Training course 2: Introduction to the R package `BayesSUR`

Marie Denis

September 28, 2020

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 introduce the R package `BayesSUR`,
- 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.

The context

`BayesSUR` is an R package which implements several Bayesian models for high-dimensional regression of multiple responses (Banterle et al., 2018).

For example it allows to analyze several responses $\mathbf{Y}$, as gene expressions, and multiple predictors $\mathbf{X}$, as miRNAs:

$$\begin{array}{c} \\ \text{Ind } 1 \\ \vdots \\ \text{Ind } n \end{array} \begin{array}{c} \begin{array}{ccc} \text{Gene } 1 & \dots & \text{Gene } q \end{array} \\ \begin{bmatrix} y_1^{(1)} & \dots & y_1^{(q)} \\ & & \\ y_n^{(1)} & \dots & y_n^{(q)} \end{bmatrix} \end{array} \sim \begin{array}{c} \\ \text{Ind } 1 \\ \vdots \\ \text{Ind } n \end{array} \begin{array}{c} \begin{array}{ccc} \text{miRNA } 1 & \dots & \text{miRNA } p \end{array} \\ \begin{bmatrix} x_{11} & \dots & x_{1p} \\ & & \\ x_{n1} & \dots & x_{np} \end{bmatrix} \end{array}$$

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

The context

Which miRNAs are related to the correlated genes 1 and 2 ?

$$
\begin{array}{c} \\ I1 \\ \vdots \\ In \end{array}
\begin{array}{c}
\begin{array}{cccccc} \text{Gene 1} & \text{Gene 2} & \dots & \dots & \text{Gene } q \end{array} \\
\begin{bmatrix} y_1^{(1)} & y_1^{(2)} & & \dots & y_1^{(q)} \\ & & & & \\ y_n^{(1)} & y_n^{(2)} & & \dots & y_n^{(q)} \end{bmatrix}
\end{array}
\sim
\begin{array}{c} \\ I1 \\ \vdots \\ In \end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{miRNA 1} & \dots & \text{miRNA } p-1 & \text{miRNA } p \end{array} \\
\begin{bmatrix} x_{11} & \dots & x_{1p-1} & x_{1p} \\ & & & \\ x_{n1} & \dots & x_{np-1} & x_{np} \end{bmatrix}
\end{array}
$$

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

The context

Which miRNAs are related to the gene q?

$$
\begin{array}{c} \\ I1 \\ \vdots \\ In \end{array}
\begin{bmatrix} \text{Gene 1} & \text{Gene 2} & \dots & \dots & \text{Gene } q \\ y_1^{(1)} & y_1^{(2)} & & \dots & y_1^{(q)} \\ & & & & \\ y_n^{(1)} & y_n^{(2)} & & \dots & y_n^{(q)} \end{bmatrix}
\sim
\begin{array}{c} \\ I1 \\ \vdots \\ In \end{array}
\begin{bmatrix} \text{miRNA 1} & \dots & \text{miRNA } p-1 & \text{miRNA } p \\ x_{11} & \dots & x_{1p-1} & x_{1p} \\ & & & \\ x_{n1} & \dots & x_{np-1} & x_{np} \end{bmatrix}
$$

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

Classical approach

- A classical approach consists in performing a multiple regression model for each outcome k:

$$\overset{\text{Gene } k}{\begin{bmatrix} y_1^{(k)} \\ \vdots \\ y_n^{(k)} \end{bmatrix}} = \overset{\text{miRNA 1}}{\begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix}} b_1^{(k)} + \overset{\text{miRNA 2}}{\begin{bmatrix} x_{12} \\ \vdots \\ x_{n2} \end{bmatrix}} b_2^{(k)} + \ldots + \overset{\text{miRNA } p}{\begin{bmatrix} x_{1p} \\ \vdots \\ x_{np} \end{bmatrix}} b_p^{(k)} + \begin{bmatrix} \varepsilon_1^{(k)} \\ \vdots \\ \varepsilon_n^{(k)} \end{bmatrix}$$

with $b^{(k)} = (b_1^{(k)}, \ldots, b_p^{(k)})'$ the p-vector of regression coefficients and $\varepsilon^{(k)} = (\varepsilon_1^{(k)}, \ldots, \varepsilon_p^{(k)})'$ the p-vector of residuals associated to the outcome k.

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

Classical approach

- A classical approach consists in performing a multiple regression model for each outcome k:

$$\overset{\text{Gene } k}{\begin{bmatrix} y_1^{(k)} \\ \vdots \\ y_n^{(k)} \end{bmatrix}} = \overset{\text{miRNA 1}}{\begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix}} b_1^{(k)} + \overset{\text{miRNA 2}}{\begin{bmatrix} x_{12} \\ \vdots \\ x_{n2} \end{bmatrix}} b_2^{(k)} + \ldots + \overset{\text{miRNA } p}{\begin{bmatrix} x_{1p} \\ \vdots \\ x_{np} \end{bmatrix}} b_p^{(k)} + \begin{bmatrix} \varepsilon_1^{(k)} \\ \vdots \\ \varepsilon_n^{(k)} \end{bmatrix}$$

with $b^{(k)} = (b_1^{(k)}, \ldots, b_p^{(k)})'$ the p-vector of regression coefficients and $\varepsilon^{(k)} = (\varepsilon_1^{(k)}, \ldots, \varepsilon_p^{(k)})'$ the p-vector of residuals associated to the outcome k.

- The selection of the relevant predictors ($b_j^{(k)} \neq 0$) is achieved through a variable selection procedure

↬ In a Bayesian framework a common approach, called Bayesian variable selection (George and McCulloch, 1993), considers a binary latent indicator vector to perform variable selection such that for the outcome k $\gamma^{(k)} = (\gamma_1^{(k)}, \ldots, \gamma_p^{(k)})'$ with $\gamma_j^{(k)} = 1$ if the predictor j is selected, 0 otherwise.

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

Multivariate models

To increase the statistical power and to consider complex relationships among and within datasets, the simultaneous analysis of the q multiple regressions has been proposed (Brown et al., 1998, 2002):

$$\begin{bmatrix} \text{Gene}_1 & \dots & \text{Gene}_q \\ y_1^{(1)} & \dots & y_1^{(q)} \\ \vdots & & \vdots \\ y_n^{(1)} & \dots & y_n^{(q)} \end{bmatrix} = \begin{bmatrix} \text{miRNA}_1 & \dots & \text{miRNA}_p \\ x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} b_1^{(1)} & \dots & b_1^{(q)} \\ \vdots & & \vdots \\ b_p^{(1)} & \dots & b_p^{(q)} \end{bmatrix} + \begin{bmatrix} \varepsilon_1^{(1)} & \dots & \varepsilon_1^{(q)} \\ \vdots & & \vdots \\ \varepsilon_n^{(1)} & \dots & \varepsilon_n^{(q)} \end{bmatrix}$$

The associated matrix notation is given by:

$$Y = XB + \varepsilon$$

$$vec(\varepsilon) \sim \mathcal{N}(0, \Sigma)$$

with Y a $n \times q$ matrix of outcome variables, X a $n \times p$ matrix of predictors for all outcomes, B a $p \times q$ matrix of regression coefficients, ε a $n \times q$ matrix of residuals, and Σ a $nq \times nq$ covariance matrix.

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

Multivariate models

Different questions raise, especially when there is complex relationships between highly structured datasets:
↬ Bayesian modeling is a flexible framework to handle those questions through prior distributions

- Need to select the same set of predictors for every response ? Need to select different predictors for every response ?

↬ Prior distribution on B or $\Gamma = \{\gamma_j^{(k)}\}_{j=1,\ldots,p;k=1,\ldots,q}$ a binary latent indicator matrix for variable selection (Jia and Xu, 2007; Bottolo et al., 2011)

- Need to assume independence or dependence among responses ? Need to estimate structure among responses?

↬ Prior distribution on Σ (Bhadra and Mallick, 2013)

- How this information may influence the selection ? Need to encourage the selection of correlated predictors? Need to encourage the selection of the same predictors for correlated responses?

↬ Prior distribution on B or Γ (Lee et al., 2017)

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

Models proposed in BayesSUR

BayesSUR considers:

- three prior distributions on Γ for variable selection,
- three prior distributions on the covariance matrix $\Sigma = C \otimes I_n$ with C a $q \times q$ covariance matrix,

$\rightsquigarrow$ Nine models are proposed.

Let $\beta = vec(B), \gamma = vec(\Gamma)$, and β_γ the associated set of non-zero regression coefficients such that

$$\beta_\gamma | \gamma, \omega \sim \mathcal{N}(0, W_\gamma^{-1})$$
$$\omega \sim \mathcal{IG}(a_\omega, b_\omega)$$

with W_γ the sub-matrix of $W = \omega^{-1} 1_{qp}$.

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

Prior distributions for variable selection

↬ Three different priors for Γ:

Independent Bernoulli prior:

$$\gamma_j^{(k)} | \omega_j^{(k)} \sim Ber(\omega_j),$$
$$\omega_j \sim \mathcal{B}eta(a_\omega, b_\omega).$$

where ω_j quantifies the probability for each predictor to be associated with any response variable.

Hotspot prior:

$$\gamma_j^{(k)} | \omega_j^{(k)} \sim Ber(\omega_j^{(k)}),$$
$$\omega_j^{(k)} = o_k \times \rho_j,$$
$$o_k \sim \mathcal{B}eta(a_0, b_0),$$
$$\rho_j \sim \mathcal{G}(a_\rho, b_\rho).$$

where o_k accounts for sparsity of response k, ρ_j the 'propensity' of predictor j to be associated with multiple responses.

Markov Random Field (MRF) prior:

$$f(\gamma | d, e, G) \propto \exp(d 1'\gamma + e\gamma' G\gamma))$$

with G an adjacency matrix containing prior information on relations between predictors and responses ↬ to consider the relationship among predictors + to associate highly correlated responses to the same predictors.

Prior distributions for the covariance matrix

↬ Three different priors for C:

Independent inverse gamma prior:

$$C = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & \dots & \sigma_q^2 \end{bmatrix}$$

with $\sigma_k^2 \sim \mathcal{IG}(a_\sigma, b_\sigma)$.

Inverse Wishart prior:

$$C \sim \mathcal{IW}(\nu, \tau I_q)$$

for a dense covariance matrix.

Hyper-inverse Wishart prior:

$$C \sim \mathcal{HIW}_{\mathcal{G}}(\nu, \tau I_q)$$

assume that multiple response variables have an underlying graph $\mathcal{G}$ encoding the conditional dependence structure between responses, for a sparse covariance matrix.

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

The models

A summary of 9 models implemented in `BayesSUR`:

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

Figure 1: Nine models across three priors of C by three priors of Γ.

with HRR: Hierarchical Related Regression, dSUR: dense Seemingly Unrelated Regression, SSUR: Sparse Seemingly Unrelated Regression.

Inference algorithm: an evolutionary Markov chain Monte Carlo (MCMC) sampler is used.

This project has received funding from The European Union'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.

The main function

The main function BayesSUR

```
> library(BayesSUR)#To download the package
> BayesSUR(data = NULL,# NULL if arguments Y and X are numeric matrices
+          Y = Y,#the response matrix
+          X = X,#the predictor matrix
+          covariancePrior = "HIW",#covariance prior: "IG", "HIW" or "IW"
+          gammaPrior = "hotspot",#gamma prior: "hotspot","MRF" or "hierarchical"
+          nIter = 10000,burnin = 5000,#nb of iterations, and burn-in
+          nChains = 2,#the number of parallel chains in
+          #the evolutionary stochastic search MCMC algorithm
+          outFilePath = "results/",# to specify the path for outputs
+          ...
+ )
```

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

Useful functions

To extract results

- `getEstimator()`: to extract the coefficients matrix $\hat{B}$, latent indicator variable $\hat{\Gamma}$ or learned structure $\hat{\mathcal{G}}$,

To plot results

- `plotEstimator()`: to visualize the three estimators/the relationship of multiple response variables with each other,
- `plotNetwork()`: to visualize the structure relations between multiple response variables and predictors,
- `plotManhattan()`: to show the number of associated response variables of each predictor.

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

Useful functions to check convergence of the MCMC sampler and the model fit

- `plotMCMCdiag()`: to print trace plots and density plots over the MCMC chains,
- `elpd()`, `plot.CPO()`: to estimate the expected log pointwise predicitve density, and to assess out-of-sample prediction accuracy/ to plot the leave-one-out cross-validation predictive density

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.

Data

Using the MOTA data sent by Ziling, a subset of mRNAs and miRNAs from the GU2 cohort were selected by applying a sparse PCA.

↬ A total of 50 genes and 50 miRNAs for 61 patients (37 cases and 24 controls) were analyzed.

From IPA a network analysis has been done for the 50 genes, and the target filter has been applied to get the targets of the 50 miRNAs

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

To download datasets

```
> data <- read.csv("lesson 2/data.csv", sep=";", header = TRUE)
> Y <- scale(as.matrix(data[,1:50]))
> X <- scale(as.matrix(data[,51:100]))
> G_miRNA_IPA <- as.matrix(read.csv("lesson 2/Graph.miRNA.RNA.IPA.csv",
+                          sep=";", header = TRUE))
> G_RNA_IPA <- as.matrix(read.csv("lesson 2/Graph.RNA.IPA.csv",
+                          sep=";", header = TRUE))
```

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

To run BayesSUR

Three different priors distributions for Γ

```
> library(BayesSUR)#To download the package
> # to run BayesSUR with gamma prior "hierarchical"
> fit_G_miRNA_SSURB <- BayesSUR(Y = Y, X = X,
+                           nIter = 50000, burnin = 10000,
+                           gammaPrior = "hierarchical",
+                           gammaInit = "MLE",
+                           outFilePath = "resultsmiRNA_SSURB_sunday/",
+                           output_CPO = TRUE)
> # to run BayesSUR with gamma prior "hotspot"
> fit_G_miRNA_SSURH <- BayesSUR(Y = Y, X = X,
+                           nIter = 50000, burnin = 10000,
+                           gammaInit = "MLE",
+                           outFilePath = "resultsmiRNA_SSURH_sunday/",
+                           output_CPO = TRUE)
> # to run BayesSUR with gamma prior "MRF"
> fit_G_miRNA_SSURMRF <- BayesSUR(Y = Y, X = X,
+                           nIter = 50000, burnin = 10000,
+                           gammaPrior = "MRF",mrfG = G_miRNA_IPA,
+                           gammaInit = "MLE",
+                           outFilePath = "resultsmiRNA_SSURMRF_sunday/",
+                           output_CPO = TRUE)
```

Under grant agreement No 840383.

To run BayesSUR

To compare models

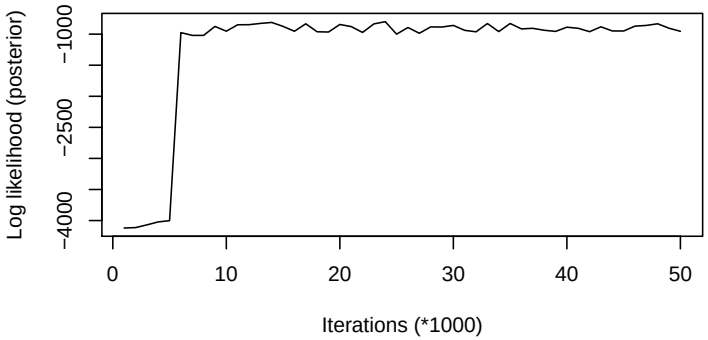
```
> # summary of the analysis
> elpd(fit_G_miRNA_SSURB, method = "waic");elpd(fit_G_miRNA_SSURH, method = "waic");

elpd.waic
-12982.03

elpd.waic
-12917.47

> elpd(fit_G_miRNA_SSURMRF, method = "waic")

elpd.waic
-12810.92
```

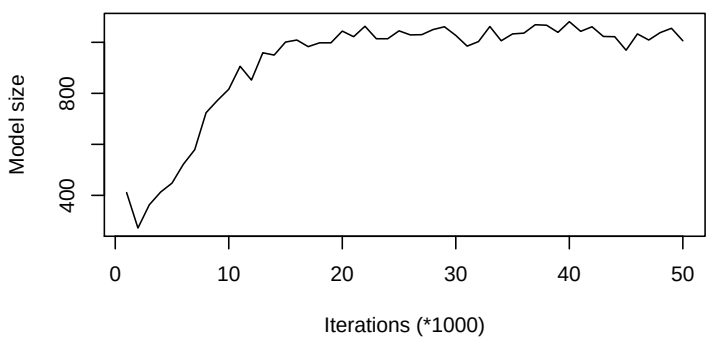
We select the model with the smallest Watanabe–Akaike information criterion (WAIC).

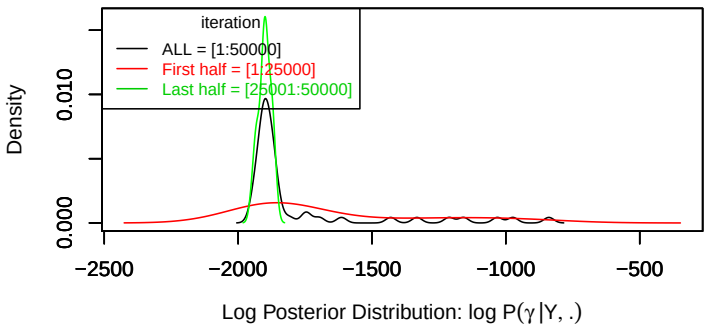
This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

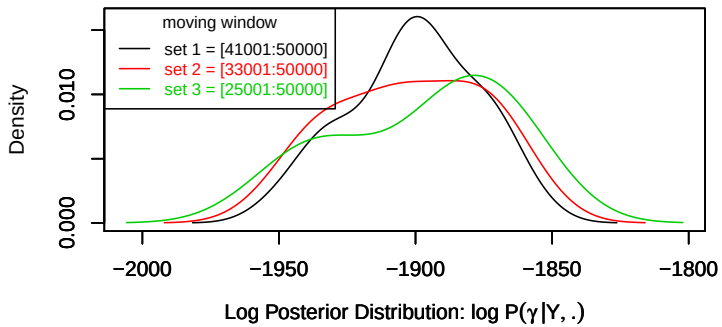
To check the convergence

```
> # summary of the analysis
> plotMCMCdiag(fit_G_miRNA_SSURB)
```









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

To run BayesSUR

To extract estimators

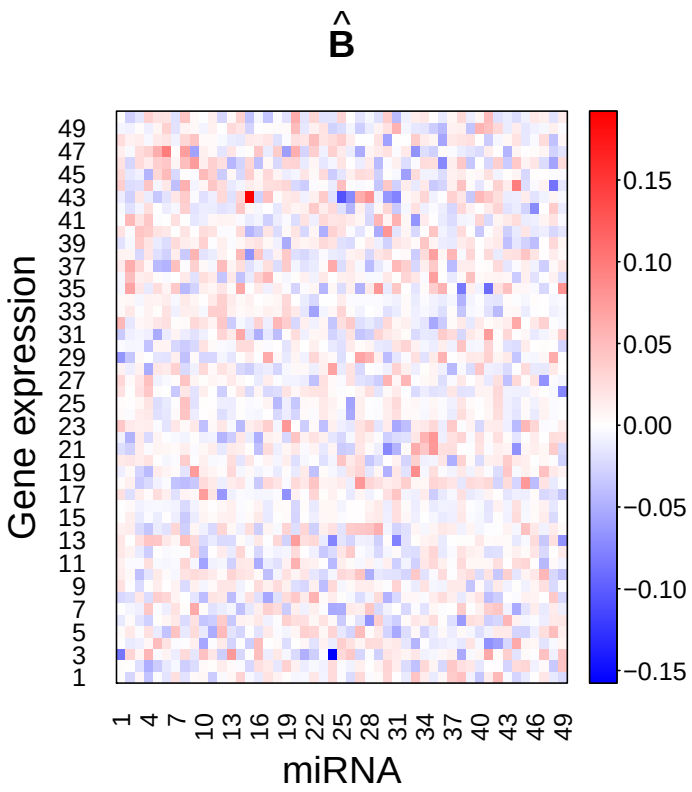
```
> # to get estimators
> gamma.hat <- getEstimator(fit_G_miRNA_SSURB, estimator = "gamma")
> b.hat <- getEstimator(fit_G_miRNA_SSURB, estimator="beta")
> graph.hat <- as.matrix(read.table("resultsmiRNA_SSURB/data_SSUR_Gy_out.txt",
+                                   quote="\"", comment.char=""))
```

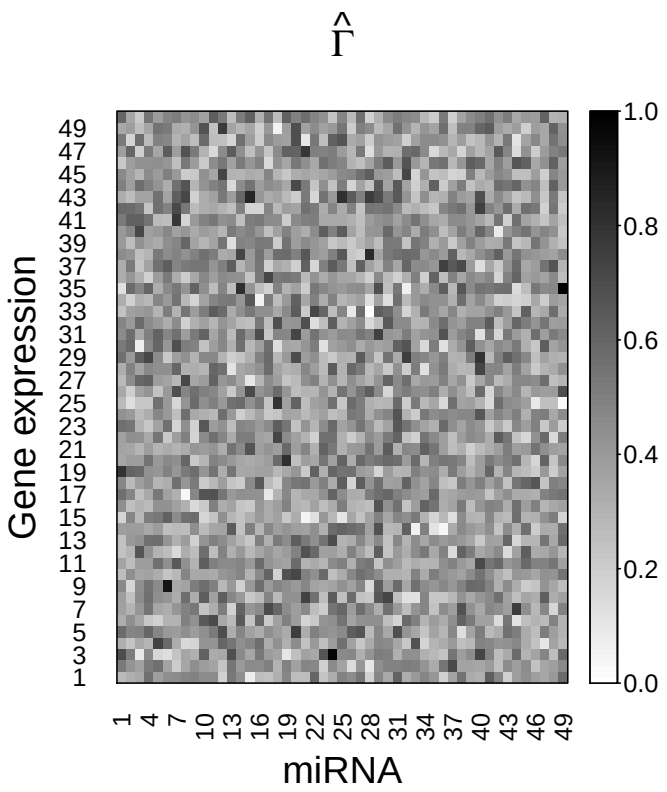
This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

To run BayesSUR

To plot estimators

```
> plotEstimator(fit_G_miRNA_SSURB, estimator = c("beta","gamma"),
+               xlab = "miRNA", ylab = "Gene expression")
```



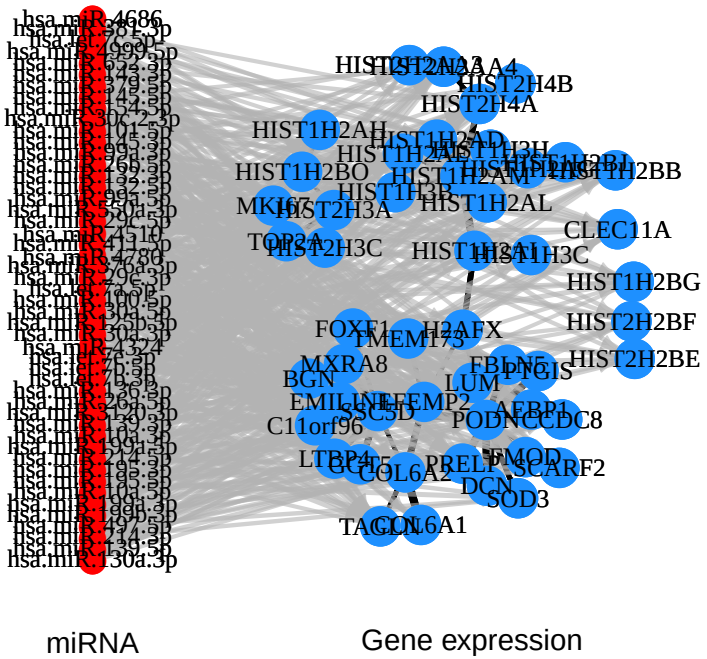


This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

To run BayesSUR

To plot network among genes and miRNAs

```
> plotNetwork(fit_G_miRNA_SSURB, name.predictors = "miRNA",
+             name.responses = "Gene expression",
+             nodesizePredictor = 10,
+             edge.weight = TRUE)
```

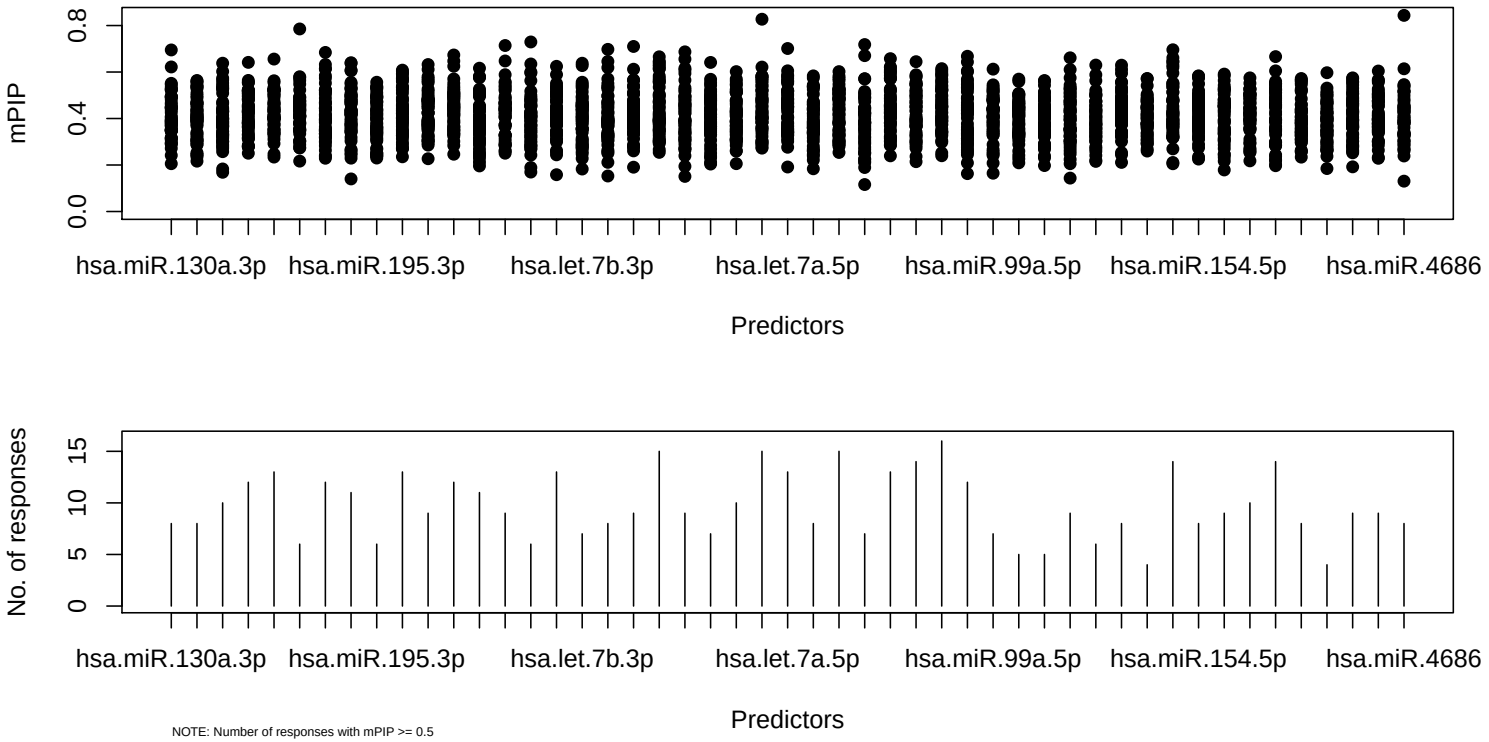


This project has received funding from The European Union's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

To run BayesSUR

To plot number of predictors per outcome

```
> plotManhattan(fit_G_miRNA_SSURB)
```

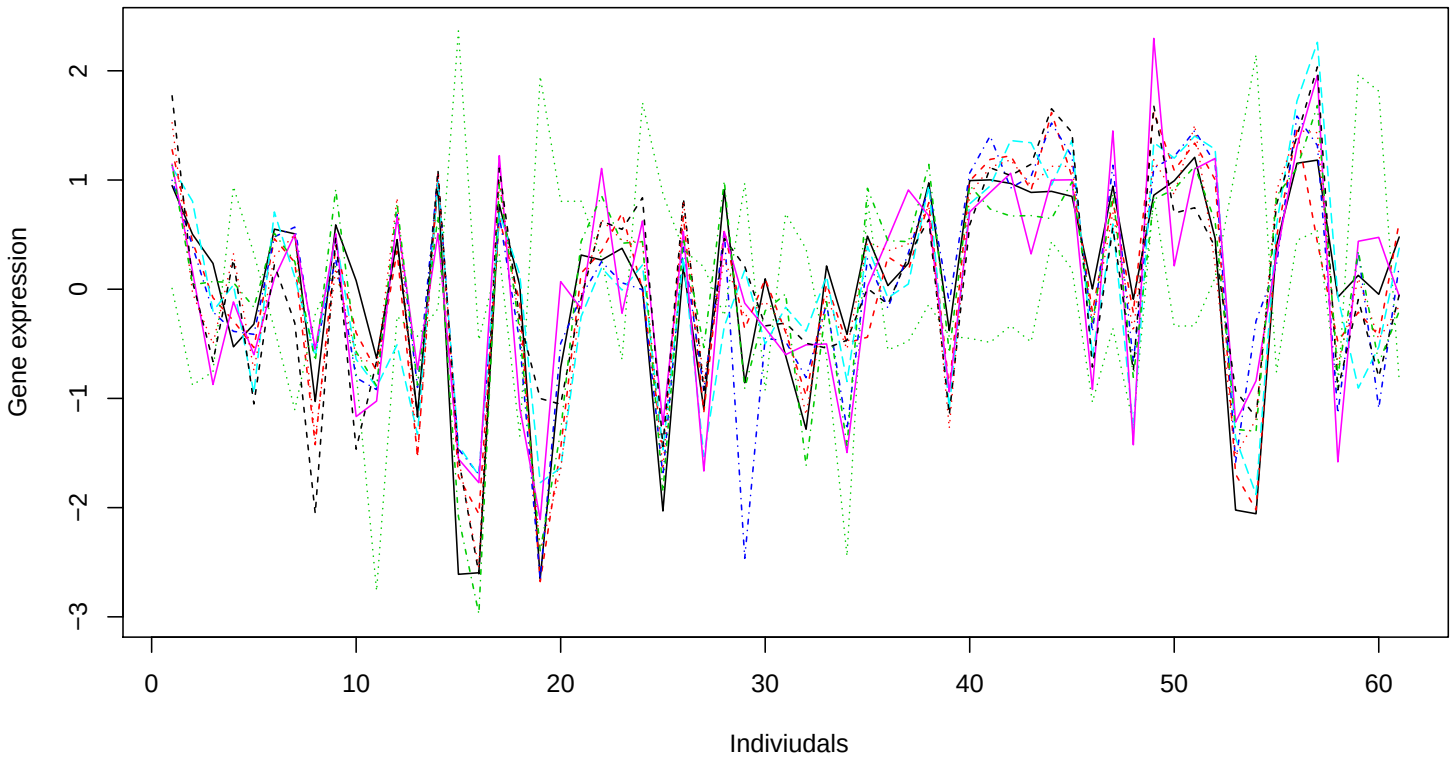


This project has received funding from The European Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

To run BayesSUR

To focus on a predictor associated with multiple outcomes

```
> # gamma.hat < 0.5  is equal to 0
> gamma.hat[gamma.hat > 0.5] <- 1; gamma.hat[gamma.hat <0.5] <- 0
> matplot(Y[,which(gamma.hat[19,]==1)], t="l", xlab = " Indiviudals",
+         ylab = "Gene expression")
```



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

To compare results with biological knowledge

```
> #to obtain incidence matrice
> # for miRNAs/genes
> TP <- sum((G_miRNA_IPA+gamma.hat)==2) ; TP
[1] 6
> TN <-  sum((G_miRNA_IPA+gamma.hat)==0); TN
[1] 1953
> FP <- sum((G_miRNA_IPA == 0 )+ (gamma.hat == 1)); FP
[1] 2894
> FN <- sum((G_miRNA_IPA == 1 )+ (gamma.hat == 0)); FN
[1] 2005
```

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

To compare results with biological knowledge

```
> # to compare with the estimated graph
> graph.hat[graph.hat > 0.5] <- 1; graph.hat[graph.hat <0.5] <- 0
> BDgraph::compare(graph.hat, G_RNA_IPA)

                Target estimate1
true positive       54     1.000
true negative     1171  1160.000
false positive       0    11.000
false negative       0    53.000
F1-score             1     0.030
specificity          1     0.991
sensitivity          1     0.019
MCC                  1     0.019
```

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

Conclusion

- The R package `BayesSUR` allows to implement 9 models taking into account the dependence structure among responses and to consider a priori information on links within and between responses and predictors for variable selection,
- Do not take into account the information estimated by the model for variable selection,
- Method does not seem to scale up for very large data it may work better with a focus on a small subset of markers.

Next training course on statistical methods integrating prior knowledge for variable selection.

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.

Banterle, M., Bottolo, L., Richardson, S., Ala-Korpela, M., Järvelin, M., and Lewin, A. (2018). Sparse variable and covariance selection for high-dimensional seemingly unrelated bayesian regression. *bioRxiv*, page 467019.

Bhadra, A. and Mallick, B. K. (2013). Joint high-dimensional bayesian variable and covariance selection with an application to eqtl analysis. *Biometrics*, 69(2):447–457.

Bottolo, L., Petretto, E., Blankenberg, S., Cambien, F., Cook, S. A., Tiret, L., and Richardson, S. (2011). Bayesian detection of expression quantitative trait loci hot spots. *Genetics*, 189(4):1449–1459.

Brown, P. J., Vannucci, M., and Fearn, T. (1998). Multivariate bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):627–641.

Brown, P. J., Vannucci, M., and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):519–536.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Jia, Z. and Xu, S. (2007). Mapping quantitative trait loci for expression abundance. *Genetics*, 176(1):611–623.

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

Lee, K. H., Tadesse, M. G., Baccarelli, A. A., Schwartz, J., and Coull, B. A. (2017). Multivariate bayesian variable selection exploiting dependence structure among outcomes: application to air pollution effects on dna methylation. *Biometrics*, 73(1):232–241.

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