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

Marie Denis and Mahlet G. Tadesse

ENAR, March 17, 2021

cirad
AGRICULTURAL RESEARCH
FOR DEVELOPMENT

GEORGETOWN UNIVERSITY

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

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

Context

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

- Disease mapping: structure in space and time for variables measured over time at adjacent locations,
- Genomic studies: dependence structure between genes obtained from biological pathways or inferred computationally (e.g., based on co-expression),

$\hookrightarrow$ Most of the dependence structures between variables may be encoded by an undirected graph $\mathcal{G}$.

Objective:

To develop a unified Bayesian variable selection for graph-structured variables providing flexibility in the amount of shrinkage and smoothness over the graph.

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

From a statistical point of view

Why incorporate the dependence structure into statistical models?

- helps the model building process,
- increases power to detect associations,
- improves the predictive power.

Why incorporate the dependence structure into variable selection methods?

- It encourages the identification of groups of dependent variables acting jointly on the response, especially those with subtle individual effects.

How to incorporate the dependence structure into variable selection methods?

- Penalized likelihood approaches
- Bayesian regularization via the specification of shrinkage priors.

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

The proposed approach

We propose to extend the approach by Faulkner and Minin (2018); Faulkner (2019) to the more general context of graph-structured variables by combining:

1. The **efficiency and flexibility** of the Horseshoe (HS) (Carvalho et al., 2010) prior (global-local shrinkage prior) in terms of selection and estimation,
2. The **appealing connection** between Gaussian Markov random field (GMRF) and undirected graphs (nonzero elements in the precision matrix correspond to edges in the graph) (Rue and Held, 2005).

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

Bayesian hierarchical models

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

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

Research and innovation programme
Under grant agreement No 840383.

Bayesian hierarchical models

To encourage the regression coefficients estimates to have opposite signs for situations where connected covariates have negative correlation

$\Rightarrow$ We propose to incorporate the sign of the sample correlation such that:

HS-GMRF-sign model

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

with $\eta_{jj'}$ the sign of the correlation between covariates j and j'

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

MCMC implementation

MCMC implementation

A Gibbs sampling algorithm is straightforward used 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 $\phi = (\phi_1, \ldots, \phi_q)' = C\beta$ 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 Unions's Horizon 2020 Research and innovation programme Under grant agreement No 840383.

Simulation studies

Objectives

- To evaluate the performances of the proposed approaches,
- 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).

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

12 simulated scenarios

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

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

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

Simulations:

- $n = 100$, $p = 140$
- $G = 14$ groups of $k = 10$ predictors,
- only the groups $g = 1, 3, 5, 8, 10$ have non-zero effects,
- $\sigma^2 = \sum_{g=1}^{G} \beta_g^2 / 5$
- repetitions: 50

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

Simulation results

Performance criteria

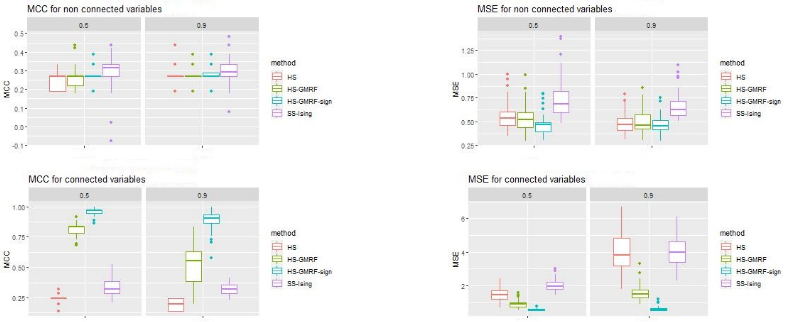
- Variable selection criteria:
  - For HS-based: variable selected if 95% HPD interval does not contain 0,
  - For SS-Ising: variable selected if marginal inclusion posterior probabilitie greater than 0.5.
- Matthews correlation coefficient (MCC),
- Mean squared error (MSE) of the regression coefficients,
- Mean squared prediction error (MSPE).

MCMC settings:

- iterations : 6000,
- burn-in: 1000.

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

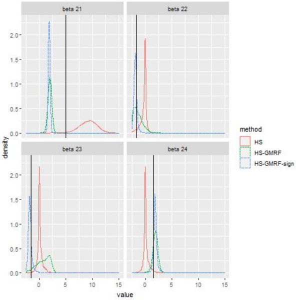
Results

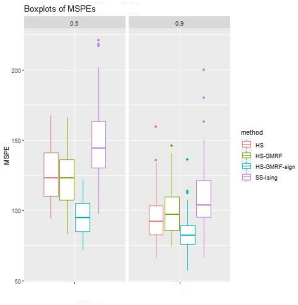


- HS-GMRF-based approaches lead to the best results in terms of MCCs and MSEs with HS-GMRF-sign outperforming the other methods,
- Performances for non-connected predictors are similar for HS and HS-GMRF-based approaches.

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

Results





- Posterior densities of β:
  - HS tends to select one representative of a group of correlated variables,
  - HS-GRMF-based approaches encourage similar values for connected variables,
- For predictive performance when correlated variables of effects of different signs: HS-GMRF-sign gives better results.

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

Results with estimated graphs

In situations where the true graph structure is not known.

- For a moderate level of correlation: graph is underestimated $\Rightarrow$ slightly poorer selection and estimation,
- For a high level of correlation: graph is overestimated $\Rightarrow$ improved selection and estimation.

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

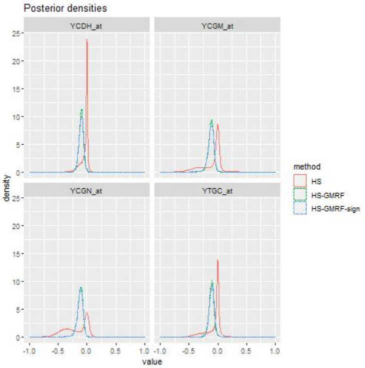
Application

Objective: To identify genes involved in the variability of riboflavin production

Data/Settings

- 71 samples, 142 gene expression
- Undirected graph with 157 edges,
- 5-fold cross-validation procedure.

Methods	MSPEs
HS	0.33
HS-GMRF	0.31
HS-GMRF-sign	0.29
SS-Ising	0.37



↬ selection of groups of correlated variables

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

Outline

1. Bibliography

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.

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.

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*.

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 Unions's Horizon 2020
Research and innovation programme
Under grant agreement No 840383.