)Vec(\mathbb{B}_{LS,g}^T), \frac{1 - D_g}{n}\mathbb{I}_{m_g} \otimes \Sigma\right) + I_g\delta_0(Vec(\mathbb{B}_g^T)),$$

where $\mathbb{B}_{LS,g}$ is the least squares estimator of $\mathbb{B}_g$, $D_g = \frac{1}{1+n\tau_g^2}$, and $I_g = p(\mathbb{B}_g = 0|rest)$

$$I_g = \frac{\pi_0}{\pi_0 + (1 - \pi_0)(\tau_g^2)^{-\frac{m_g(q-1)}{2}}(1 + n\tau_g^2)^{-\frac{m_g}{2}} \exp\left\{(1 - D_g)nTr[\Sigma^{-1}\mathbb{B}_{LS,g}^T\mathbb{B}_{LS,g}]\right\}}$$

- The marginal posterior distribution of β_{ij}^g is also a spike and slab distribution,

$$\beta_{ij}^g|\mathbb{X}, \mathbb{Y} \sim (1 - I_g)N\left((1 - D_g)\widehat{\beta}_{LS,ij}^g, \frac{1 - D_g}{n}\Sigma_{jj}\right) + I_g\delta_0(\beta_{ij}^g),$$

where Σ_{jj} is the j-th diagonal element of Σ.

Posterior median as a soft thresholding estimator

- The resulting median is a soft thresholding estimator defined by

$$\widehat{\beta}_{ij}^{Med,g} = \mathrm{Med}(\beta_{ij}^{g}|\mathbb{X},\mathbb{Y}) = \mathrm{sgn}\left(\widehat{\beta}_{LS,ij}^{g}\right)\left((1-D_g)|\widehat{\beta}_{LS,ij}^{g}| - \frac{\sqrt{\Sigma_{jj}}}{\sqrt{n}}Q_g\sqrt{1-D_g}\right)_+,$$

where z_+ denotes the positive part of z and $Q_g = \phi^{-1}(\frac{1}{2(1-\min(\frac{1}{2},l_g))})$.

- For a univariate response ($q = 1$) the matrix Σ reduces to the scalar σ^2, and our result matches the previous work of Xu et al (2015).
- In the multivariate frequentist setting, Li et al. (2015) proposed an iterative algorithm with similar soft thresholding function to incorporate group structure in estimating the regression estimates.

Oracle property

- Let $\mathbb{B}^0, \mathbb{B}_g^0, \beta_{ij}^{0,g}$ denote the true values of $\mathbb{B}, \mathbb{B}_g, \beta_{ij}^g$, respectively.
- The index vector of the true model as $\mathcal{A} = (I(\|Vec(\mathbb{B}_g)\|_2 \neq 0), g = 1, \ldots, G)$,
- The index vector of the model selected by certain thresolding estimator $\widehat{\mathbb{B}}_g$ as $\mathcal{A}_n = (I(\|Vec(\widehat{\mathbb{B}}_g)\|_2 \neq 0), g = 1, \ldots, G)$.
- Model selection consistency is attained if and only if $\lim_n P(\mathcal{A}_n = \mathcal{A}) = 1$.

Oracle property

- Under orthogonal design the median thresholding estimator has oracle property.
- **Theorem**: Assume orthogonal design matrix, i.e., $\mathbb{X}^T\mathbb{X} = n\mathbb{I}_p$. Suppose $\sqrt{n}\tau_{g,n}^2 \to \infty$ and $\log(\tau_{g,n}^2)/n \to 0$ as $n \to \infty$, for $g = 1, \ldots, G$, then the median thresholding estimator has oracle property, that is, variable selection consistency,

$$\lim_{n\to\infty} \mathrm{P}(\mathcal{A}_n^{Med} = \mathcal{A}) = 1$$

and asymptotic normality,

$$\sqrt{n}\left(Vec(\widehat{\mathbb{B}}_{\mathcal{A}}^{Med}) - Vec(\mathbb{B}_{\mathcal{A}}^{0})\right) \xrightarrow{d} N(\mathbf{0}, \Sigma \otimes \mathbb{I}).$$

Gibbs Sampler: full posterior distribution

$$p(\mathbb{B}, \tau^2, \Sigma, \pi_0 | \mathbb{Y}, \mathbb{X}) \quad \propto \quad p(\mathbb{Y}|\mathbb{B}, \tau^2, \Sigma, \pi_0) \times p(\mathbb{B}|\tau^2, \Sigma, \pi_0)$$
$$\times p(\tau^2) \times p(\Sigma) \times p(\pi_0),$$

where

$$p(\mathbb{Y}|\mathbb{B}, \tau^2, \Sigma, \pi_0) \propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} Tr\left[(\mathbb{Y} - \mathbb{X}\mathbb{B})\Sigma^{-1}(\mathbb{Y} - \mathbb{X}\mathbb{B})^T\right]\right\},$$

$$p(\mathbb{B}|\tau^2, \Sigma, \pi_0) = \prod_{g=1}^{G} p(\mathbb{B}_g|\tau_g^2, \Sigma, \pi_0),$$

$$p(\mathbb{B}_g|\tau_g^2, \Sigma, \pi_0) \propto (1-\pi_0)(2\pi)^{-\frac{qm_g}{2}} (\tau_g^2)^{-\frac{qm_g}{2}} |\Sigma|^{-\frac{m_g}{2}} \exp - \left\{\frac{1}{2\tau_g^2} Tr\left[\mathbb{B}_g \Sigma^{-1} \mathbb{B}_g^T\right]\right\} I[\mathbb{B}_g \neq 0]$$
$$+\pi_0 \delta_0(Vec(\mathbb{B}_g^T)),$$

$$p(\tau_1, \ldots, \tau_g) \propto \prod_{g=1}^{G} (\lambda^2)^{\frac{qm_g+1}{2}} (\tau_g^2)^{\frac{qm_g+1}{2}-1} \exp\left(-\frac{\lambda^2 m_g}{2} \tau_g^2\right),$$

$$p(\pi_0) \propto \pi_0^{a-1}(1-\pi_0)^{b-1},$$

$$p(\Sigma) \propto |\Sigma|^{-\frac{d+q+1}{2}} \exp\left\{-\frac{1}{2} Tr(Q\Sigma^{-1})\right\}$$

Conditional posterior distribution

Let $\mathbb{B}_{(g)}$ denote the $\mathbb{B}$ matrix without the gth group, and $\mathbb{X}_{(g)}$ denote the covariate matrix corresponding to $\mathbb{B}_{(g)}$, that is,

$$\mathbb{X}_{(g)} = (\mathbb{X}_1, \ldots, \mathbb{X}_{g-1}, \mathbb{X}_{g+1}, \ldots, \mathbb{X}_G)$$

where $\mathbb{X}_g$ is the design matrix corresponding to $\mathbb{B}_g$.

- Conditional posterior distribution of $\mathbb{B}_g$: spike and slab distribution
- Conditional posterior distribution of $\alpha_g^2 = \frac{1}{\tau_g^2}$
  - Inverse Gamma if $\mathbb{B}_g = 0$
  - Inverse Gaussian if $\mathbb{B}_g \neq 0$
- Conditional posterior distribution of Σ: Inverse Wishart
- Conditional posterior distribution of π_0: Beta distribution

Multivariate Sparse Group Selection with Spike and Slab Prior (MBSGS-SS)

- Reparametrize the coefficients matrices to tackle the two kinds of sparsity separately:

$$\mathbb{B}_g = \boldsymbol{V}_g^{\frac{1}{2}} \widetilde{\mathbb{B}}_g, g = 1, \ldots, G;\ j = 1, \ldots, m_g,$$

  with $\boldsymbol{V}_g^{\frac{1}{2}} = \mathrm{diag}\{\tau_{g_1}, \ldots, \tau_{gm_g}\},\ \tau_{gj} \geqslant 0$ and where $\widetilde{\mathbb{B}}_g$, when nonzero, follow $Vec(\widetilde{\mathbb{B}}_g^T) \sim N_{m_g q}(\mathbf{0}, \mathbb{I}_{m_g} \otimes \Sigma)$.

- Diagonal element of $\boldsymbol{V}_g^{\frac{1}{2}}$ control the magnitude of the elements of $\mathbb{B}_g$.

Multivariate Sparse Group Selection with Spike and Slab Prior (MBSGS-SS)

- To select variables at the group level, we assume the multivariate spike and slab prior for each $Vec(\widetilde{\mathbb{B}}_g^T)$:

$$Vec(\widetilde{\mathbb{B}}_g^T|\Sigma, \tau_g, \pi_0) \overset{ind}{\sim} (1-\pi_0)N_{m_g q}(0, \mathbb{I}_{m_g} \otimes \Sigma) + \pi_0 \delta_0(Vec(\widetilde{\mathbb{B}}_g^T))$$

- Note that when $\tau_{gj} = 0$, the j-th row of $\mathbb{B}_g$ is essentially dropped out of the model even when $\widetilde{\mathbb{B}}_g^j \neq 0$.
- So in order to choose variables within each relevant group, we assume the following spike and slab prior for each τ_{gj}:

$$\tau_{gj} \overset{ind}{\sim} (1-\pi_1)N^+(0, s^2) + \pi_1 \delta_0(\tau_{gj}), \quad g = 1, \ldots, G;\ j = 1, \ldots, m_g,$$

where $N^+(0, s^2)$ denotes a normal $N(0, s^2)$ distribution truncated below at 0.

Prior Specification

- We assume an Inverse Wishart prior for $\Sigma \sim \mathrm{IW}(d, Q)$
- We assume conjugate beta hyper-priors for π_0 and π_1.
- We use a conjugate inverse gamma prior for $s^2 \sim \text{Inverse Gamma}(1, t)$,
- We stimate t with the Monte Carlao EM algorithm. For the k-th EM update,

$$t^{(k)} = \frac{1}{\mathrm{E}_{t^{(k-1)}}\left[\frac{1}{s^2}|\mathbb{Y}\right]},$$

where the posterior expectation of $\frac{1}{s^2}$ is estimated from the Gibbs samples based on $t^{(k-1)}$.

Gibbs Sampler: full posterior distribution

- Joint posterior

$$
\begin{aligned}
&p(\widetilde{\mathbb{B}}, \tau^2, \Sigma, \pi_0, \pi_1, s^2 | \mathbb{Y}, \mathbb{X}) \\
&\quad \propto |\Sigma|^{-n/2} \exp\left\{ -\frac{1}{2} Tr\left[\left(\mathbb{Y} - \sum_{g=1}^{G} \mathbb{X}_g \boldsymbol{V}_g^{\frac{1}{2}} \widetilde{\mathbb{B}}_g \right) \Sigma^{-1} \left(\mathbb{Y} - \sum_{g=1}^{G} \mathbb{X}_g \boldsymbol{V}_g^{\frac{1}{2}} \widetilde{\mathbb{B}}_g \right)^T \right] \right\} \\
&\quad \times \prod_{g=1}^{G} (1-\pi_0)(2\pi)^{-\frac{qm_g}{2}} |\Sigma|^{-\frac{m_g}{2}} \exp\left\{ -\frac{1}{2} Tr\left[\widetilde{\mathbb{B}}_g \Sigma^{-1} \widetilde{\mathbb{B}}_g^T \right] \right\} I[\widetilde{\mathbb{B}}_g \neq 0] \\
&\qquad + \pi_0 \delta_0(Vec(\widetilde{\mathbb{B}}_g^T)) \\
&\quad \times \prod_{g=1}^{G} \prod_{j=1}^{m_g} \left[(1-\pi_1) 2 (2\pi s^2)^{-\frac{1}{2}} exp\left\{ -\frac{\tau_{gj}^2}{2s^2} \right\} I[\tau_{gj} > 0] + \pi_1 \delta_0(\tau_{gj}) \right] \\
&\quad \times |\Sigma|^{-\frac{d+q+1}{2}} \exp\left\{ -\frac{1}{2} Tr(Q\Sigma^{-1}) \right\} \\
&\quad \times \pi_0^{a_1 - 1} (1-\pi_0)^{a_2 - 1} \\
&\quad \times \pi_1^{c_1 - 1} (1-\pi_1)^{c_2 - 1} \\
&\quad \times t (s^2)^{-2} \exp\left\{ -\frac{t}{s^2} \right\}
\end{aligned}
$$

Simulation Studies

- Comparison to Lasso, group Lasso, sparse goup Lasso for univariate setting
  - R package **glmnet** and **SGL**
- Comparison to Lasso for Mulivariate setting
  - R package **glmnet**
- Bayseian approaches
  - BGL-SS (Bayesian Group Lasso with Spike and Slab prior)
  - BSGS-SS (Bayesian Sparse Group selection with spike and slab priors)
  - MBGL-SS for multivariate setting
  - MBSGL-SS for multivariate setting
  - running 20000 iterations in which the first 10000 are burn-ins.

Simulation results: summary

- Simulations results suggest that the multivariate Bayesian group lasso with spike and slab prior is strongly influenced by a combination of different group size structures and high correlation between predictors.
- The multivariate Bayesian sparse group selection with spike and slab prior does not suffer in this situation.
- Most powerful method for variable selection and prediction performance in the presence of group structure data
- All numerical results from our article could be reproduced using our R package `MBSGS` available on CRAN.

Computation

- The current version of our package runs for example:
  - a MBSGS-SS model in around 2 minutes
  - a MBGL-SS model in around 1 minute for a model with 20 groups of 5 variables
  - for a sample size (n = 900) with 20000 iterations including 10000 for the burnin.
- Further improvements of the code such as parallelization over the group structure are in progress to speed up the computational time for tacking Big Data sets
- In this context of genetics studies, some further extensions of our model are under investigation such as integrating different group penalties given a biological prior of the pathways or different distribution priors for each group.

Application: Aim and Data

- Identify a parsimonious set of predictors that explains the joint variability of gene expression in four tissues (adrenal gland, fat, heart, and kidney).
- 770 SNPs in 29 inbred rats as a predictor matrix (n = 29, p = 770)
- 29 measured expression levels in the 4 tissues as the outcome (q = 4).

Application

	Correlation				Summary statistics	
	ADR	Fat	Heart	Kidney	Mean	Variance
ADR	1.00	0.46	0.44	0.70	4.72	0.07
Fat		1.00	0.24	0.42	8.23	0.09
Heart			1.00	0.44	8.79	1.61
Kidney				1.00	6.65	0.07

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Group size	74	67	63	60	39	45	52	43	31	51	21	26	33	22	15	27	18	30	34	19

- chromosome information defines the group structure of the predictor matrix
- ran our MBGL-SS and MBSGS-SS models using this group structure
- The multivariate lasso selects 69 SNPs which come from the 20 chromosomes.
- The MBGL-SS selects only the two first groups corresponding to the SNPs from chromosomes 1 and 2.

Application: MBSGS-SS results, 32 SNPs from 8 chromosomes

Chromosome	SNP Name	ADR	Fat	Heart	Kidney
2	D2Rat147	0.00553	0.00238	-	0.00329
2	D2Rat222	0.00442	0.00116	-	0.00305
2	D2CebrP476s2	0.00123	-	-	-
2	D2Rat69	0.00715	0.01748	0.00730	0.00620
2	D4Ucsf2	0.00054	-	-	-
2	D7Cebr205s3	0.00246	-	0.00950	0.00461
3	D7Cebr14C16s2	0.00209	0.00326	-	0.00049
4	D7Rat112	0.00035	0.00001	-	-
4	D7Rat19	0.01113	0.01800	0.03680	0.01828
4	Cyp11b2	0.00075	0.00374	-	0.00394
7	D10Ntr32	0.00123	-	0.01112	0.00143
7	D10Rat31	0.00031	0.00573	0.00442	0.00316
7	D10Cebr39s2	0.00280	0.00490	0.00821	0.00586
7	Es13	0.00539	-	0.00924	0.00419
7	D10Rat226	0.00415	0.00006	0.00987	0.00372
7	D14Rat36	0.00036	-	0.03076	-
7	D14Cebrp312s2	0.00004	-	0.05427	-
10	**D14Mit3**	**0.04963**	**0.05415**	**0.33434**	**0.07491**
10	D15Rat21	0.00937	0.00569	0.03140	0.01704
10	D19Utr1	0.00149	0.00297	0.00251	0.00487
10	Ednra	0.00026	-	-	-
10	D2Mit16	-	0.00077	-	-
10	D2Rat70	-	0.00190	-	-
10	D3Cebr204s4	-	0.00042	-	-
14	D4Rat49	-	0.00102	0.00092	0.00401
14	D7Mit6	-	0.00002	-	-
14	D10Rat102	-	0.00112	-	-
14	D4Rat252	-	-	-0.00184	-
14	Myc	-	-	0.00669	-
15	D10Mit3	-	-	0.00104	-
19	D14Rat8	-	-	0.00058	-
19	D14Rat52	-	-	0.00361	-

Application: results

- SNP D14Mit (from chromosome 10) has been previously identified
- Highest estimate (0.334) for the heart tissue
- the posterior standard deviation of the regression parameter for each selected non-zero median estimate was in the range 0.11 to 0.64.
- SNP D14Mit3 estimate was 0.334 with posterior standard deviation 0.639
- Importance of chromosomes could be investigated using an estimate of the probability of inclusion

Chromosome	1	2	3	4	5	6	7	8	9	10
EPI	0.00	1.00	1.00	1.00	0.72	0.00	1.00	0.00	0.00	1.00
Chromosome	11	12	13	14	15	16	17	18	19	20
EPI	0.83	0.12	0.46	1.00	0.93	0.88	0.79	0.59	0.89	0.40

Concluding remarks: summary

- Bayesian methods for group-sparse modeling in the context of a multivariate correlated response variable.
- Our models are based on spike and slab type priors which facilitate variable selection.
- Importance to include the group size information in the shrinkage part of our model.
- We have shown that the posterior median estimator could both select and estimate the regression coefficients.
- Simulation results suggest very good performance of the posterior median estimator for variable selection and prediction error.
- This estimator obtains similar results as the highest probability model in terms of true and false positive rates.

References

- Li, Y., Nan, B., and Zhu, J. (2015). *Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure.* Biometrics, 71(2): 354–363.

- Liquet, B., Lafaye de Micheaux, P., Hejblum, B., and Thiebaut, R. (2016). *Group and Sparse Group Partial Least Square Approaches Applied in Genomics Context.* Bioinformatics, 3(1): 35–42.

- Liquet, B. and Sutton, M. (2016). *MBSGS: Multivariate Bayesian Sparse Group Selection with Spike and Slab*. R package version 1.0.0. URL http://CRAN.R-project.org/package=MBSGS

- Liquet, B., Mengersen, K., Pettitt, A. N. and Sutton, M. (2016). *Bayesian Variable Selection Regression Of Multivariate Responses For Group Data* Bayesian Analysis. Volume 12, Number 4 (2017), 1039-1067.

- Received the Lindley award from ISBA Prize Committee, delivered during ISBA World Meeting in Edinburgh 2018

- Xu, X. and Ghosh, M. (2015). *Bayesian Variable Selection and Estimation for Group Lasso.* Bayesian Analysis, 10(4): 909–936.

Extension to Pleotropic mapping for genome-wide association studies using group variable selection

- Pleiotropy: genetic variants which affect multiple different complex diseases
- Example: genetic variants which affect both Breast and Thyroid cancer.
- Results from GWAS suggest that complex diseases are often affected by many variants with small effects (known as polygenicity)

- Aims:
  - statistical method to leverage pleiotropic effects
  - incoporate prior pathway knowledge to increase statistical power and identify important risk variants.

Model for multiple GWAS studies

- Suppose we have data from K independent GWAS datasets, $\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K$, where $\mathcal{D}_k = (\{y_1, x_1\}, \ldots, \{y_{n_k}, x_{n_k}\})$
- $y_{ik} \in \{0, 1\}$ denotes the phenotype of the kth study
- $x_{ik} \in \mathbb{R}^p$ is the vector with corresponding p SNPs.
- Logistic regression model

$$y_{ik} \sim \text{Bernoulli}\left(g^{-1}(\nu_{ik})\right)$$
$$\nu_{ik} = x_{ik}^T \boldsymbol{\beta}_{\cdot k}$$

  for $k = 1, \ldots, K$, where $g(\cdot)$ is the logistic link function
- $\boldsymbol{\beta}_{\cdot k} \in \mathbb{R}^p$ the regression coefficients for the kth GWAS.
- Let $\boldsymbol{\beta}_{j\cdot} \in \mathbb{R}^K$, $j = 1, \ldots, p$, the vector of K regression coefficients corresponding to the jth SNP over the K GWAS.

Group Structure

- SNPs can be partitioned into G groups (genes)
- Let $\pi_g, g = 1, \ldots, G$ the set of SNPs contained in the gth group with $p_g = |\pi_g|$.
- Matrix of all regression coefficients as $\mathbf{B} = (\boldsymbol{\beta}_{\cdot 1}, \ldots, \boldsymbol{\beta}_{\cdot K}) = (\boldsymbol{\beta}_{1\cdot}, \ldots, \boldsymbol{\beta}_{p\cdot})^T$.

Frequentist Approach

- The log likelihood for the combined datasets:

$$p(\mathcal{D} \mid \mathbf{B}) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} L\left(y_{ik} \boldsymbol{\beta}_k^T x_{ik}\right) \quad \text{where} \quad L(x) = -\ln(1 + e^{-x})$$

- The penalised likelihood estimate

$$\widehat{\mathbf{B}} = \underset{\mathbf{B} \in \mathbb{R}^{p \times K}}{\operatorname{argmin}} \left\{ - \sum_{k=1}^{K} \sum_{i=1}^{n_k} L\left(y_{ik} \boldsymbol{\beta}_k^T x_{ik}\right) + \lambda_1 \|\mathbf{B}\|_{G_{2,1}} + \lambda_2 \|\mathbf{B}\|_{\ell_{2,1}} \right\} \quad (6)$$

- $G_{2,1}$-norm penalty $\|\mathbf{B}\|_{G_{2,1}} = \sum_{g=1}^{G} \sqrt{\sum_{i \in \pi_g} \sum_{j=1}^{K} \beta_{ik}^2}$
- $\ell_{2,1}$-norm penalty $\|\mathbf{B}\|_{\ell_{2,1}} = \sum_{i=1}^{p} \sqrt{\sum_{k=1}^{K} \beta_{ik}^2}$ respectively.
- The $G_{2,1}$-norm fixes the group structure across studies and encourages sparsity at a gene level.
- The $\ell_{2,1}$-norm which allows sparsity within a group.

Inference

- Inference using the alternating direction method of multipliers algorithm (ADMM).
- Novel approach for identifying pleiotropic effects as it accounts for gene specific and SNP specific effects using a variable selection approach.
- The method is only capable of producing a point estimate of **B** and acurrate estimation of the variance for these parameters is not easily given.

Bayesian Logistic regression with multivariate spike and slab prior: LogitMBGL-SS

- Let $\gamma = (\gamma_1, \ldots, \gamma_p)^T$ indicate the association status for SNPs where $\gamma_j = 1$ indicates that the jth SNP is associated to all K traits.
- Spike and slab prior for the jth SNP $\boldsymbol{\beta}_{j\cdot} \in \mathbb{R}^K$,

$$
\begin{aligned}
\boldsymbol{\beta}_{j\cdot} &\sim (1-\gamma_j)\mathcal{N}_K(0, \tau_j^2 \mathbf{V}) + \gamma_j \delta_0(\boldsymbol{\beta}_{j\cdot}) \\
\tau_j^2 &\sim \text{Gamma}\left(\frac{K+1}{2}, \frac{\lambda}{2}\right), \\
\mathbf{V} &\sim IW(d, Q), \\
\gamma_j &\sim \text{Bernolli}(\alpha_0) \\
\alpha_0 &\sim \text{Beta}(a, b)
\end{aligned}
$$

  for $j = 1, \ldots, p$, where $\delta_0(\boldsymbol{\beta}_j)$ denotes a point mass at $\mathbf{0} \in R^K$.
- Here, $\mathbf{V} \in \mathbb{R}^{K \times K}$ is a covariance matrix modeling the covariance of the SNP effect on the traits.

Extension

- Should perform well when the SNPs are independent.
- GWAS datasets: strong correlations that can occur between SNPs within the same gene.
- Solution: reparameterise the coefficeints to handle the sparsity at a gene grouping level and individual feature level separately.
- $\tau \in \mathbb{R}^p$ to model individual sparsity
- $\mathbf{b}^{(g)} \in \mathbb{R}^{p_g K}$ with $\mathbf{b}^{(g)} = (\mathbf{b}_1^{(g)^T}, \ldots, \mathbf{b}_{p_g}^{(g)^T})$ where $\mathbf{b}_j^{(g)} \in \mathbb{R}^K$ for group sparsity.

$$\beta_{j\cdot} = \tau_j \mathbf{b}_j^{(g)}, \qquad \text{where } \tau_j \geqslant 0, \quad \text{for all } j \in \pi_g.$$

Bayesian Logistic regression using multivariate sparse group selection with spike and slab priors

$$\boldsymbol{\beta}_{j\cdot} = \tau_j \mathbf{b}_j^{(g)}, \qquad \text{where } \tau_j \geqslant 0, \quad \text{for all } j \in \pi_g.$$

We assume the following multivariate spike and slab

$$\begin{aligned}
\mathbf{b}^{(g)} &\sim (1-\alpha_0)\mathcal{N}_{p_g K}(0, \mathbb{I}_{p_g} \otimes \mathbf{V}) + \alpha_0 \delta_0(\mathbf{b}^{(g)}) \\
\tau_j &\sim (1-\alpha_1)\mathcal{N}^{+}\left(0, s^2\right) + \alpha_1 \delta_0(\tau_j), \\
\alpha_0 &\sim \text{Beta}(a_1, a_2) \\
\alpha_1 &\sim \text{Beta}(c_1, c_2) \\
s^2 &\sim \text{InvGamma}(1, t)
\end{aligned}$$

for $j \in \pi_g$ and $g = 1, \ldots, G$

Signal recovery:

(i) GMT: Grouped multi-task penalised logistic regression ($\lambda_1 > 0, \lambda_2 = 0$) using $G_{2,1}$-norm

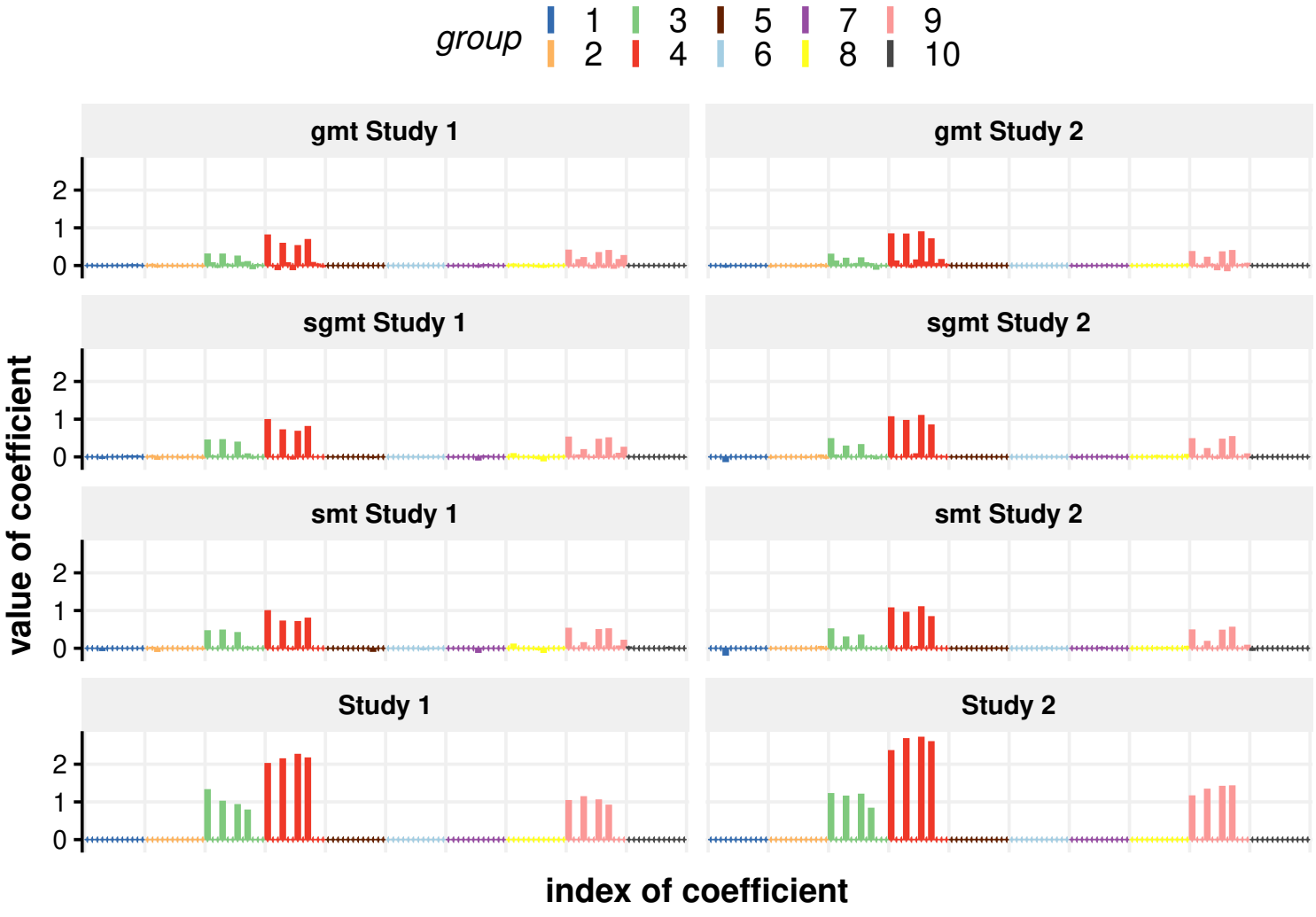
(ii) SMT: Sparse multi-task penalid penalised logistic regression ($\lambda_1 = 0, \lambda_2 > 0$) using $\ell_{2,1}$-norm

(iii) SGMT: Sparse group multi-task penalised logistic regression ($\lambda_1 > 0, \lambda_2 > 0$)

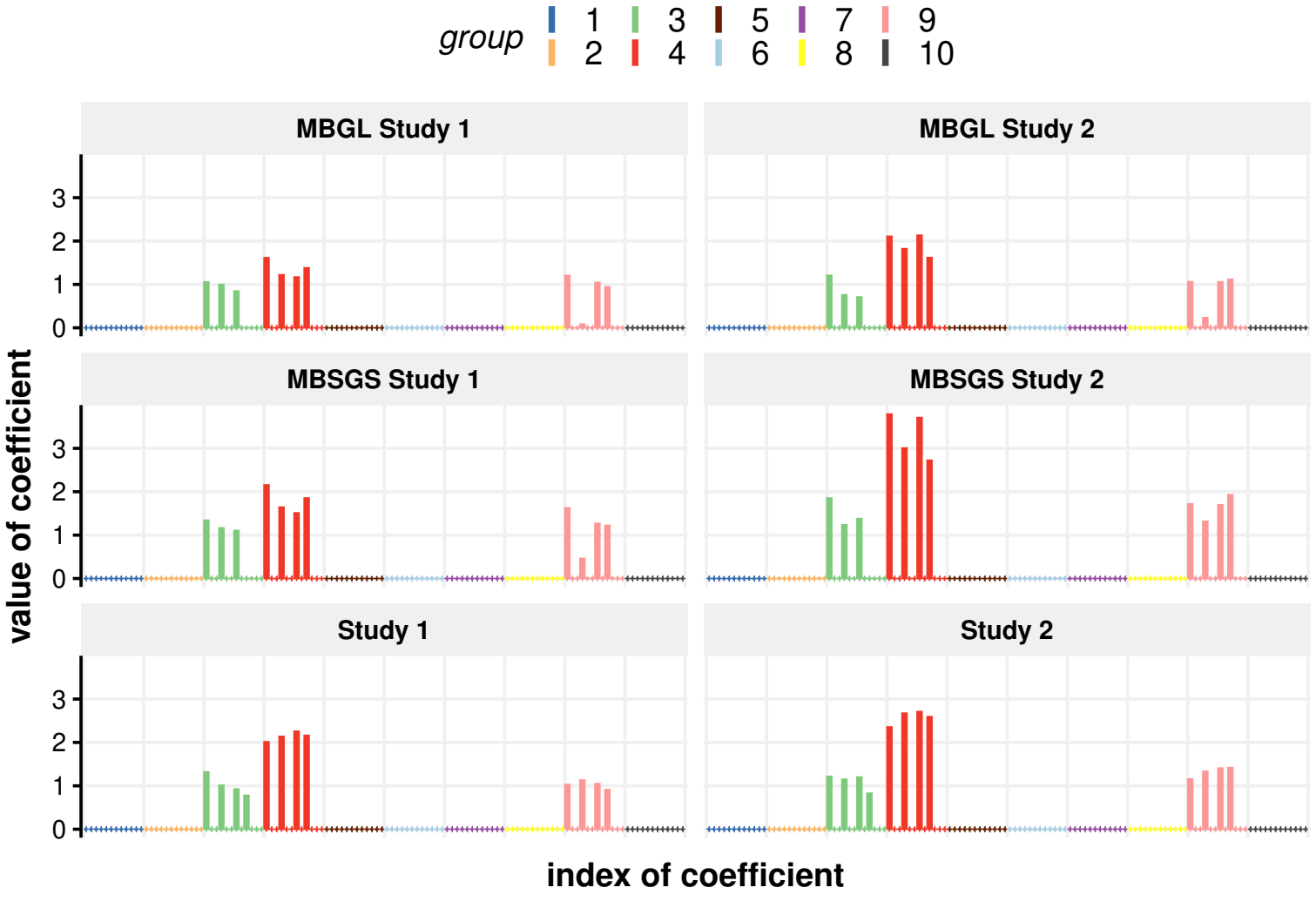
(iv) LOGITMBGL: Bayesian logistic regression using multivariate group lasso with spike and slab prior

(v) LOGITMBSGS: Bayesian logistic regression using multivariate sparse group selection with spike and slab prior

Results: Frequentist



Results: Bayesian



Main Conclusion on the simulation studies

- The penalised approaches perform reasonably well in variable selection but the reconstructed signal is underestimated.
- In general, the penalised likelihood methods suffer in terms of false negatives, selecting more variables to be nonzero than the Bayesian methods.
- The Bayesian methods perform the best in terms of signal recovery measured by the ℓ_1 error and variable selection performance metrics.
- The penalised likelihood approaches are computationally efficient using alternating direction method of multipliers algorithm
- Simulation results suggest that when computationally possible the Bayesian estimators should be used.
- The multivariate Bayesian sparse group selection with spike and slab prior performed the best in terms of signal recovery.
- The Bayesian method provides a natural method for quantifying the variability of the estimated coefficients.

What Next ?

- Application on real data: case/control studies
  - Breast Cancer and Thyroide Cancer
  - Thyroide Cancer (482 case, 463 control)
  - Breast Cancer (1172 case, 1125 control)
  - 6677 SNPs from 618 genes from 10 non-overlapping gene pathways.

What Next ?

- Application on real data: case/control studies
  - Breast Cancer and Thyroide Cancer
  - Thyroide Cancer (482 case, 463 control)
  - Breast Cancer (1172 case, 1125 control)
  - 6677 SNPs from 618 genes from 10 non-overlapping gene pathways.

Looking for one postdoc position to fill for 2 years granted by “la ligue contre le Cancer”

https://lma-umr5142.univ-pau.fr/fr/vie-du-laboratoire/recrutement/post-doctorat.html