Structured variable selection with continuous shrinkage priors

Marie Denis

in collaboration with B. Heuclin, F. Mortier, M. Tadesse, S. Tisné

October 29, 2021

cirad
AGRICULTURAL RESEARCH FOR DEVELOPMENT

GEORGETOWN UNIVERSITY

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

Outline

Outline

Context

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

- Between observations in the response variable:
    - Structure in space and/or time (Dynamic linear model, state-space model, spatio-temporal model,...)
    - Structure induced by grouping factors (Linear mixed model,...)
    - ...
- Between predictors:
    - Dependence structure between genes belonging to the same biological pathways or co-expressed (Gaussian graphical model,...),
    - Dependence structure between covariates collected over years,
    - ...

Such structures need to be taken into account into statistical models

Context

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

- Genetic/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,
- ..

Need to use statistical approaches preventing ill-posed problems (**non-invertible matrix**, **overfitting**) and leading to **parsimonious** models

Context

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

- Between observations in the response variable:
    - Structure in space and/or time (Dynamic linear model, state-space model, spatio-temporal model,...)
    - Structure induced by grouping factors (Linear mixed model,...)
    - ...
- Between predictors:
    - Dependence structure between genes belonging to the same biological pathways or co-expressed (Gaussian graphical model,...),
    - Dependence structure between variables collected over years,
    - ...

Such structures need to be taken into account into statistical models

Linear model context

$$Y = X\beta + \varepsilon,\ \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

with

- $Y = (y_1, \ldots, y_n)'$ the n-vector of observations,
- X the $n \times p$ matrix of predictors which may be structured and/or of high dimension,
- $\beta = (\beta_1, \ldots, \beta_p)'$ the p-vector of coefficients,
- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ the n-vector of residuals,
- σ^2 the residual variance.

$\looparrowright$ To estimate parameters β and σ^2

Classical regression technique

Ordinary Least Square (OLS) regression

To minimize the lost function $L^{OLS}(\beta) = \| Y - X\beta \|^2$:

$$\hat{\beta}^{OLS} = (X'X)^{-1}(X'Y)$$

But

- In presence of structures between predictors (as collinearity): $(X'X)^{-1}$ close to singularity and so, $\hat{\beta}^{OLS}$ not accurate
- When the number of predictors is high: $\hat{\beta}^{OLS}$ does not perform well in unseen datasets (overfitting), does not provide parsimonious models (Hadamard, 1902) and in very high dimension $(X'X)^{-1}$ not invertible

↪ Need to use **regularization methods**

Regularization methods

Consist in introducing additional information into the problem:

- By imposing constraints as in ANOVA,
- By adding a penalty term to the minimization of the loss function as in **penalized regressions** (Ridge (Hoerl and Kennard, 1970), Lasso (Tibshirani, 1996),..),
- By specifying a dependence structure for effects of variables,
- ...

Bayesian approach: a natural framework

Bayesian approaches

In Bayesian framework additional information integrating into models via prior distributions

$\looparrowright$ 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:

To present **shrinkage** priors integrating **structure** information to select structured variables

Outline

Shrinkage priors

Two classes of shrinkage priors:

- **Spike-and-slab priors**: Discrete mixture of two distributions (Mitchell and Beauchamp, 1988; George and McCulloch, 1997)
- **Continuous shrinkage priors**: Unimodal continuous distributions (Bayesian Lasso prior, Horseshoe prior, Elastic-Net prior, ...) (Kyung et al., 2010; Carvalho et al., 2008)

Spike-and-slab prior

- Introduction of γ:

$$\gamma_j = \begin{cases} 1 \text{ if variable } j \text{ is selected} \\ 0 \quad \text{otherwise} \end{cases}$$

$$\beta_j|(\gamma_j = 1) \sim p_{Slab}(\beta_j)$$

$$\beta_j|(\gamma_j = 0) \sim p_{Spike}(\beta_j)$$

- The estimation of $\mathbb{P}(\gamma_j = 1|Y)$ gives access to the a posteriori probability of variable selection

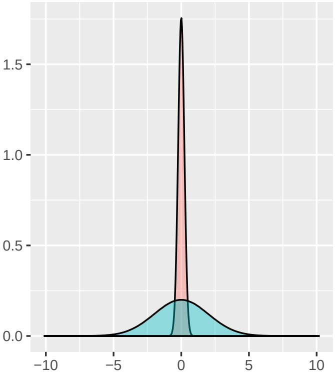


Figure 1: Spike-and-Slab prior distribution. Slab part in blue and spike part in red

Continuous shrinkage prior

Bayesian version of penalized approaches:

$$\beta_j | \tau^2, \omega_j \sim \mathcal{N}(0, \tau^2 \omega_j^2) \;\; j = 1, \ldots, p$$

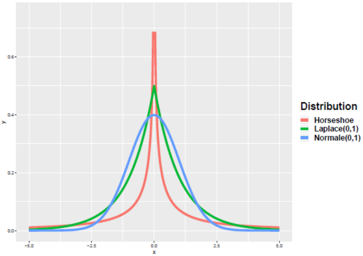
$$\tau^2, \omega_j^2 \sim \mathcal{F}(\tau^2; \omega_j^2)$$

where

- $\mathcal{F}$ is a distribution to specify

  $\looparrowright$ Bayesian Lasso prior, **Global-local priors**, ...

Figure 2: Continuous shrinkage prior distributions



Horseshoe prior (Carvalho et al., 2009)

A global-local prior with $\tau \sim \mathcal{C}^+(0,1)$ and $\omega_j \sim \mathcal{C}^+(0,1)$ $j = 1, \ldots, p$

- τ controls the global shrinkage
- ω_j controls the individual shrinkage: allows large signals to escape from the overall shrinkage

Outline

Structure

$$Y = X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon,\ \varepsilon \sim \mathcal{N}_n(0, \sigma^2)$$

- X the $n \times p$ matrix of predictors which may be structured and/or of high dimension,

$\looparrowright$ $\beta = (\beta_1, \ldots, \beta_p)' \sim \mathcal{N}_p(0, \Sigma)$ with Σ related to structure between variables

$\looparrowright$ **Context dependent**

Examples

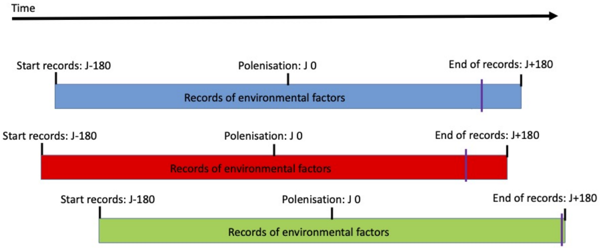
- $X_{t-1} \not\perp\!\!\!\perp X_t$: $\Sigma = AR(\rho)$ with ρ autoregressive parameter
- X_i's belong to a same group (pathways in genomic, genes in genetic, ...)

Outline

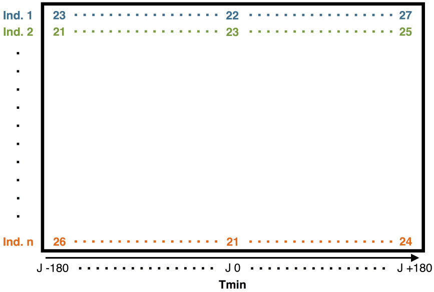
Biological motivation

Objective: to understand the impact of environmental variables on the process of fruit abscission in oil palm.
Dataset provided by "le Centre de Recherches Agricoles-Plantes Pérennes (CRA-PP)" (Tisné et al., 2020)

Oil palm: fruit abscission process

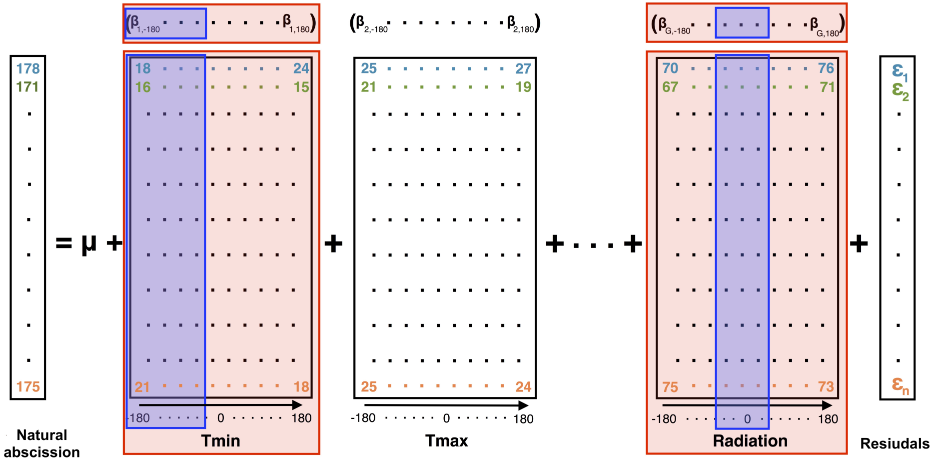


- **Climatic factors:**
  Temperature, rainfall, ...
- **Ecophysiological factors:**
  Evapotranspiration, photosynthesis, ...



Oil palm: fruit abscission process

To identify the environmental variables and the time periods affecting the oil palm fruit abscission process



Statistical questions

Linear model:

$$\boldsymbol{y} = \mu + \sum_{g=1}^{G} \boldsymbol{X_g} \boldsymbol{\beta_g} + \varepsilon,\ \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$$

Selection:

- To identify environmental variables $\rightarrow$ Selection of grouped variables
  $\hookrightarrow$ Selection of grouped parameters: $(\beta_{g,-180}, \ldots, \beta_{g,180})' = (0, \ldots, 0)'$?
- To identify time periods $\rightarrow$ Selection of variables
  $\hookrightarrow$ Selection of parameters $\beta_{g,t} = 0$?

Double structure:

- Grouped variables
  $\looparrowright$ Integration of group structure
- Repeated measures over time for each environmental variables
  $\looparrowright$ High correlation between successive variables: to integrate natural order of variables on regression coefficients $\boldsymbol{\beta}_g = (\beta_{g,-180}, \ldots, \beta_{g,180})'$

Group fused horseshoe prior

To combine the horseshoe and the fused (Tibshirani et al., 2005) priors.

$$\boldsymbol{\beta_g} | \tau^2, \lambda_g^2, \boldsymbol{\omega_g^2}, \boldsymbol{\upsilon_g^2}, \sigma^2 \sim \mathcal{N}_T\left(0, \sigma^2 \left(\frac{\boldsymbol{D_g'}\boldsymbol{\Omega_g^{-1}}\boldsymbol{D_g}}{\tau^2\lambda_g^2} + \boldsymbol{\Upsilon_g^{-1}}\right)^{-1}\right)$$

$$\tau,\ \lambda_g,\ \omega_{g,t},\ \upsilon_{g,t} \sim \mathcal{C}^+(0,1)$$

1st matrix: to penalize differences while allowing abrupt changes

- τ^2: global hyperparameter
- λ_g^2: local hyperparameters specific to each group (environmental variables)
- $\omega_{g,t}^2$: local hyperparameter specific to difference $\beta_{g,t+1} - \beta_{g,t}$

2nd matrix: to add an additional level of penalization to shrink towards zero small effects

- $\upsilon_{g,t}^2$ local hyperparameter specific to the effect $\beta_{g,t}$

$\hookrightarrow$ Extension of priors: *sparse group horseshoe* (Xu et al., 2016), *fusion horseshoe* (Faulkner and Minin, 2018), *fused Laplace* (Kyung et al., 2010)

Inference: Gibbs algorithm (Markov chain Monte Carlo method). R code available at: https://github.com/Heuclin/GroupFusedHorseshoe

Simulation study

- Different profiles over time: smooth or with abrupt changes
- Comparisons with usual approaches:

↬ Sparse PLS, Elastic-net, Penalized regression with composite MCP penalty, Penalized regression with composite MCP + ridge

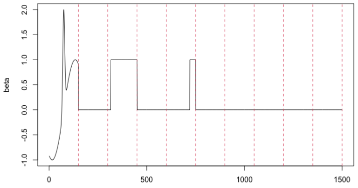


Figure 4: Simulated profiles for G=10 groups over T=150 time points

Results

- Group fused horseshoe outperforms other approaches in terms of selection, estimation, and prediction

↬ Robust: no sensitive to the ratio sample size / number of variables

↬ Stable: low variability between repetitions

Application on oil palm

Data

- 1,173 bunches (statistical unit)
- Outcome (y): number of days from pollination to fruit drop
- 5 climatic variables: Tmax, Tmin, Relative air humidity (RH), Rainfall (R), Solar radiation (SR)
- 5 ecophysiological variables: Maximum daily vapor pressure deficit (VPD), Fraction of transpirable soil water (FTSW), Supply-demand ratio (SD), Daily reproductive demand (DRD)
- 121 time points for each environmental variables: $p = 1,210$ predictors greater than $n = 1,173$ observations

Application on oil palm

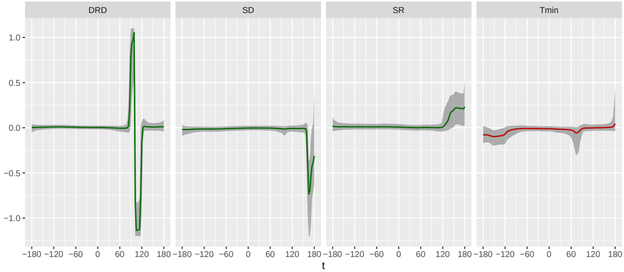


Figure 5: Estimated coefficient profiles for DRD, SD, SR, Tmin. Gray shadows represent the 95% credible interval.

- Identification of 4 environmental variables: DRD, SD, SR, Tmin
- Identification of relevant time periods
  - Tmin: smooth effect during the inflorescence development
  - DRD and SD: punctual effects at the end of the fruit bunch development
  - SR: smooth effect at the end of the fruit bunch development

Outline

Motivations

Most of the dependence structures between variables may be encoded by an undirected graph
↬ We propose to extend the approach proposed 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) prior in terms of selection and estimation:
2. 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 thus leading to sparse matrices and to smooth coefficients over the graph with possible abrupt changes

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

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

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

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.

$Y = \sum_{g=1}^{G} \boldsymbol{X_g}\beta_g + \varepsilon$ with $X_{i,g} = (X_{i;g1}, \ldots, X_{i;gk})' \sim \mathcal{N}_k(0, \Sigma_g)$ 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 half of groups with Σ_g and

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

Simulations

- $G = 14$ groups of $k = 10$ predictors,
- Only groups $g = 1, 3, 5, 8, 10$ have non-zero effects,
- $\sigma^2 = \sum_{g=1}^{G} \beta_g^2 / 5$
- Repetitions: 50

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).

Results

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 $\rho = 0.9$

Results

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

	MCC		MSE	
	Connected	Non-connected	Connected	Non-connected
	$\Sigma_{g,\text{half}}$ $\rho = 0.5$			
HS-GMRF	**0.956** ($\pm$ 0.033)	0.277 ($\pm$ 0.039)	**0.558** ($\pm$ 0.061)	**0.469** ($\pm$ 0.111)
HS-GMRF-nosign	0.810 ($\pm$ 0.053)	0.264 ($\pm$ 0.057)	0.913 ($\pm$ 0.202)	0.542 ($\pm$ 0.151)
HS	0.237 ($\pm$ 0.038)	0.244 ($\pm$ 0.054)	1.464 ($\pm$ 0.374)	0.553 ($\pm$ 0.139)
SS-Ising	0.332 ($\pm$ 0.062)	**0.295** ($\pm$ 0.096)	2.028 ($\pm$ 0.372)	0.744 ($\pm$ 0.208)
	$\Sigma_{g,\text{half}}$ $\rho = 0.9$			
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

Results

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
$\Sigma_{g,\mathrm{half}}$	$\rho = 0.5$	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

Results

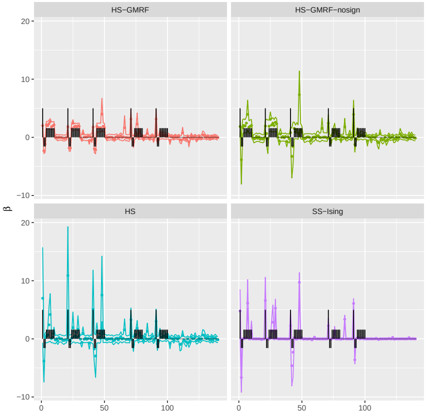


Figure 6: Estimated coefficients along with 80% HPD intervals in one simulated replication ($\Sigma_{g,\text{half}}$, $\rho = 0.9$).

- HS and SS-Ising tend to select one representative of a group of correlated variables,
- HS gives wide HPD intervals,
- 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.

Outline

Conclusion/Perspectives

Conclusion

Two priors proposed to select structured variables:

- Efficient combination and extension of existing approaches to deal with high-dimensional structured predictors,
- Flexibility and efficiency of horseshoe prior, via the local and global shrinkage hyperparameters, to handle different structures and to select relevant variables and/or groups of variables,
- Encompass a broad type of dependence structures : applicable in various applications (varying coefficient models, near infrared spectroscopy context (NIRS), QTL mapping, ...)

Perspectives

- To integrate prior knowledge on strengths of connections,
- To integrate dependence structures between observations,
- To extend to multivariate case (Y multivariate),
- To consider a multi-dimensional indexation.

Outline

1. Introduction
   - Context
   - Selection
   - Structure
2. Group fused horseshoe prior
3. Horseshoe Gaussian Markov field prior
4. Conclusion/Perspectives
5. Bibliography

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2008). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling Sparsity via the Horseshoe. page 8.

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.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55.

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.

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

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.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tisné, S., Denis, M., Domonhédo, H., Pallas, B., Cazemajor, M., Tranbarger, T. J., and Morcillo, F. (2020). Environmental and trophic determinism of fruit abscission and outlook with climate change in tropical regions. *Plant-Environment Interactions*, 1(1):17–28.

Xu, Z., Schmidt, D. F., Makalic, E., Qian, G., and Hopper, J. L. (2016). Bayesian grouped horseshoe regression with application to additive models. In *Australasian Joint Conference on Artificial Intelligence*, pages 229–240. Springer.