Training course 3: Bayesian variable selection approach for structured variables.

Marie Denis and Mahlet G. Tadesse

January 25, 2021

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Outline

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Introduction

Objectives are:

- to present a Bayesian variable selection approach developed for structured variables,
- to present results on simulated datasets,
- to apply on a subset of your data.

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Context

Dependence among covariates may be induced by various factors in different applications:

- Disease mapping: structure in space and time for covariates measured over time at adjacent locations,
- Genetic: spatial dependence among molecular markers positioned at adjacent positions along genome,
- Genomic: dependence among genes belonging to same pathways,
- Functional MRI: spatial dependence among measures in connected regions in the brain,
- ...

↬ Most of these dependence structures may be depicted by an **undirected graph**.

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Statistical point of view

Such information may:

- help in the model building process,
- increase the statistical power to detect true associations,
- improve the predictive power.

Objective

To develop statistical methods incorporating the graph structure to simultaneously select the relevant variables and estimate their associated effects.

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Existing methods

In frequentist context:

- Approaches based on penalized likelihood methods (Li and Li, 2008, 2010):

↬ smoothness at the regression coefficient level by integrating a graph information.

In Bayesian context:

- Spike-and-slab prior on the regression coefficients with an Ising prior (prior considering graph information) on the variable selection indicators (Li and Zhang, 2010; Vannucci and Stingo, 2010),
- Priors on shrinkage parameters for global-local priors (Kowal et al., 2019)

↬ Smoothness at the variable selection indicator or at the shrinkage parameter levels,

- Bayesian version of the frequentist approaches (Zhou and Zheng, 2013).

The proposed approach:

We propose combining a Gaussian Markov random field (GMRF) prior with a global-local shrinkage prior to select graph-structured variables in the same spirit as Faulkner and Minin (2018).

Outline

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Statistical model

Linear regression model

$$y = X\beta + \varepsilon$$

where

- $y = (y_1, \ldots, y_n)'$ a n-dimensional vector of observations,
- $\beta = (\beta_1, \ldots, \beta_p)'$ a p-dimensional vector of regression coefficients,
- X a $n \times p$ matrix of predictor variables,
- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ a n-vector of residuals assumed to follow a multivariate Gaussian distribution $\mathcal{N}_n(0, \sigma^2 I_n)$.

Graph information for the p predictors:

$\Rightarrow$ $\mathcal{G}$ an **undirected graph** such that $\mathcal{G} = \bigcup_{i=1}^{I} \mathcal{G}_i = \bigcup_{i=1}^{I} (V_i, E_i)$ is a disjoint union of I graphs where V_i is a finite set of vertices (or predictors) associated to the graph i, and E_i is the associated set of edges.

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Statistical model

Bayesian hierarchical model

$$
\begin{aligned}
y|\beta,\sigma^2 &\sim \mathcal{N}_n(X\beta,\sigma^2 I_n)\\
\beta_j-\beta_{j'}|\tau^2_{jj'},\,\lambda^2 &\sim \mathcal{N}(0,\lambda^2\tau^2_{jj'}) \text{ for } (j,j')\in\bigcup_{i=1}^{I} E_i\\
\beta_j|\tau^2_j,\,\lambda^2 &\sim \mathcal{N}(0,\lambda^2\tau^2_j) \text{ for } j\in\mathcal{S}\\
\tau_{jj'} &\sim \mathcal{C}^+(0,1) \text{ for } (j,j')\in\bigcup_{i=1}^{I} E_i\\
\tau_j &\sim \mathcal{C}^+(0,1) \text{ for } j\in\mathcal{S}\\
\lambda &\sim \mathcal{C}^+(0,\sigma)\\
\sigma^2 &\sim \mathcal{IG}(s,r)
\end{aligned}
$$

With $\mathcal{S}$ is the set of indices associated to one representative of each graph, $\mathcal{C}^+$ the half-Cauchy distribution, $\mathcal{IG}$ the inverse-gamma distribution.

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Statistical models

HS-GMRF

$$\beta_j - \beta_{j'} | \tau^2_{jj'}, \lambda^2 \sim \mathcal{N}(0, \lambda^2 \tau^2_{jj'}) \text{ for } (j, j') \in \bigcup_{i=1}^{I} E_i$$

- $\tau^2_{jj'}$: local shrinkage parameters providing flexibility in the amount of shrinkage and smoothness of the coefficients of the connected variables,
- λ^2: global parameter performing shrinkage on all coefficients.

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Statistical models

HS-GMRF

$$\beta_j - \beta_{j'} | \tau^2_{jj'}, \lambda^2 \sim \mathcal{N}(0, \lambda^2 \tau^2_{jj'}) \text{ for } (j, j') \in \bigcup_{i=1}^{I} E_i$$

- $\tau^2_{jj'}$: local shrinkage parameters providing flexibility in the amount of shrinkage and smoothness of the coefficients of the connected variables,
- λ^2: global parameter performing shrinkage on all coefficients.

HS-GMRF-sign

$$\beta_j - \delta_{jj'} \beta_{j'} | \tau^2_{jj'}, \lambda^2 \sim \mathcal{N}(0, \lambda^2 \tau^2_{jj'}) \text{ for } (j, j') \in \bigcup_{i=1}^{I} E_i$$

- $\delta_{jj'}$: the correlation sign among X_j and X'_j.

Both models are inferred with a Markov chain Monte Carlo algorithm.

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Outline

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Outline

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Simulated datasets

Objective

To evaluate the performance of the proposed approaches for varying number of connected predictors (q), varying level of correlation among them (ρ), varying values of regression coefficients (3 simulation models), and two designs (1 and 2).

Design	ρ	q	**Design**	ρ	q
1	0.5	low (126 edges)	2	0.5	low (63 edges)
1	0.5	high (238 edges)	2	0.5	high (119 edges)
1	0.9	low (126 edges)	2	0.9	low (63 edges)
1	0.9	high (238 edges)	2	0.9	high (119 edges)

Design 1: all covariates are connected / **Design 2**: half of the covariates are connected and the remaining are not.

A total of 24 scenarios were analyzed by four methods: horseshoe (HS) , HS-GMRF, HS-GMRF-sign, spike-and-slab with Ising prior for $n = 100$ observations and $p = 140$ covariates. HS model refers to the model assuming a HS prior (Carvalho et al., 2010) on the regression coefficients (no structure information).

Simulated datasets

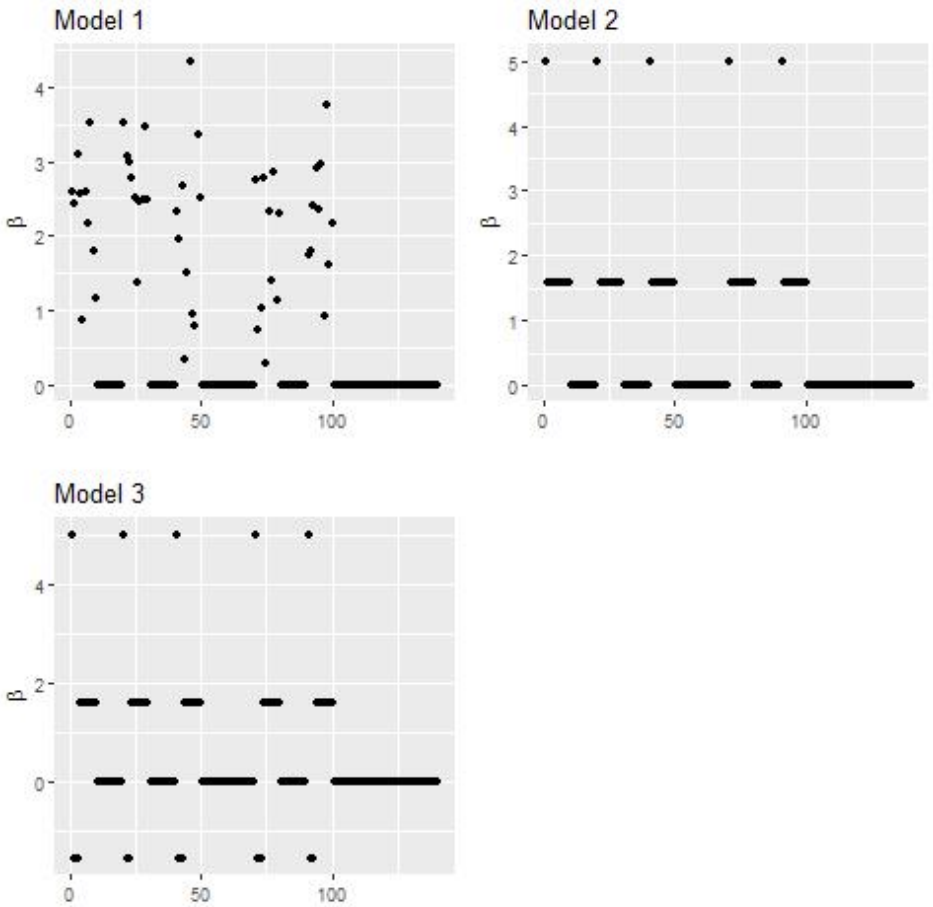
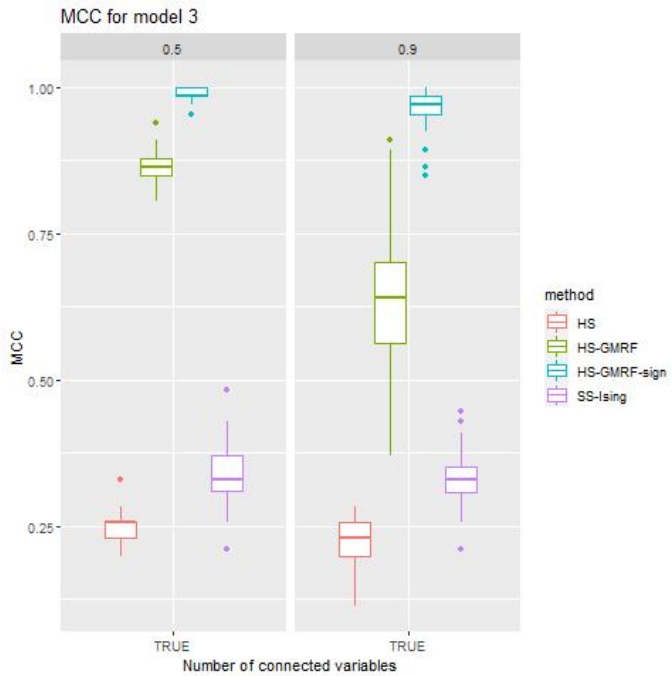


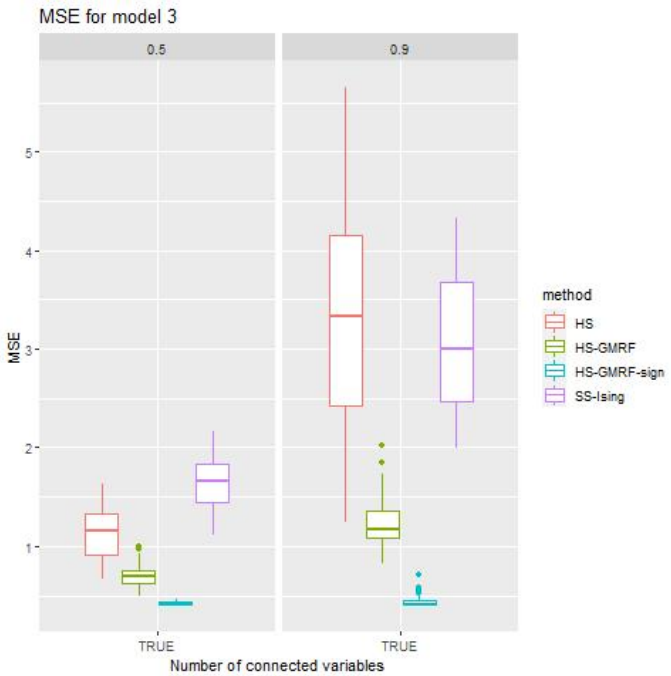
Figure 1: Simulated βs for the 140 covariates under simulation models 1, 2, and 3.

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Results

Boxplots of Matthews correlation coefficients (MCC) (a combined measure of the overall variables selection accuracy) on the left and mean squared errors (MSE) for model simulation 3 under design 1:

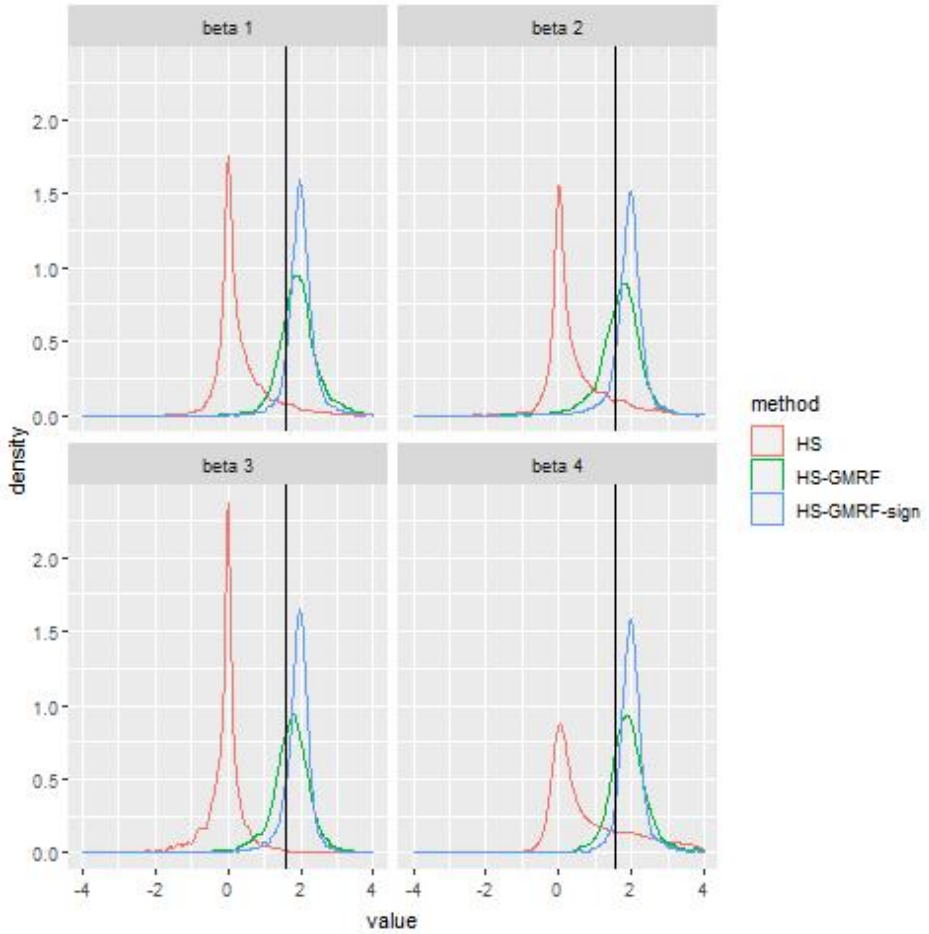




This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Results

Posterior densities for four βs for HS, HS-GMRF, and HS-GMRF-sign models:



This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Results

By taking into account the structure information among the covariates the proposed method gives better results in terms of selection and estimation for all scenarios compared to the other approaches.

- **Higher number of edges**: helps the selection and estimation however when the correlation is high it encourages similar estimated values for the coefficients associated to highly correlated predictors

$\Rightarrow$ Hypothesis that highly correlated predictors have similar effects.

- **Integrating the correlation** sign among the predictors improves the results.
- The selection and estimation of regression coefficients for the non connected variables is the same as the HS prior $\Rightarrow$ **flexibility** in the shrinkage due to both local and global parameters.
- In terms of interpretation this approach allows identifying groups of correlated variables $\Rightarrow$ results **more interpretable**.

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Outline

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Two real datasets

Two datasets were analyzed:

- Application 1: The study of riboflavin production based on gene expressions.
- Application 2: The study of fat content based on near-infrared absorbency spectra of meat samples.

- Graph obtained from partial correlations.
- Graph obtained from *a priori* knowledge.

- Identification of groups of correlated genes, better predictions (results obtained with a cross-validation procedure).
- Identification of wavelengths, better predictions (results obtained with a cross-validation procedure).

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Outline

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Context

Multi-Omic inTegrative Analysis (MOTA) approach was developed to analyze multi-omic data for ranking candidate disease biomarkers (Fan et al., 2020).

Data

- Three sets of omic data generated from the same set of samples: metabolomics (M), glycomics (G), and proteomics (P).
- Three study cohorts (TU, GU1, GU2).

Principle

- Network obtained from intra-omic and inter-omic connections

↬ MOTA score combines z-scores obtained from p-values calculated using Student's t-tests.

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Ranking of HCC-associated metabolites

Table 3. Biomarker candidates overlapping between the GU1, TU, and combined TU and GU1 datasets ranked by *t*-test, iDINGO, and MOTA.

Rank	GU1 Cohort	TU Cohort	GU1+TU Cohort	No. of Overlaps
		Ranking using **Student *t*-Test (*p*-Value)**		
1	ethanolamine	glutamic acid	ethanolamine	
2	phenylalanine	lactic acid	sorbose	
3	sorbose	alpha tocopherol	citric Acid	
4	pyroglutamic acid	valine	isoleucine	
5	glycine	ethanolamine	threitol	2
6	linoleic acid	alpha-D-glucosamine 1-phosphate	ribose	
7	creatinine	norvaline	malic acid	
8	lauric acid	citric Acid	phenylalanine	
9	ribitol /arabitol	norleucine	stearic acid	
10	threitol	sorbose	trans-aconitic acid	
		Ranking using **iDINGO**		
1	linoleic acid	norvaline	valine	
2	isoleucine	cystine	ethanolamine	
3	leucine	sorbose	butanediol	
4	proline	tagatose	ribose	
5	ethanolamine	isoleucine	glycine	
6	valine	trans-3-hydroxy-L-proline	sorbose	2
7	glutamic acid	N,N-dimethyl-1-4-phenylenediamine	tyrosine	
8	sorbose	cholesterol	malic acid	
9	aspartic acid	butanediol	isoleucine	
10	glycine	arachidic acid	tagatose	
		Ranking using **MOTA**		
1	tyrosine	alpha tocopherol	alpha tocopherol	
2	alpha tocopherol	tyrosine	ethanolamine	
3	pyroglutamic acid	ethanolamine	glycine	
4	glycine	creatinine	lactic acid	
5	ethanolamine	tyramine	creatinine	
6	phenylalanine	mimosine	tyrosine	4
7	citric acid	lactic acid	cholesterol	
8	threitol	cholesterol	tyramine	
9	tyramine	threitol	citric Acid	
10	aspartic acid	ribose	isoleucine	

Note: Metabolite candidates that appeared in the top-10 ranked lists of all three cohorts are highlighted with the same color.

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Results for metabolomics data

Results from MOTA select metabolites which were not identified by the other approaches:

- MOTA considers information from all the three platforms as well as the relationships among them

↬ Example of **tyrosine** which is identified by MOTA but has a high p-value with Student t-test $\Rightarrow$ **tyrosine** highly connected with low-p-value proteins and glycans

The proposed approach may be an alternative approach to select significant markers while considering relationships among the three sets of omic data through an undirected graph.

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Data

Focus on TU cohort with a total of 66 metabolites, 82 glycans, and 100 proteins for 86 patients. For the following we will note X the scaled 86×248 matrix of predictors such that:

$$X = [M, G, P]$$

and Y the 86 vector of binary observations with 0 for CIRR and 1 for HCC.

Objective is to identify predictors related to the disease status:

- ↬ Multivariable approaches to analyze simultaneously all markers using penalized approaches such as Lasso or horseshoe (HS) prior

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Univariate analysis

As a first step we also perform univariate analyses for X the scaled matrix of the three sets of omic data:

```
> res.student <- apply(X, 2, function(c) t.test(c~Y)$p.value)
> fdr.student <- p.adjust(res.student, method =  "fdr")
> head(sort(fdr.student))
```

```
[33032] L-glutamic acid 2                    P00747                    P02743
           0.0001570735              0.0221825895              0.0221825895
                 P13598                    P19320                    P29622
           0.0221825895              0.0221825895              0.0221825895
```

```
> head(sort(fdr.student[1:ncol(X.tu)]))  # focus on metabolites
```

```
                     [33032] L-glutamic acid 2
                                  0.0001570735
            [107689] L-(+) lactic acid [6.851]
                                  0.0317582528
                    [2116] alpha tocophereol
                                  0.0903087788
                  [700] ethanolamine [9.879]
                                  0.0993613578
[740] alpha-D-glucosamine 1-phosphate¬† [16.61]
                                  0.1021115952
                                    L-Valine 2
                                  0.1116471870
```

Mutivariable analysis

Lasso regression with the package glmnet

```
> library(glmnet)
> lambdas <- 10^seq(2, -3, by = -.1)
> # cross-validation procedure
> cv_lasso <- cv.glmnet(X, Y, alpha = 1, lambda = lambdas,
+                       family = "binomial")
> # largest value of lambda such that error is
> # within 1 standard error of the minimum
> optimal_lambda <- cv_lasso$lambda.1se
> lasso_model <- glmnet(X, Y, alpha = 1, lambda = optimal_lambda,
+                       family = "binomial")
> which(lasso_model$beta[,1] != 0)
[33032] L-glutamic acid 2
                       1
```

- Only one metabolite is selected

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Mutivariable analysis

HS regression with the package bayesreg

```
> library(bayesreg)
> HS.model <- bayesreg(Y~X, data = data.frame(Y, X),
+                      prior = "hs",n.samples = 10000,
+                      burnin = 5000, model = "binomial")
> library(coda)
> # posterior means
> beta.hs <- rowMeans(HS.model$beta[,-(1:5000)])
> # Highest posterior density (HPD) intervals
> ci <- t(apply(HS.model$beta[,-(1:5000)], 1, function(r) HPDinterval(as.mcmc(r), prob =0.95)))
> colnames(X)[which(apply(sign(ci), 1, prod) == 1)]
character(0)
```

- No metabolite is selected

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

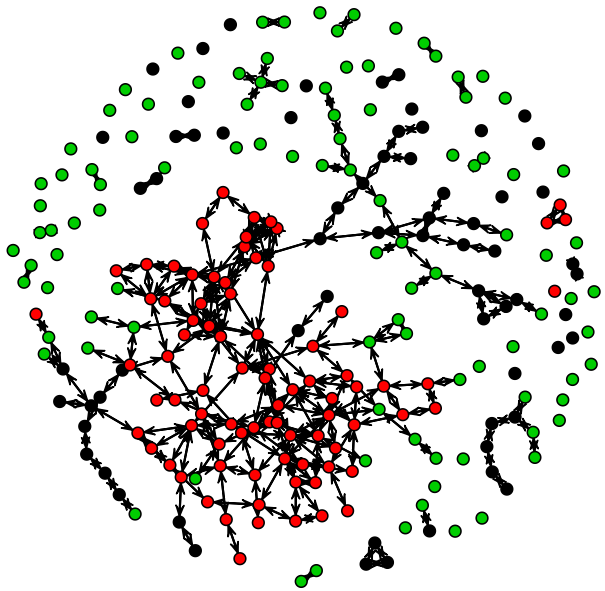
Undirected graph

In the objective to infer an undirected graph, the partial correlations were computed. A threshold of 0.45 was applied to obtain the following network:

Network

```
> library(network)
> plot(network(G), vertex.col = col.data)
```

black = Metabolomics, red = Glycomics, green = Proteomics



This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

HS-GMRF-sign model

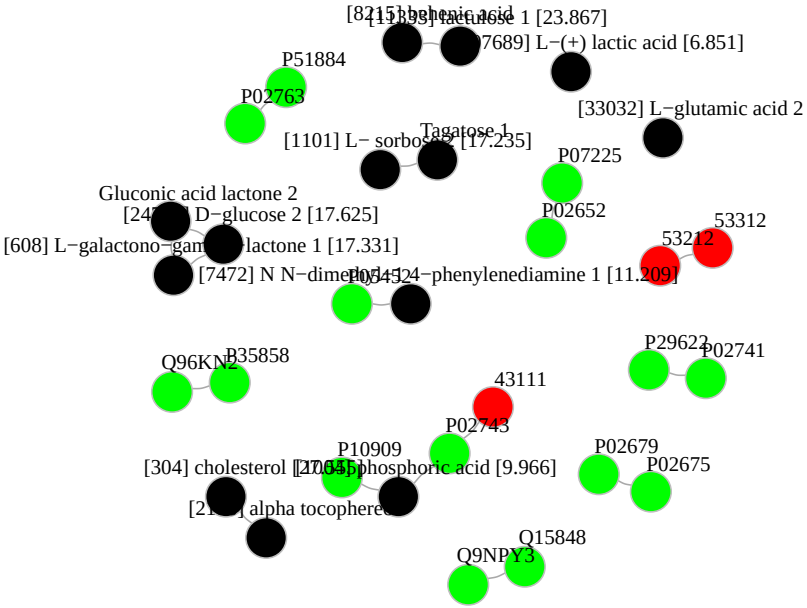
HS-GMRF-sign model was applied:

```
> MRF_prob <- MRF_prob(Z = as.numeric(Y)-1, X = X, C = C,iter = iter)
> burn <- 5000
> iter <- 10000
> # posterior means
> beta.MRF <- colMeans(MRF_prob$theta_store[-(1:burn),][seq(1,iter-burn, by = 10),])
> # Highest posterior density (HPD) intervals
> ci.prob.95 <- t(apply(MRF_prob$theta_store[-(1:burn),][seq(1,iter-burn, by = 10),],
+                   2, function(r) HPDinterval(as.mcmc(r), prob = 0.95)))
> head(colnames(X)[which(apply(sign(ci.prob.95), 1, prod) == 1)])
[1] "[33032] L-glutamic acid 2"          "[107689] L-(+) lactic acid [6.851]"
[3] "[2116] alpha tocophereol"           "Tagatose 1"
[5] "[1101] L- sorbose 2 [17.235]"       "[304] cholesterol [27.555]"
> ci.prob.90 <- t(apply(MRF_prob$theta_store[-(1:burn),][seq(1,iter-burn, by = 10),],
+                   2, function(r) HPDinterval(as.mcmc(r), prob = 0.90)))
> #colnames(X)[which(apply(sign(ci.prob.90), 1, prod) == 1)]
```

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

HS-GMRF-sign model

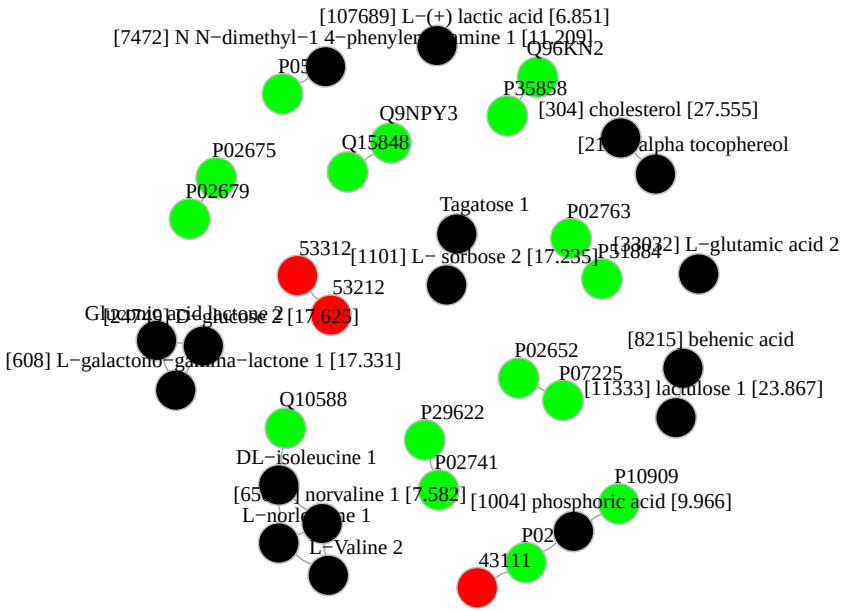
Network for variables selected using a 95% HPD:



This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

HS-GMRF-sign model

Network for variables selected using a 90% HPD:



This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

HCC-associated metabolites identified with the proposed approach with 90% HPD

Table 3. Biomarker candidates overlapping between the GU1, TU, and combined TU and GU1 datasets ranked by *t*-test, iDINGO, and MOTA.

Rank	GU1 Cohort	TU Cohort	GU1+TU Cohort	No. of Overlaps
		Ranking using **Student *t*-Test (*p*-Value)**		
1	ethanolamine	glutamic acid	ethanolamine	
2	phenylalanine	lactic acid	sorbose	
3	sorbose	alpha tocopherol	citric Acid	
4	pyroglutamic acid	valine	isoleucine	
5	glycine	ethanolamine	threitol	2
6	linoleic acid	alpha-D-glucosamine 1-phosphate	ribose	
7	creatinine	norvaline	malic acid	
8	lauric acid	citric Acid	phenylalanine	
9	ribitol /arabitol	norleucine	stearic acid	
10	threitol	sorbose	trans-aconitic acid	
		Ranking using **iDINGO**		
1	linoleic acid	norvaline	valine	
2	isoleucine	cystine	ethanolamine	
3	leucine	sorbose	butanediol	
4	proline	tagatose	ribose	
5	ethanolamine	isoleucine	glycine	2
6	valine	trans-3-hydroxy-L-proline	sorbose	
7	glutamic acid	N,N-dimethyl-1-4-phenylenediamine	tyrosine	
8	sorbose	cholesterol	malic acid	
9	aspartic acid	butanediol	isoleucine	
10	glycine	arachidic acid	tagatose	
		Ranking using **MOTA**		
1	tyrosine	alpha tocopherol	alpha tocopherol	
2	alpha tocopherol	tyrosine	ethanolamine	
3	pyroglutamic acid	ethanolamine	glycine	
4	glycine	creatinine	lactic acid	
5	ethanolamine	tyramine	creatinine	4
6	phenylalanine	mimosine	tyrosine	
7	citric acid	lactic acid	cholesterol	
8	threitol	cholesterol	tyramine	
9	tyramine	threitol	citric Acid	
10	aspartic acid	ribose	isoleucine	

Note: Metabolite candidates that appeared in the top-10 ranked lists of all three cohorts are highlighted with the same color.

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Conclusion/Perspectives

Conclusion

- The proposed method allows to integrate *a priori* knowledge on the structure among the predictors through an undirected graph.
- Results in terms of selection and prediction are better than approaches which do not integrate any information or which integrate information at the variable selection indicator level.

Perspectives

- The proposed method may be considered for analyzing multi omic data thanks to the integration of relationships among the covariates.
- To extend to model with a longitudinal response.

This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Outline

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Fan, Z., Zhou, Y., and Ressom, H. W. (2020). Mota: Network-based multi-omic data integration for biomarker discovery. *Metabolites*, 10(4):144.

Faulkner, J. R. and Minin, V. N. (2018). Locally adaptive smoothing with markov random fields and shrinkage priors. *Bayesian analysis*, 13(1):225.

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2019). Dynamic shrinkage processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):781–804.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182.

Li, C. and Li, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The annals of applied statistics*, 4(3):1498.

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American statistical association*, 105(491):1202–1214.

Vannucci, M. and Stingo, F. C. (2010). Bayesian models for variable selection that incorporate biological information. *Bayesian Statistics*, 9:1–20.

This project has received funding from The European Union's Horizon 2020 Research and Innovation programme under grant agreement No 840383.

Zhou, H. and Zheng, T. (2013). Bayesian hierarchical graph-structured model for pathway analysis using gene expression data. *Statistical applications in genetics and molecular biology*, 12(3):393–412.

This project has received funding from The European Unions's Horizon 2020 Research and Innovation programme Under grant agreement No 840383.