Global-local shrinkage prior for variable selection in graph-structured models.

Marie Denis and Mahlet G. Tadesse

SSC2022, May 30, 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Context

Dependence between variables may be induced by various factors in different applications:

- Genomic studies: dependence structure between genes obtained from biological pathways or inferred computationally (e.g., based on co-expression),
- Environmental studies: dependence structure between covariates collected over years,
- ...

In many domains high-dimensional data are generated: the number of variables p may be greater than the number of observations n:

- Genomic studies: high-throughput technologies provide genetic/genomic information on the whole genome,
- Environmental studies: high-throughput technologies provide regular and intense monitoring of phenotypic traits over time,
- ...

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Context

Dependence between variables may be induced by various factors in different applications:

- Genomic studies: dependence structure between genes obtained from biological pathways or inferred computationally (e.g., based on co-expression),
- Environmental studies: dependence structure between covariates collected over years,
- ...

In many domains high-dimensional data are generated: the number of variables p may be greater than the number of observations n:

- Genomic studies: high-throughput technologies provide genetic/genomic information on the whole genome,
- Environmental studies: high-throughput technologies provide regular and intense monitoring of phenotypic traits over time,
- ...

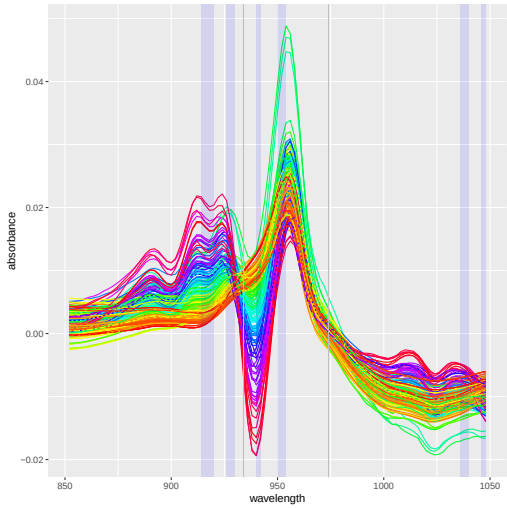
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Context

For example in linear model context:

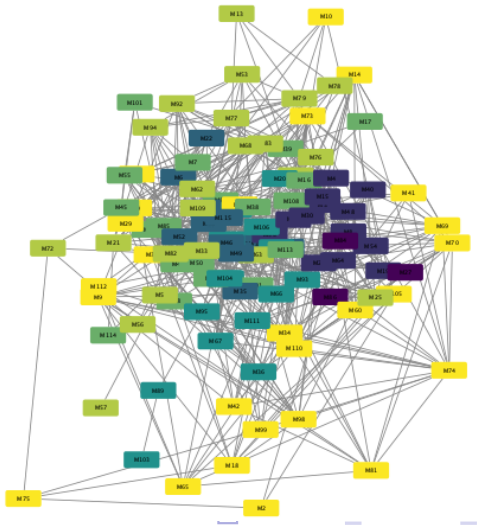
↬ Spectrometric data:

$$\underbrace{\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}}_{\text{Fat content measured on n individuals}} \sim \underbrace{\begin{bmatrix} X_{11} & X_{1p} \\ \vdots & \vdots \\ X_{n1} & X_{np} \end{bmatrix}}_{\text{Spectra sampled at p wavelengths}}$$



↬ Genomic data:

$$\underbrace{\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}}_{\text{Production measured on n individuals}} \sim \underbrace{\begin{bmatrix} X_{11} & X_{1p} \\ \vdots & \vdots \\ X_{n1} & X_{np} \end{bmatrix}}_{\text{Expression values of p genes}}$$



project has received funding from the
opean Union's Horizon 2020 research
innovation programme under the
ie Skłodowska-Curie grant agreement
840383.

Context

Need to use statistical approaches incorporating such structures and dealing with $p >> n$

↬ **Regularization methods**

- To help the model building process by reducing the model complexity of models
- To prevent ill-posed problems (non-invertible matrix, overfitting)
- To lead to parsimonious models

In the following we will focus on **Bayesian approaches**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Bayesian approaches

In Bayesian framework additional information is integrated into models via prior distributions

$\hookrightarrow$ Regularization is done by specifying **specific priors**

Selection

To shrink towards zero small coefficients while leaving large signals large: **Shrinkage** priors

Structure

Priors with a **variance-covariance matrix** related to structure information between variables

Objective

Taking advantage that most of the dependence between variables may be encoded by an undirected graph $\mathcal{G}$

$\hookrightarrow$ To propose **shrinkage** priors integrating **graph structure** information to select graph-structured variables

Shrinkage priors

Two classes of shrinkage priors:

- **Spike-and-slab priors**: Discrete mixture of two distributions (Mitchell and Beauchamp, 1988; ?)

- **Continuous shrinkage priors**: Unimodal continuous distributions (Bayesian Lasso prior, Horseshoe prior, Elastic-Net prior, ...) (Kyung et al., 2010; ?)

↬ the class of global-local priors (Carvalho et al., 2010; Polson and Scott, 2010): a scale mixture of Gaussian distributions with the mixing density depending on two hyperparameters to control the **global** shrinkage and the **local** deviations

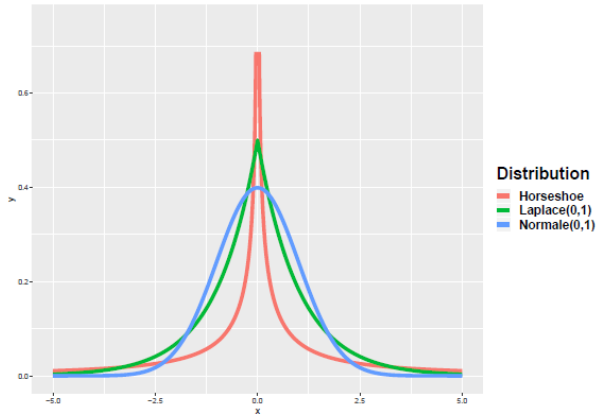
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

The proposed approach

We propose to combine:

- A global-local prior, the horseshoe (HS) prior (Carvalho et al., 2010), for its **efficiency and flexibility** in terms of selection and estimation

↪ allows to shrink towards zero small coefficients while allowing large signals to escape from the overall shrinkage



- With a Gaussian Markov random field (GMRF) for its **appealing connection** with undirected graphs (Rue and Held, 2005)

↪ allows to impose the dependence structure between the parameters via the precision matrix of a conditionally Gaussian prior

↪ An extension of the approach by Faulkner and Minin (2018); Faulkner (2019) to the more general context of graph-structured variable selection.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Bayesian hierarchical model

We assume that $\mathcal{G} = \bigcup_{i=1}^{I} \mathcal{G}_i = \bigcup_{i=1}^{I} (V_i, E_i)$ a disjoint union of I subgraphs and $\mathcal{S}$ the set of indices associated to one representative of each of the I subgraphs.

HS-GMRF model

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \\
\beta_j - s_{jj'}\beta_{j'}|\tau^2_{jj'},\, \lambda^2 &\sim \mathcal{N}(0, \lambda^2\tau^2_{jj'}) \text{ for } (j, j') \in \bigcup_{i=1}^{I} E_i \\
\beta_j|\tau^2_j,\, \lambda^2 &\sim \mathcal{N}(0, \lambda^2\tau^2_j) \text{ for } j \in \mathcal{S} \\
\tau_{jj'} &\sim \mathcal{C}^+(0,1) \text{ for } (j, j') \in \bigcup_{i=1}^{I} E_i;\ \tau_j \sim \mathcal{C}^+(0,1) \text{ for } j \in \mathcal{S} \\
\lambda|\sigma &\sim \mathcal{C}^+(0, \sigma);\ \sigma^2 \sim \mathcal{IG}(a_0, b_0)
\end{aligned}
$$

with $s_{jj'} = \text{sign}\{\text{cor}(X_j, X_{j'})\}$ to encourage regression coefficients of negatively correlated variables to take opposite signs.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Simulation study

Objectives

- To evaluate the performances of the proposed approach with and without incorporating the sign of the sample correlation (HS-GMRF and HS-GMRF-nosign),
- To compare the results with two other approaches: the HS and the spike-and-slab with Ising prior (SS-Ising) (Smith and Fahrmeir, 2007; Li and Zhang, 2010) and when the true graph is known and unknown.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Simulation study

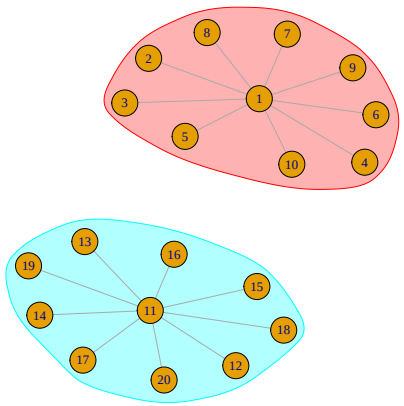
$$Y = \sum_{g=1}^{G} \boldsymbol{X_g} \beta_g + \varepsilon \text{ with } X_{i,g} = (X_{i;g1}, \ldots, X_{i;gk})' \sim \mathcal{N}_k(0, \Sigma_g) \text{ and } \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

12 simulated scenarios

- Two covariance structures
- Two levels of correlation ($\rho = 0.5, 0.9$)
- Three regression coefficients

Simulations

- $G = 14$ groups of $k = 10$ predictors with 5 groups with non-zero effects,
- $\sigma^2 = \sum_{g=1}^{G} \beta_g^2 / 5$
- Repetitions: 50



↬ Focus on the scenario where half of groups with Σ_g and

$$\beta_g = (5, -\frac{5}{\sqrt{10}}, -\frac{5}{\sqrt{10}}, \underbrace{\frac{5}{\sqrt{10}}, \ldots, \frac{5}{\sqrt{10}}}_{k-3})$$

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Simulation study

Performance criteria

- Variable selection criteria:

↬ For HS-based: variable selected if 95% HPD interval does not contain 0,

↬ For SS-Ising: variable selected if marginal inclusion posterior probabilitie greater than 0.5.

- Matthews correlation coefficient (MCC),
- Mean squared error (MSE) of the regression coefficients,
- Mean squared prediction error (MSPE).

MCMC settings:

- Iterations : $6,000$,
- Burn-in: $1,000$.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Results using the true graph

Table 1: Average MCC, MSE and MSPE (with SE) over 50 simulated replications.

		MCC	MSE	MSPE
$\Sigma_{g,\text{half}}$ $\rho = 0.5$	HS-GMRF	**0.708** ($\pm$ 0.018)	**0.513** ($\pm$ 0.067)	**94.871** ($\pm$ 13.632)
	HS-GMRF-nosign	0.624 ($\pm$ 0.034)	0.728 ($\pm$ 0.155)	122.188 ($\pm$ 21.609)
	HS	0.240 ($\pm$ 0.041)	1.009 ($\pm$ 0.200)	126.252 ($\pm$ 19.657)
	SS-Ising	0.323 ($\pm$ 0.054)	1.386 ($\pm$ 0.204)	149.294 ($\pm$ 27.384)
$\Sigma_{g,\text{half}}$ $\rho = 0.9$	HS-GMRF	**0.668** ($\pm$ 0.046)	**0.541** ($\pm$ 0.089)	**84.954** ($\pm$ 14.485)
	HS-GMRF-nosign	0.444 ($\pm$ 0.117)	1.038 ($\pm$ 0.259)	99.123 ($\pm$ 17.694)
	HS	0.219 ($\pm$ 0.038)	2.243 ($\pm$ 0.551)	95.219 ($\pm$ 19.279)
	SS-Ising	0.312 ($\pm$ 0.048)	2.359 ($\pm$ 0.437)	109.387 ($\pm$ 23.713)

- HS-GMRF-based approaches lead to the best results in terms of MCCs, MSEs, and MSPEs,
- HS-GMRF outperforms HS-GMRF-nosign especially when $\rho = 0.9$

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Results using the true graph

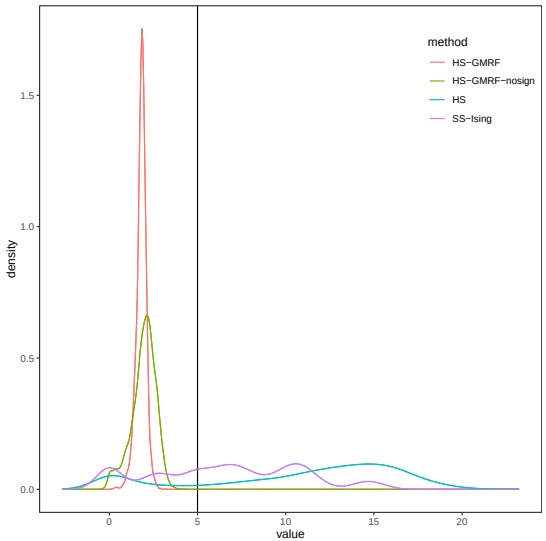
Table 2: Average MCC and MSE for connected and non-connected covariates over 50 simulated replications.

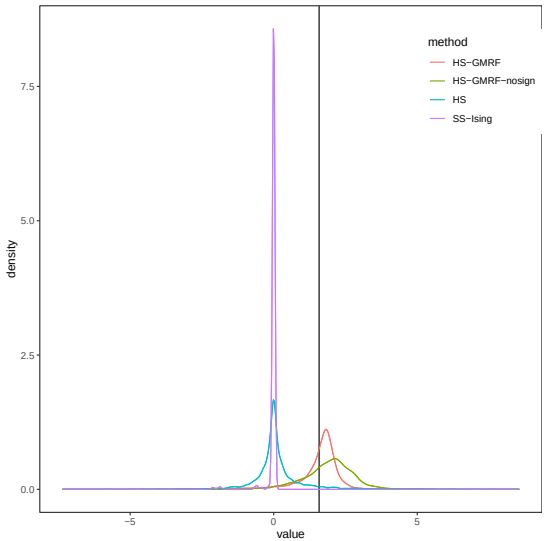
	MCC		MSE	
	Connected	Non-connected	Connected	Non-connected
	$\Sigma_{g,\text{half}}$ $\rho = 0.9$			
HS-GMRF	**0.883** ($\pm$ 0.078)	0.278 ($\pm$ 0.049)	**0.611** ($\pm$ 0.138)	**0.470** ($\pm$ 0.091)
HS-GMRF-nosign	0.526 ($\pm$ 0.177)	0.265 ($\pm$ 0.053)	1.582 ($\pm$ 0.465)	0.495 ($\pm$ 0.112)
HS	0.188 ($\pm$ 0.043)	0.271 ($\pm$ 0.046)	3.998($\pm$ 1.105)	0.488 ($\pm$ 0.103)
SS-Ising	0.310 ($\pm$ 0.047)	**0.304** ($\pm$ 0.081)	4.055 ($\pm$ 0.855)	0.662 ($\pm$ 0.135)

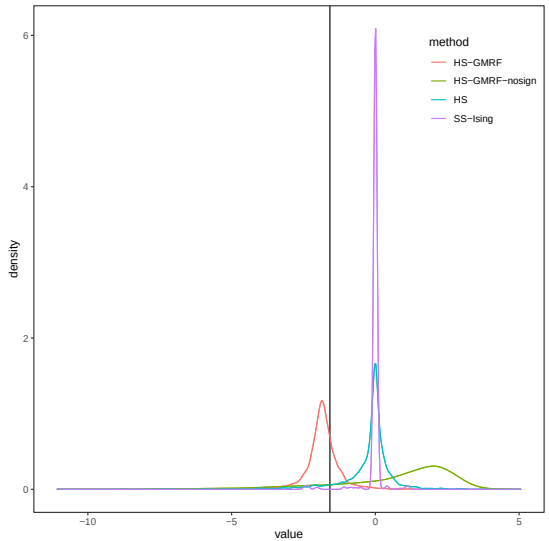
- Performances for non-connected predictors are similar for HS and HS-GMRF-based approaches.
- For connected variables the integration of the dependence structure helps to select variables with small effects

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Results using the true graph



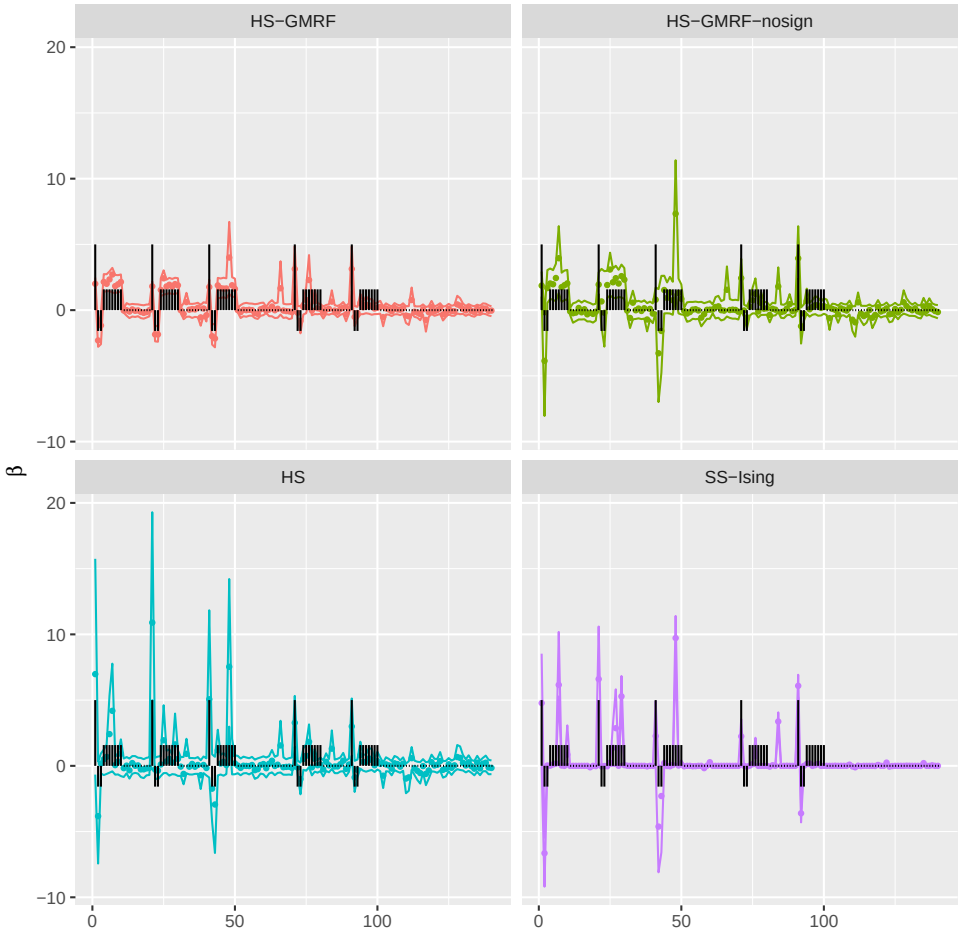




- HS leads to a bimodal posterior density or a distribution concentrated around 0 with large tails
- HS-GMRF-based approaches give narrower posterior densities away from 0
- For $\beta = -5/\sqrt{(10)}$ HS-GMRF-nosign spreads out posterior density around the average of β's

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Results using the true graph



Estimated coefficients along with 80% HPD intervals in one simulated replication ($\Sigma_{g,\text{half}}$, $\rho = 0.9$).

- HS-GMRF-based approaches give similar estimates for highly correlated covariates,
- HS-GMRF yields narrower HPD intervals with good coverage and fairly accurate estimates for regression coefficients with opposite signs.
- HS and SS-Ising tend to select one representative of a group of correlated variables,
- HS gives wide HPD intervals,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Results using an estimated graphs

Graph structure may not be known and may need to be estimated. Graphical Lasso approach used to estimate the graph (Friedman et al., 2008)

- HS-GMRF-based approaches outperform the other approaches,
- For moderate correlation: graph is underestimated $\Rightarrow$ slightly poorer selection and estimation for the HS-GMRF-based approaches than with the true graph,
- For high correlation: graph is overestimated $\Rightarrow$ improved selection for the HS-GMRF-based approaches compared to the true graph but with an overesmoothing of the regressions coefficients

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Shoot growth in *Arabidopsis thaliana*

Objective

To identify molecular markers involved in the rosette compactness phenotype using 358 plants:

- A total of 486 molecular markers along five chromosomes
- Undirected graph: block diagonal matrix where blocks are associated to each chromosome and are tridiagonal matrices
- 5-fold cross-validation procedure

Methods	CV-MSPE	Selected genes
HS-GMRF	1.13	42 (95% HPD) 67 (90% HPD)
HS	1.16	0(95% HPD) 0 (90% HPD)
SS-Ising	1.28	0 (PPI $>$ 0.5) 24 (PPI $>$ 0.2) 121 (PPI $>$ 0.1)
Lasso	1.22	71

↬ HS-GMRF yields to the smallest CV-MSPE

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Shoot growth in *Arabidopsis thaliana*

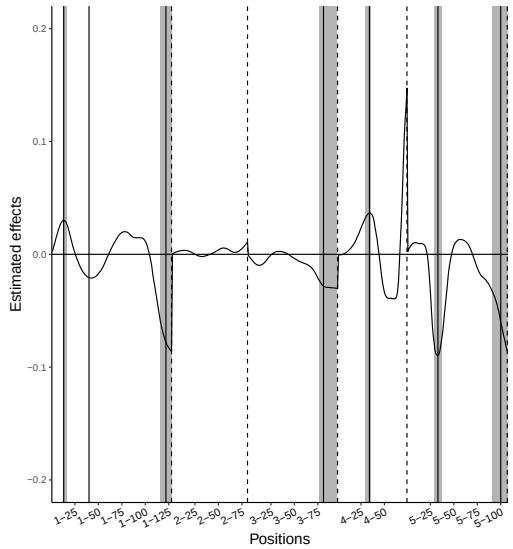


Figure 1: Estimates along with selected genomic regions (shaded areas) using HS-GMRF and QTLs selected by Marchadier et al. (2019) (vertical solid lines). The vertical dashed lines delimit the chromosomes.

- HS-GMRF picks contiguous markers ↬ Selection of genomic regions

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Riboflavin production

Objective

To identify gene expressions involved in the variability of riboflavin production using data on 71 samples

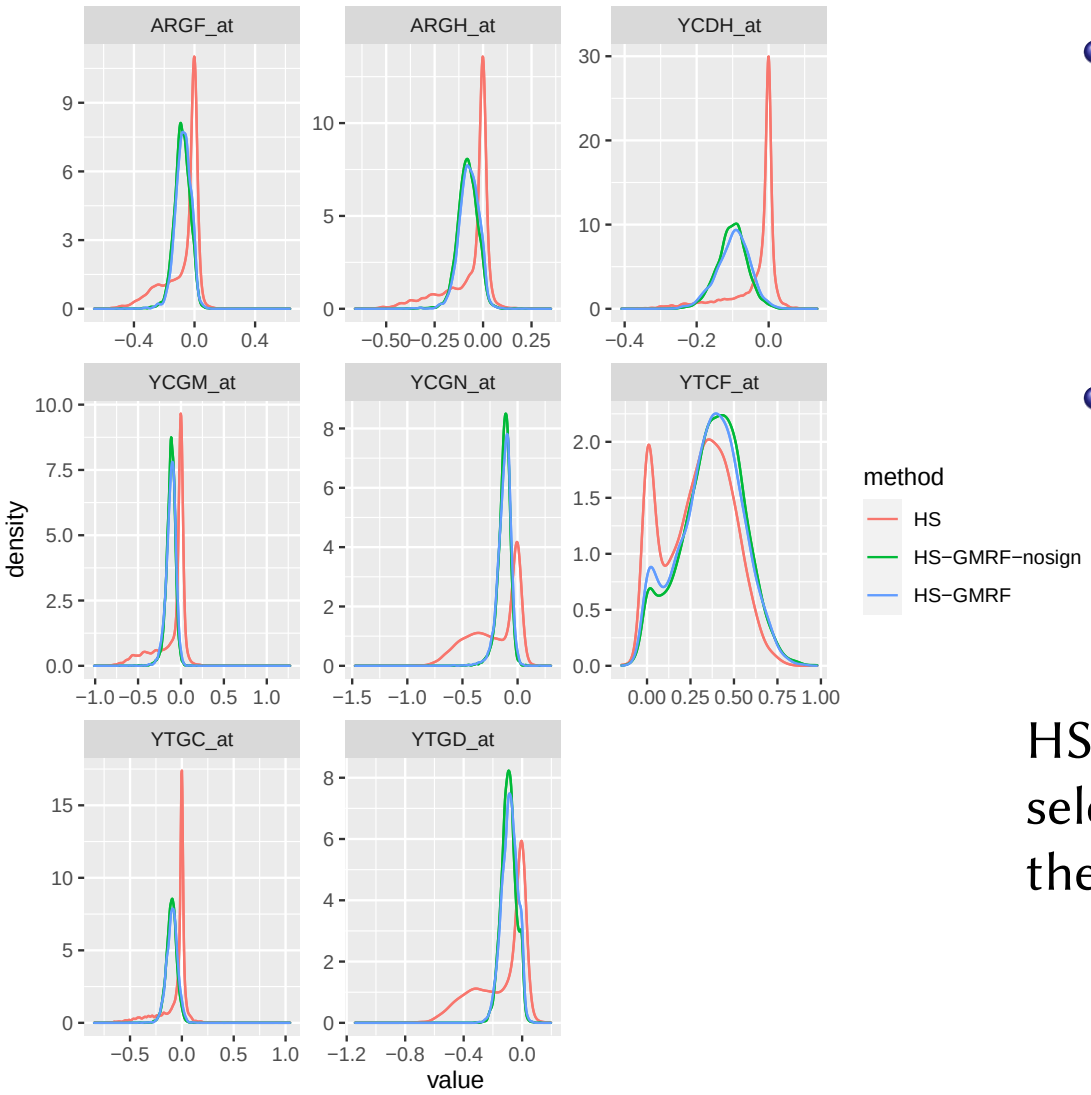
- A total of 142 gene expressions considered
- Estimation of an undirected graph with 157 edges
- 5-fold cross-validation procedure

Methods	CV-MSPE	Selected genes
HS-GMRF	**0.29**	4 (90% HPD) 8 (80% HPD)
HS-GMRF-nosign	0.31	4(90% HPD) 6 (80% HPD)
HS	0.33	0(90% HPD) 0 (80% HPD)
SS-Ising	0.37	21 ($PPI > 0.5$) 16 ($PPI > 0.8$)
Lasso	0.41	16

↬ HS-GMRF yields the smallest CV-MSPE

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Riboflavin productionation



For moderate non-zero effects:

- HS estimates densities concentrated around 0 with long tails or bimodal densities with one of the modes around 0,
- HS-GMRF-based methods estimate unimodal densities or bimodal densities with the mode around 0 less than with HS.

HS-GMRF-based approaches select groups of genes involved in the same biological pathway.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Application

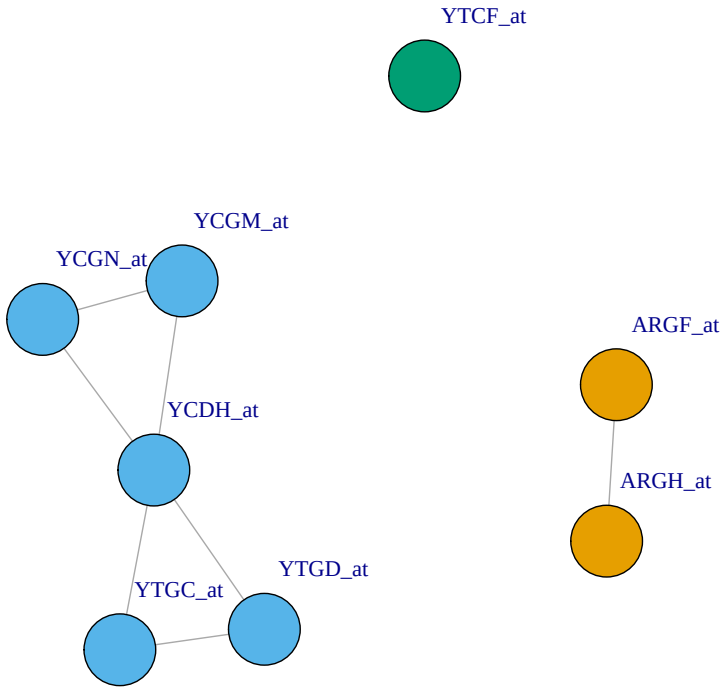


Figure 2: The estimated network

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Conclusion/Perspective

The proposed approaches allow to:

- consider a broad type of dependence structures,
- achieve flexibility in the estimation and the selection due to the local and global shrinkage hyperparameters,
- need to consider the sign of the sample correlation,
- give better predictive performances notably by selecting groups of connected variables,
- give good results even when true graph is unknown and needs to be estimated.

Limitation:

- tend to encourage similar values fo connected variables, especially for highly correlated variables or overestimated graphs.

For future research:

- Extension to non-Gaussian distributions,
- Integration of prior knowledge on strengths of connections between variables.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Thanks for your attention !

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

Faulkner, J. R. (2019). *Adaptive Bayesian Nonparametric Smoothing with Markov Random Fields and Shrinkage Priors*. PhD thesis.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411.

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.

Makalic, E. and Schmidt, D. F. (2016). High-dimensional bayesian regularised regression with the bayesreg package. *arXiv preprint arXiv:1611.06649*.

Martínez-Beneito, M. A. and Botella-Rocamora, P. (2019). *Disease Mapping: From Foundations to Multidimensional Modeling*. CRC Press.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9(501-538):105.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Smith, M. and Fahrmeir, L. (2007). Spatial bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association*, 102(478):417–431.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

MCMC implementation

MCMC implementation

A Gibbs sampling algorithm is straightforward to fit the hierarchical models:

- by using the parametrization of a half-Cauchy as a mixture of inverse-gamma distributions (Makalic and Schmidt, 2016),
- by introducing a q-dimensional vector $\phi = (\phi_1, \ldots, \phi_q)' = C\beta$ (Martínez-Beneito and Botella-Rocamora, 2019) where $q = |E| + |\mathcal{S}|$ and C is a contrast matrix such that:

$$\phi \sim \mathcal{N}_q(0, \Sigma_\phi),$$

with $\Sigma_\phi = \text{diag}(\lambda^2 \boldsymbol{\tau^2})$.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.

Results using the true graph

Table 3: Coverage probability (CP) and width of 95% HPD intervals averaged over the 50 simulated replications.

			CP of 95% HPD	Width of 95% HPD
$\Sigma_{g,\text{half}}$	$\rho = 0.5$	HS-GMRF	0.923 ($\pm$0.026)	2.047 ($\pm$0.188)
		HS-GMRF nosign	0.931 ($\pm$0.027)	2.712 ($\pm$0.231)
		HS	0.894 ($\pm$0.037)	2.871 ($\pm$0.278)
		SS-Ising	0.751 ($\pm$0.026)	0.656 ($\pm$0.117)
	$\rho = 0.9$	HS-GMRF	0.928 ($\pm$0.019)	2.415 ($\pm$0.248)
		HS-GMRF nosign	0.922 ($\pm$0.031)	3.212 ($\pm$0.284)
		HS	0.908 ($\pm$0.05)	3.255 ($\pm$0.419)
		SS-Ising	0.773 ($\pm$0.029)	0.927 ($\pm$0.181)

- CPs similar for HS-based approaches but wider intervals for HS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 840383.