Bayesian Penalty Mixing with the Spike-and-Slab LASSO

Veronika Ročková[1] and Ed George[2]

Symposium on Data Science and Statistics in
Honor of Edward J. Wegman
Reston, Virginia
May 18, 2018

[1]Chicago Booth
[2]Wharton, UPenn

Introducing the Spike-and-Slab LASSO

For known $\boldsymbol{X}_{n\times p}$ with standardized columns $||\boldsymbol{X}_j||^2 = n$, suppose

$$\boldsymbol{Y} = \boldsymbol{X}_{n\times p}\,\boldsymbol{\beta}_0 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \mathrm{I}_n),$$

where $||\boldsymbol{\beta}_0||_0 = q$, p large, $q << p$. Goal is the recovery of $\boldsymbol{\beta}_0$.

$\rightsquigarrow$ *Popular Bayesian approach:* **Point-Mass Spike-and-Slab Prior**

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) = \textstyle\prod_{i=1}^{p}[\gamma_i \phi(\beta_i \mid \lambda) + (1-\gamma_i)\delta_0(\beta_i)],$$

$$\phi(\beta_i \mid \lambda) \equiv \frac{\lambda}{2} e^{-\lambda|\beta_i|}, \qquad \gamma_1, \ldots, \gamma_p \mid \theta \ \ \textit{iid} \sim \texttt{Bern}(\theta), \quad \theta \sim \pi(\theta)$$

- Ideal posterior concentration
- MCMC posterior simulation slow for p large

$\rightsquigarrow$ *Popular penalized-likelihood approach:* **LASSO**

$$\pi(\boldsymbol{\beta} \mid \lambda) = \textstyle\prod_{i=1}^{p} \phi(\beta_i \mid \lambda)$$

- Fabulously fast identification of the mode
- Problematic bias and posterior issues

The Essence of the LASSO

- Select the "best" LASSO estimator of the form

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ -\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 - \lambda \sum_{i=1}^{p} |\beta_i| \right\}$$

  for an increasing sequence of λ values.

- Each $\widehat{\boldsymbol{\beta}}$ is a posterior mode under $\pi(\boldsymbol{\beta} \mid \lambda) = \prod_{i=1}^{p} \phi(\beta_i \mid \lambda)$, an iid prior.

- As λ increases, all $\hat{\beta}_i$'s are uniformly shrunk more, and larger subsets of $\hat{\beta}_i$'s are thresholded to zero.

- As $\lambda \to \infty$, $\pi(\boldsymbol{\beta} \mid \lambda) \to \delta_{\mathbf{0}}(\boldsymbol{\beta})$, a point mass at $\mathbf{0}$.

- *Dynamic Posterior Exploration* is made practical by fast convex optimization.

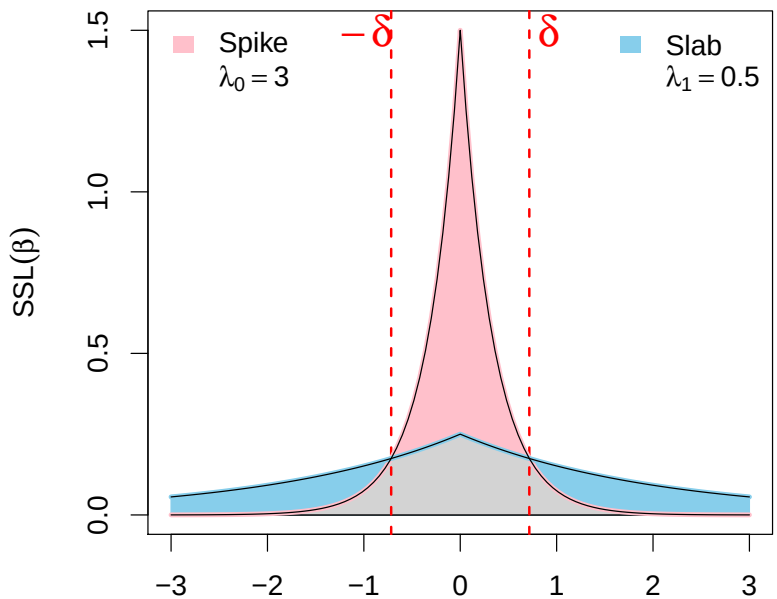
Hybrid Idea: The Spike-and-Slab LASSO Prior

A mixture of two LASSO priors with penalties λ_1 and λ_0

$$\pi_{SSL}(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) = \prod_{i=1}^{p} [\gamma_i \phi(\beta_i \mid \lambda_1) + (1 - \gamma_i)\phi(\beta_i \mid \lambda_0)]$$

$$\gamma_1, \ldots, \gamma_p \mid \theta \ \ \textit{iid} \sim \texttt{Bern}(\theta), \quad \theta \sim \pi(\theta)$$

- λ_1 **small**: slab distribution holds large coefficients steady
- λ_0 **large**: spike distribution thresholds small coefficients
- θ controls the sparsity



The Essence of the Spike-and-Slab LASSO

- Select the "best" *SSL* (Spike-and-Slab LASSO) estimator of the form

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ -\frac{1}{2} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \log \pi_{SSL}(\boldsymbol{\beta}) \right\}$$

  for an increasing sequence of λ_0 values, with λ_1 fixed at a small value.

- Each $\widehat{\boldsymbol{\beta}}$ is a posterior mode under the $\pi_{SSL}(\boldsymbol{\beta})$ mixture prior.

- As λ_0 increases, small $\hat{\beta}_i$'s are thresholded to zero by the "spike" while large ones are held steady by the "slab".

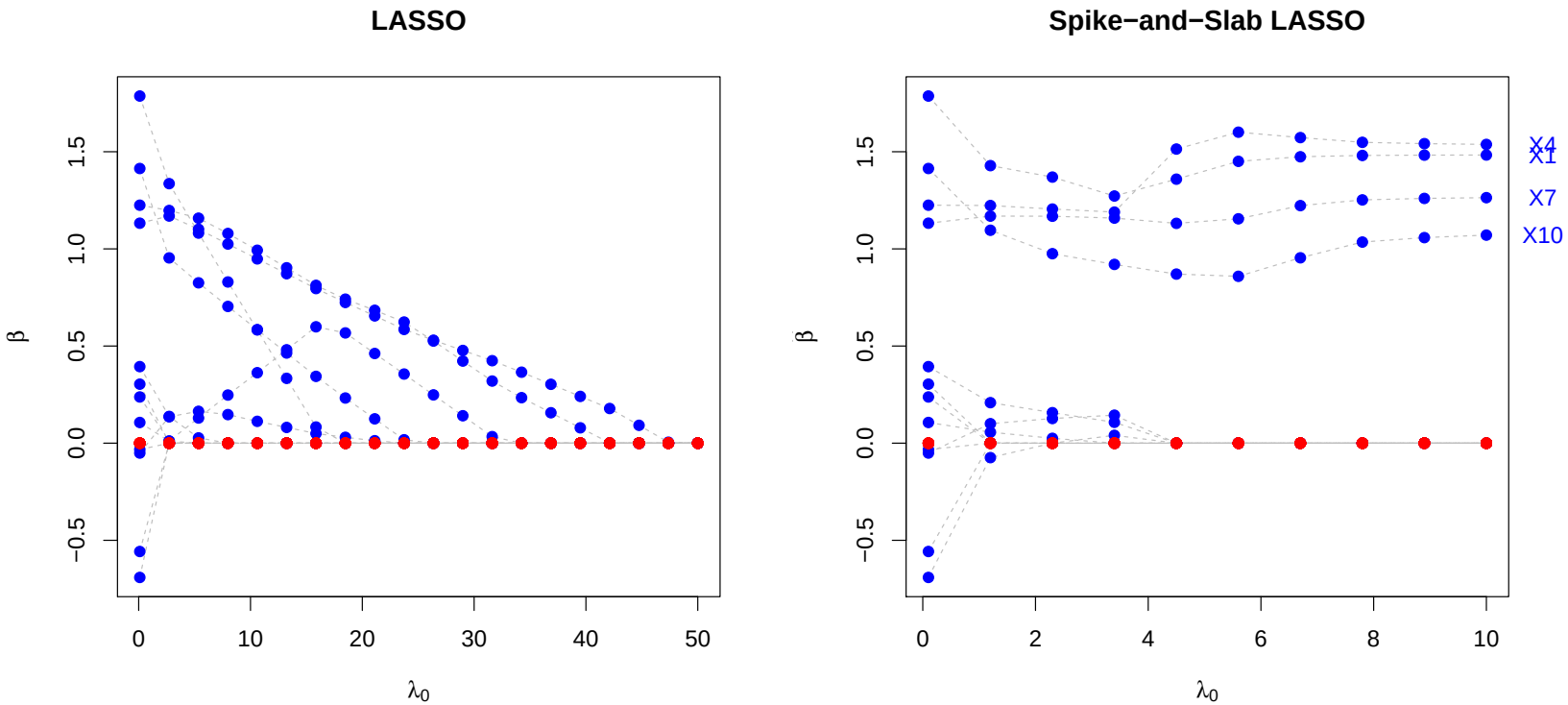
- Simultaneous *variable selection* and (nearly) *unbiased estimation*, occurring directly in the $\boldsymbol{\beta}$ space.

- As $\lambda_0 \to \infty$, $\pi_{SSL}(\boldsymbol{\beta})$ goes to the point mass spike-and-slab prior.

- *Dynamic Posterior Exploration* is made practical by fast non-convex optimization.

The LASSO and the Spike-and-Slab LASSO in Action

- Consider $p = 12$ and $n = 50$
- 4 indep groups of correlated ($\rho_{ij} = 0.9$) predictors $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$
- $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\beta_0, I_n)$, where $\beta_0 = (\underbrace{1.3, 0, 0}, \underbrace{1.3, 0, 0}, \underbrace{1.3, 0, 0}, \underbrace{1.3, 0, 0})'$.



- *LASSO never correct. After cross validation, 4 false positives.*
- *Spike-and-Slab LASSO path stabilizes at the correct model.*

What is new?

Other non-convex regularizers such as MCP and SCAD serve to mitigate the bias of the LASSO. However, in comparison:

(1) *Spike-and-Slab LASSO is a* **hierarchical Bayes procedure**

- $\rightsquigarrow$ Underlying latent model indicators $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$
- $\rightsquigarrow$ $\pi(\boldsymbol{\gamma})$ can be used to target regions of interest

(2) *Spike-and-Slab LASSO penalty is* **non-separable**

- $\rightsquigarrow$ θ adapts to the unknown sparsity of $\boldsymbol{\beta}_0$
- $\rightsquigarrow$ Automatic hyper-parameter tuning (avoids cross-validation)
- $\rightsquigarrow$ Automatic adjustment for multiplicity
- $\rightsquigarrow$ Coordinate ascent for a non-separable regularizer

The Separable *SSL* Penalty

(When θ is fixed)

Focusing First on $\pi_{SSL}(\boldsymbol{\beta} \mid \theta)$

- Recall the full SSL prior

$$\pi_{SSL}(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) = \prod_{i=1}^{p} [\gamma_i \phi(\beta_i \mid \lambda_1) + (1 - \gamma_i)\phi(\beta_i \mid \lambda_0)]$$

$$\gamma_1, \ldots, \gamma_p \mid \theta \overset{iid}{\sim} \texttt{Bern}(\theta), \quad \theta \sim \pi(\theta)$$

- This prior is a mixture of Laplace priors both within and across the coordinates of $\boldsymbol{\beta}$.
- Such **Bayesian penalty mixing** yields penalization that adaptively tailors shrinkage effects to the underlying $\boldsymbol{\beta}_0$
- To better understand this, let's integrate out the γ_i's, and first focus on $\pi_{SSL}(\boldsymbol{\beta} \mid \theta)$, treating θ as if it were fixed and known.

The Separable *SSL* Penalty

- The conditional *SSL* prior is an independent product

$$\pi_{SSL}(\boldsymbol{\beta} \mid \theta) = \prod_{i=1}^{p} [\theta \phi(\beta_i \mid \lambda_1) + (1 - \theta)\phi(\beta_i \mid \lambda_0)]$$

- Here, the latent γ_i indicators have been margined out.

- The conditional *SSL* estimator is the solution to

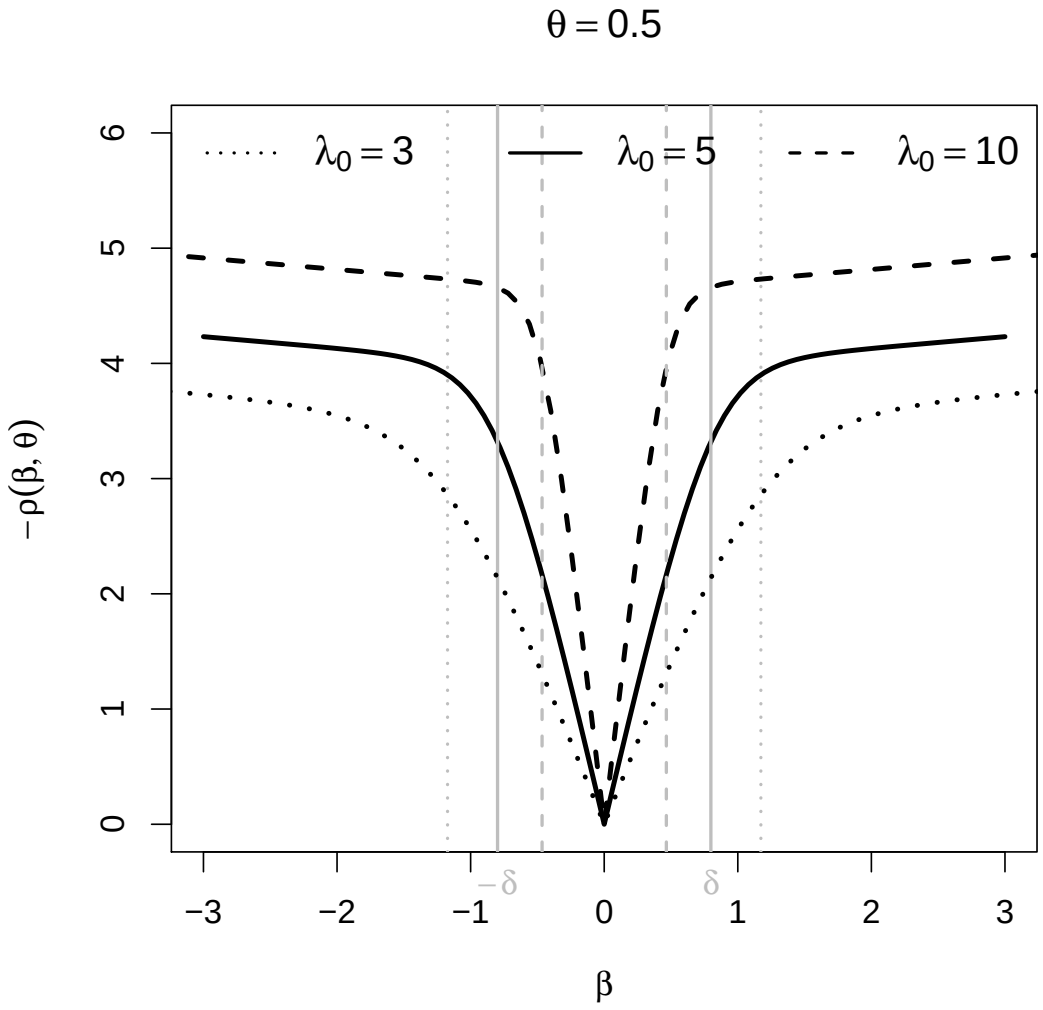
$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ -\frac{1}{2} ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \log \pi_{SSL}(\boldsymbol{\beta} \mid \theta) \right\}$$

- This *SSL* penalty is a **separable** sum of component penalties

$$\log \pi_{SSL}(\boldsymbol{\beta} \mid \theta) = \sum_{i=1}^{p} \log[\theta \phi(\beta_i \mid \lambda_1) + (1 - \theta)\phi(\beta_i \mid \lambda_0)]$$

The Separable *SSL* Penalty

- Each component of the *SSL* penalty is a smooth mix of two LASSO-like penalties



The Adaptive Effect of Bayesian Penalty Mixing

Via the first order conditions for $\widehat{\boldsymbol{\beta}}$, the derivative of the penalty determines the amount of shrinkage,

$$\begin{aligned}\frac{\partial \log \pi(\beta_i \mid \theta)}{\partial |\beta_i|} &= p^\star_\theta(\beta_i)\frac{\partial \log \phi(\beta_i \mid \lambda_1)}{\partial |\beta_i|} + [1 - p^\star_\theta(\beta_i)]\frac{\partial \log \phi(\beta_i \mid \lambda_0)}{\partial |\beta_i|} \\ &= -\left[p^\star_\theta(\beta_i)\,\lambda_1 + [1 - p^\star_\theta(\beta_i)]\,\lambda_0\right] \equiv -\lambda^\star_\theta(\beta_i)\end{aligned}$$

where

$$p^\star_\theta(\beta_i) = P(\gamma_i = 1 \mid \beta_i, \theta) = \frac{\theta\phi(\beta_i \mid \lambda_1)}{\theta\phi(\beta_i \mid \lambda_1) + (1-\theta)\phi(\beta_i \mid \lambda_0)}$$

is the conditional probability that β_i was drawn from $\phi(\beta_i \mid \lambda_1)$.

- $\lambda^\star_\theta(\beta_i)$ is an adaptive convex combination of λ_1 and λ_0.
- $\lambda^\star_\theta(\beta_i)$ puts more weight on the slab penalty λ_1 when β_i is large, and puts more weight on the spike penalty λ_0 when β_i is small.

SSL is a “Self-adaptive LASSO"

Let $z_j = \boldsymbol{X}_j' \boldsymbol{e}_j$ where $\boldsymbol{e}_j = \boldsymbol{Y} - \sum_{k \neq j} \boldsymbol{X}_k \widehat{\beta}_k$. By the first order conditions

- *The LASSO mode satisfies*

$$\widehat{\beta}_j = \frac{1}{n}[|z_j| - \lambda]_+ \text{sign}(z_j).$$

- Constant penalty regardless of the size of $|z_j|$ - Toooo bad!

- *The Spike-and-Slab LASSO mode satisfies*

$$\widehat{\beta}_j = \frac{1}{n}[|z_j| - \lambda_\theta^\star(\widehat{\beta}_j)]_+ \text{sign}(z_j).$$

- “Self-adaptive" property of the shrinkage term - Wonderful!
- Immediately suggests optimization by coordinate-wise ascent!

Refined Characterization of the Global Mode

- As $\lambda_0 \to \infty$, the posterior becomes multimodal (non-concave), and

$$\widehat{\beta}_j = \frac{1}{n}[|z_j| - \lambda^\star_\theta(\widehat{\beta}_j)]_+ \mathrm{sign}(z_j)$$

  is not sufficient to characterize the global mode.

- Further refinement reveals the *SSL* global mode $\widehat{\boldsymbol{\beta}}$ to be a thresholding rule satisfying

$$\widehat{\beta}_j = \begin{cases} 0 & \text{when} \quad |z_j| \le \Delta \\ \frac{1}{n}[|z_j| - \lambda^\star_\theta(\widehat{\beta}_j)]_+ \mathrm{sign}(z_j) & \text{when} \quad |z_j| > \Delta. \end{cases}$$

  where

$$\Delta \approx \sqrt{2n \log\left[1 + \frac{\lambda_0}{\lambda_1}\frac{1-\theta}{\theta}\right]} + \lambda_1$$

- $\widehat{\boldsymbol{\beta}}$ is a blend of hard and soft thresholding.
- The selection threshold Δ drives the minimax properties of the mode and can be calibrated through suitable choices of $(\lambda_0, \lambda_1, \theta)$.

The Non-Separable Fully Bayes *SSL* Penalty (When θ is random)

The Limitations of Separable Penalties

- Separable penalties $\mathit{Pen}(\boldsymbol{\beta}) = \sum_{i=1}^{p} \mathit{pen}(\beta_i)$ are limited by their inability to adapt to common features across the components of $\boldsymbol{\beta}$.

- This includes the ℓ_1 LASSO penalty, $\mathit{pen}(\beta_i) = -\lambda|\beta_i|$, the ℓ_0, ℓ_2, SCAD, MCP penalties, the separable *SSL* penalty and many more.

- Such separable penalties implicitly assume iid priors, namely $\pi(\boldsymbol{\beta} \mid \eta) = \prod_{i=1}^{p} \pi(\beta_i \mid \eta)$ with some (possibility multivariate) hyperparameter η.

Borrowing Strength via Non-Separable Penalties

- Moving beyond such penalties, **exchangeable priors** from mixing over η

$$\pi(\boldsymbol{\beta}) = \int \prod_{i=1}^{p} \pi(\beta_i \mid \eta)\pi(\eta)\mathrm{d}\,\eta$$

  yield **non-separable penalties** that "borrow strength" across $\beta_1, \ldots, \beta_p$.

- The adaptive potential of $\log \pi(\boldsymbol{\beta})$, from such **hierarchical Bayesian penalty mixing**, is reflected by the adaptive nature of its derivative

$$\frac{\partial \log \pi(\boldsymbol{\beta})}{\partial |\beta_i|} = \int \frac{\partial \log \pi(\boldsymbol{\beta} \mid \eta)}{\partial |\beta_i|} \pi(\eta \mid \boldsymbol{\beta})\mathrm{d}\,\eta.$$

- Such constructions require penalty components that correspond to proper priors, ruling out penalties such as SCAD and MCP.

The Bayesian LASSO: An Exchangeable Attempt

- *Park and Casella (2007)* propose the **Bayesian LASSO**

$$\pi(\boldsymbol{\beta} \mid \lambda) = \prod_{j=1}^{p} \textit{Laplace}(\beta_j \mid \lambda), \quad \lambda \sim \pi(\lambda)$$

  recommending it for posterior median estimation via MCMC.

- But from a penalized likelihood perspective, its limitations for modal estimation are exposed.

- Under $\pi(\lambda)$, the Bayes LASSO posterior mode $\widehat{\boldsymbol{\beta}}$ turns out to be the solution to

$$\widehat{\beta}_j = \frac{1}{n}[|z_j| - \mathrm{E}(\lambda \mid \widehat{\boldsymbol{\beta}})]_+ \mathrm{sign}(z_j).$$

- *Adaptive, but uniform shrinkage for all coordinates - Too bad!*

- $\mathrm{E}(\lambda \mid \widehat{\boldsymbol{\beta}})$ *cannot be calibrated to obtain minimax rates - Toooo bad!*

The Non-Separable Fully Bayes *SSL* Penalty

- Mixing $\pi_{SSL}(\boldsymbol{\beta} \mid \theta)$ over $\pi(\theta)$, the components of $\boldsymbol{\beta}$ become apriori dependent.

$$\pi_{SSL}(\boldsymbol{\beta}) = \int_0^1 \prod_{j=1}^{p} [\theta\phi(\beta_i \mid \lambda_1) + (1-\theta)\phi(\beta_i \mid \lambda_0)] \, \mathrm{d}\, \pi(\theta).$$

- The *SSL* penalty $\log \pi_{SSL}(\boldsymbol{\beta})$ is now **non-separable**, and the lack of a closed form for $\pi_{SSL}(\boldsymbol{\beta})$ complicates its tractability.
- Fortunately, a revealing and simple form can still be obtained for its derivative.
- It is useful to focus on the i^{th} direction, while keeping all other coordinates fixed at $\boldsymbol{\beta}_{\setminus i}$

Further Adaptivity From Bayesian Penalty Mixing

- *The derivative of* $\log \pi(\boldsymbol{\beta})$ *now reveals doubly adaptive penalization that borrows strength across coordinates*

$$\frac{\partial \log \pi(\boldsymbol{\beta})}{\partial |\beta_i|} = -\lambda^\star(\beta_i; \boldsymbol{\beta}_{\setminus i}),$$

where

$$\lambda^\star(\beta_i; \boldsymbol{\beta}_{\setminus i}) = p^\star(\beta_i; \boldsymbol{\beta}_{\setminus i})\, \lambda_1 + [1 - p^\star(\beta_i; \boldsymbol{\beta}_{\setminus i})]\, \lambda_0$$

and

$$p^\star(\beta_i; \boldsymbol{\beta}_{\setminus i}) \equiv \int_0^1 p^\star_\theta(\beta_i) \pi(\theta \mid \boldsymbol{\beta}) \mathrm{d}\,\theta.$$

- By averaging over $\pi(\theta \mid \boldsymbol{\beta})$, the shrinkage term is given an opportunity to **borrow strength** and learn about the sparsity level of $\boldsymbol{\beta}$. - Hooray!

A Surprising and Useful Simplification!

$$p^\star(\beta_i; \boldsymbol{\beta}_{\setminus i}) = p^\star_{\theta_i}(\beta_i), \quad \theta_i = \mathsf{E}[\theta \mid \boldsymbol{\beta}_{\setminus i}]$$

Implications for the Global Mode

- Building on the separable case, the global mode satisfies

$$\widehat{\beta}_i = \begin{cases} 0 & \text{when} \quad |z_i| \le \Delta_i \\ \frac{1}{n}[|z_i| - \lambda^{\star}_{\widehat{\theta}_i}(\widehat{\beta}_i)]_+ \operatorname{sign}(z_i) & \text{when} \quad |z_i| > \Delta_i. \end{cases}$$

  where $\widehat{\theta}_i = \mathsf{E}[\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus i}]$, and

$$\Delta_i \approx \sqrt{2\, n \log\left[1 + \frac{\lambda_0}{\lambda_1} \frac{1 - \mathsf{E}(\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus i})}{\mathsf{E}(\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus i})}\right]} + \lambda_1.$$

- $\widehat{\beta}_i$ is now a *doubly adaptive blend of soft and hard thresholding* with adaptive coordinate-specific thresholds Δ_i.
- When $\theta \sim \mathcal{B}(1, D\,p)$ for some $D > 0$,

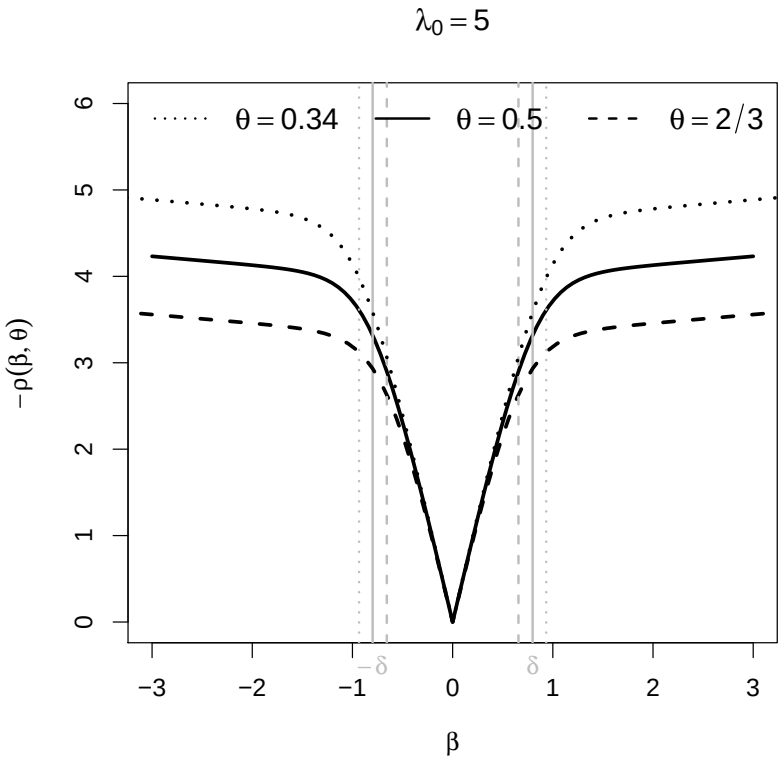
$$\mathsf{E}(\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus i}) \sim \frac{\widehat{q}}{p}, \qquad \widehat{q} = \|\widehat{\boldsymbol{\beta}}\|_0$$

- More refined tuning of the Δ_i for improved minimax rates becomes available through suitable choices of (λ_0, λ_1).

Automatic Multiplicity Control

Assume $p = 2$ with $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, and suppose $\theta \sim \mathcal{B}(1, 1)$.

The univariate SSL penalty for 1^{st} *direction, while* **keeping** β_2 **fixed:**



$\rightsquigarrow$ If $\beta_2 = 0 \rightarrow \mathrm{E}[\theta \mid \beta_2 = 0] = 0.34$ Δ_i goes up

$\rightsquigarrow$ If $\beta_2 = 4 \rightarrow \mathrm{E}[\theta \mid \beta_2 = 4] = 2/3$ Δ_i goes down

The Spike-and-Slab LASSO:

Implementation

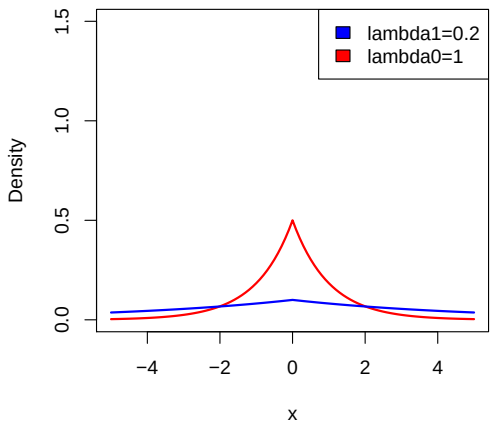
Dynamic Posterior Exploration

LASSO:

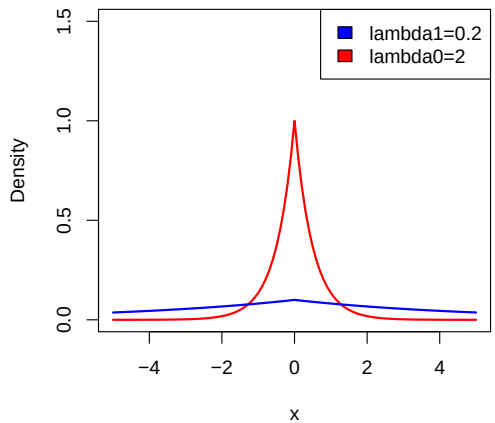
A path-following algorithm indexed by a sequence of Laplace priors

Spike-and-Slab LASSO:

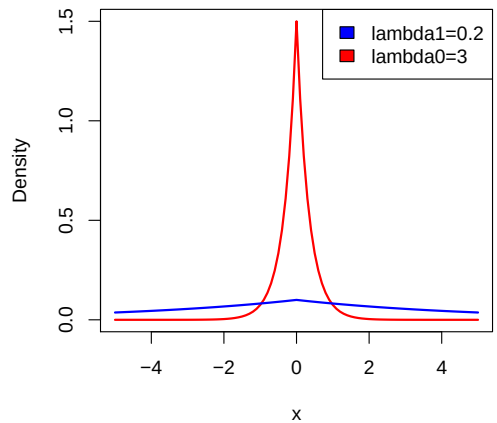
A path-following algorithm indexed by a sequence of Laplace **mixtures**



$\lambda_0 = 1$



$\lambda_0 = 2$



$\lambda_0 = 3$

EMVS

Find $(\widehat{\gamma}, \widehat{\theta}) = \arg\max_{(\boldsymbol{\beta},\theta)} \pi(\boldsymbol{\beta}, \theta \mid \boldsymbol{Y})$ iteratively with an

EM algorithm by treating γ as missing data

- **E-step:** Let $\boldsymbol{\beta}^{(k)}$ and $\theta^{(k)}$ be the most recent updates of $(\boldsymbol{\beta}, \theta)$. For $j = 1, \ldots, p$ compute $\lambda^{\star}_{\theta^{(k)}}(\beta_j^{(k)})$ where

$$\lambda^{\star}_{\theta}(\beta_j) = p^{\star}_{\theta}(\beta_j)\,\lambda_1 + [1 - p^{\star}_{\theta}(\beta_j)]\,\lambda_0$$

- **M-step:** An adaptive LASSO regression

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ -\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 - \sum_{j=1}^{p} \lambda^{\star}_{\theta^{(k)}}(\beta_j^{(k)})|\beta_j| \right\}$$

Update the weight

$$\theta^{(k+1)} = \frac{p^{\star}_{\theta^{(k)}}(\beta_j^{(k)}) + a - 1}{a + b + p - 2}$$

Refined Coordinate Ascent

Refined Coordinate Ascent:

(Mazumder et al. (2011), Breheny and Huang (2011))

Targeted towards local maxima that are global maximizers in each direction

Beginning with $\boldsymbol{\beta}^{(0)}$, proceed iteratively with

$$\widehat{\beta}_j^{(k)} = \frac{\mathbb{I}(|z_j| > \Delta_j)}{n} \left[|z_j| - \lambda^{\star}_{\theta_j} \left(\widehat{\beta}_j^{(k-1)} \right) \right]_+ \operatorname{sign}(z_j)$$

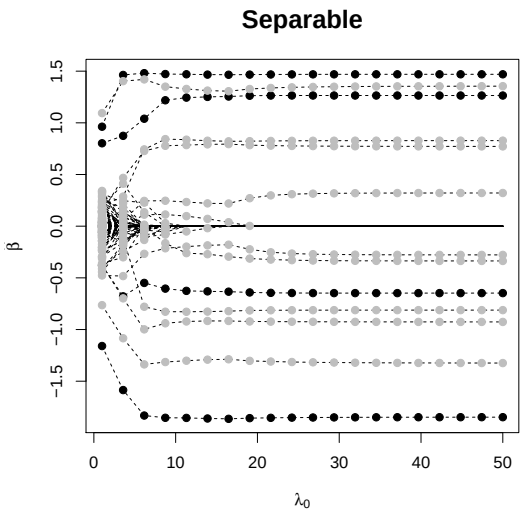
$$\theta_j = \mathsf{E}\left[\theta \,\middle|\, \widehat{\boldsymbol{\beta}}_{\setminus j}^{(k-1)} \right]$$

where $z_j = \boldsymbol{X}_j' \boldsymbol{e}_j$ for $\boldsymbol{e}_j = \boldsymbol{Y} - \sum_{k \neq j} \boldsymbol{X}_k \widehat{\beta}_k$.

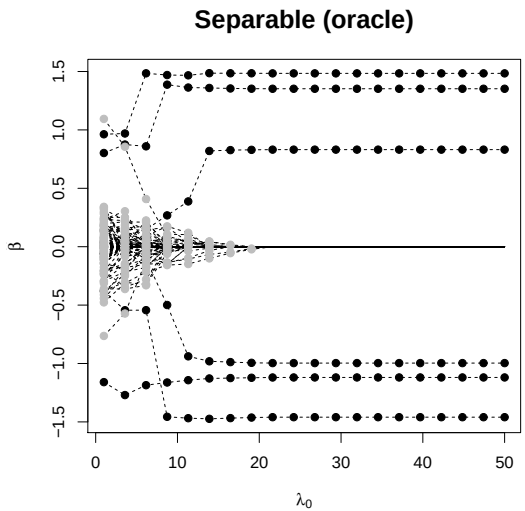
Our non-separable penalty requires only a simple additional step!

The Spike-and-Slab LASSO in Action

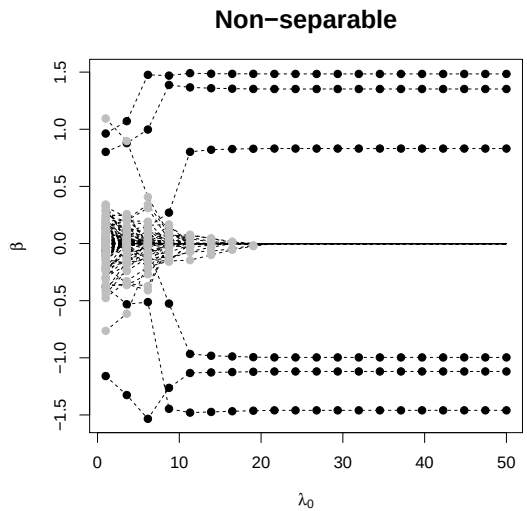
- $p = 1000$ and $n = 100$
- 50 indep groups of 20 correlated ($\rho_{ij} = 0.9$) predictors $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$
- β_0 assigned $q = 6$ nonzero entries $\frac{1}{\sqrt{3}}(-2.5, -2, -1.5, 1.5, 2, 2.5)$, one within each of the first 6 blocks



$\theta = 0.5$



$\theta = 6/1000$

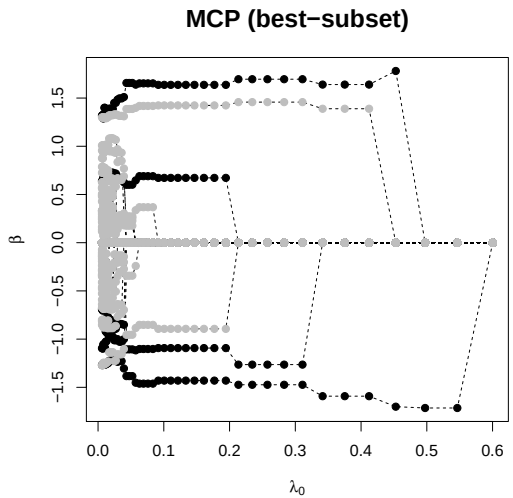


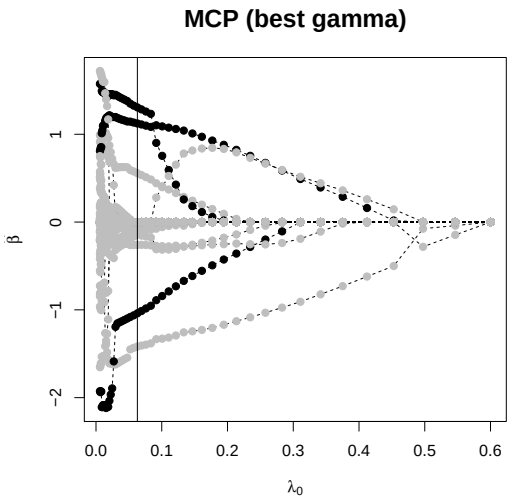
$\theta \sim \mathcal{B}(1, p)$

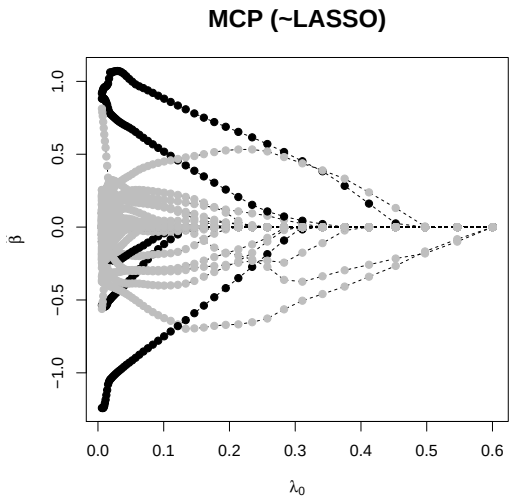
The non-separable SSL adapts and mimics the oracle choice!

Comparison with the MCP Penalty

$$pen_{MCP}(\beta) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2\gamma} & \text{if} \quad |\beta| \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & \text{if} \quad |\beta| > \gamma\lambda \end{cases}$$







Hard Tresholding
$\gamma \approx 1$

Best Cross-validated
$\gamma = 4.185$

Towards the LASSO
$\gamma = 8.5$

As opposed to the Spike-and-Slab LASSO:

Cross-validation needed over a two-dimensional grid of values (λ, γ)

The regularization path does not stabilize.

A Simulation Comparison with Competing Methods

			Correlated Block Design						
	λ_1	θ	MSE	FDR	FNR	$\widehat{q}$	TRUE	TIME	HAM
SSL	1	$\frac{6}{1000}$	3.21	0.253	0.253	6	21	0.34	3.04
SSL	0.1	$\mathcal{B}(1,p)$	3.32	0.255	0.257	5.99	23	0.69	3.07
SSL	1	$\mathcal{B}(1,p)$	3.33	0.26	0.26	6	22	0.48	3.12
Horseshoe			3.19	0.246	0.417	4.64	1	465.84	3.64
EMVS			5.89	0.074	0.688	2.02	0	0.78	4.28
MCP*			6.77	0.563	0.483	7.09	1	2.04	6.89
Adaptive-LASSO			2.79	0.549	0.192	10.75	2	5.37	7.05
SSL	1	0.5	5.98	0.574	0.31	9.71	2	0.33	7.43
SCAD*			8.39	0.77	0.57	11.2	0	0.52	12.04
LASSO			3.47	0.845	0.113	34.35	0	0.74	29.71

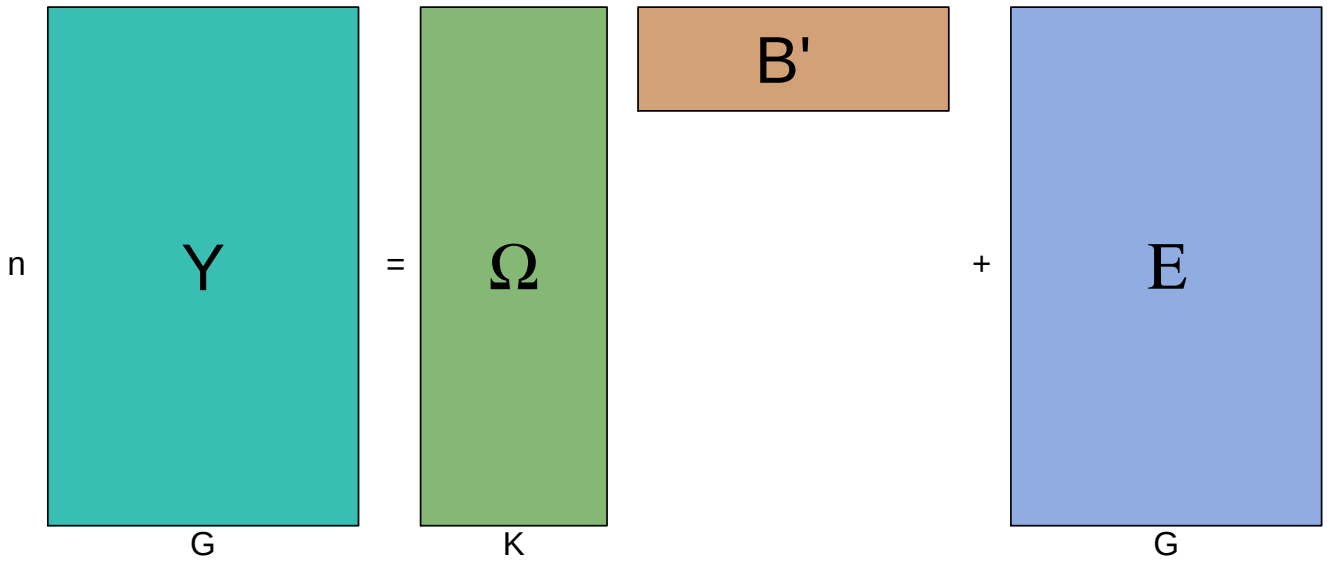
Table: Simulation study using 100 repetitions; MSE (average mean squared error), FDR (false discovery rate), FNR (false non-discovery rate), DIM (average size of the model), TRUE (# true model detected), TIME (average execution time in seconds), HAM (average Hamming distance); Methods have been sorted based on the Hamming distance. ($\star$: `ncvreg` implementation using cross-validation over a one-dimensional grid with a default value of the second tuning parameter).

Fast Bayesian Factor Analysis With The Spike-and-Slab LASSO

An SSL Application to Bayesian Factor Analysis

Generic factor model for **fixed number** K of latent factors:

$$\boldsymbol{Y}_i \mid \boldsymbol{\omega}_i, \boldsymbol{B}, \Sigma \overset{\text{ind}}{\sim} \mathcal{N}_G\left(\boldsymbol{B}\boldsymbol{\omega}_i, \Sigma\right), \quad \boldsymbol{\omega}_i \sim \mathcal{N}_K(\mathbf{0}, \mathrm{I}_K) \quad 1 \leq i \leq n,$$



$\rightsquigarrow$ $\boldsymbol{E} = [\epsilon_1, \ldots, \epsilon_n]'$ with $\epsilon_i \overset{\text{ind}}{\sim} \mathcal{N}_G(\mathbf{0}, \Sigma)$, $\Sigma = \text{diag}\{\sigma_j^2\}_{j=1}^G$

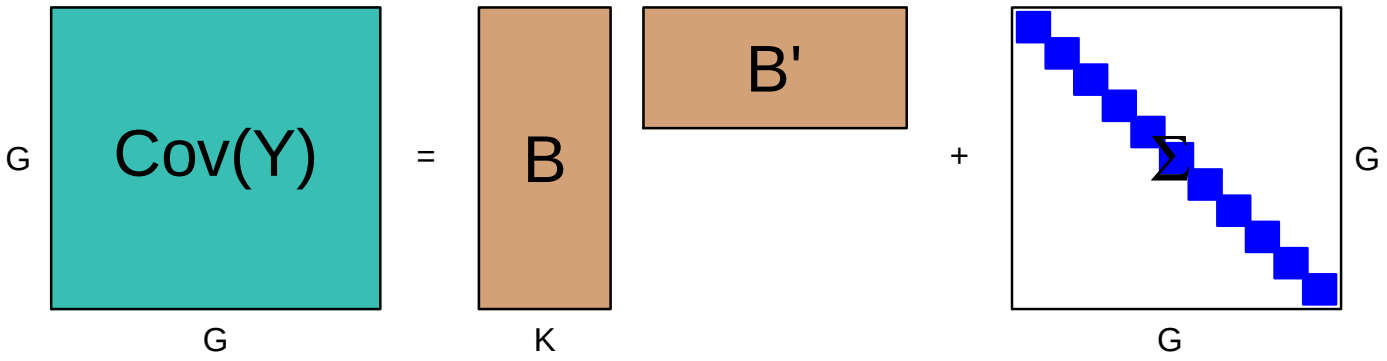
$\rightsquigarrow$ $\boldsymbol{\Omega} = [\omega_1, \ldots, \omega_n]'$: latent factors

$\rightsquigarrow$ $\boldsymbol{B} = (b_{jk})_{j,k=1}^{G,K}$: factor loadings

An SSL Application to Bayesian Factor Analysis

Integrating out $\boldsymbol{\omega}_i \sim \mathcal{N}_K(\mathbf{0}, \mathrm{I}_K)$ yields

$$f(\boldsymbol{y}_i \mid \boldsymbol{B}, \boldsymbol{\Sigma}) = \mathcal{N}_G(\mathbf{0}, \boldsymbol{B}\boldsymbol{B}' + \boldsymbol{\Sigma}),\ 1 \leq i \leq n.$$



- Because $\boldsymbol{B}\boldsymbol{B}' = (\boldsymbol{B}\boldsymbol{P})(\boldsymbol{B}\boldsymbol{P})'$, for any orthogonal matrix $\boldsymbol{P}$, the likelihood is invariant under factor rotation.
- Components of $\boldsymbol{B}$ are unidentifiable.
- Effective factor cardinality K is unknown.

The Prior and Algorithm Underlying Our Approach

An SSL-IBP prior on infinite-dimensional $\boldsymbol{B} = \{\beta_{jk}\}_{j,k}^{G,\infty}$ *which anchors on sparsity inducing factor orientations.*

- A **Spike-and-Slab LASSO (SSL)** prior

$$\pi(\beta_{jk}|\gamma_{jk}) \sim \gamma_{jk}\phi(\beta_{jk}|\lambda_1) + (1-\gamma_{jk})\phi(\beta_{jk}|\lambda_0),$$

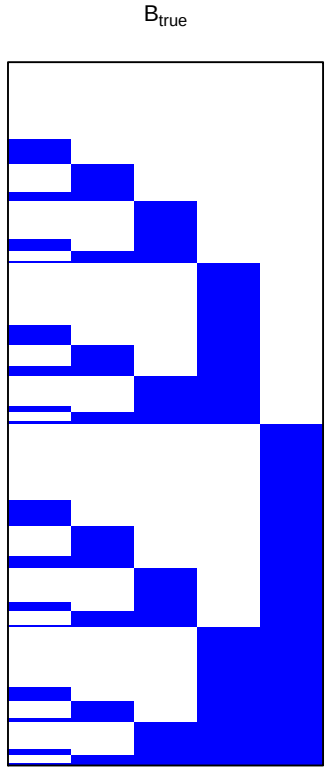
controlled by an **Indian Buffet Process (IBP)** on 0-1 $\gamma'_{jk}s$

$$\gamma_{jk} \sim \texttt{Bern}[\theta_{(k)}], \quad \theta_{(k)} = \prod_{l=1}^{k} \nu_l, \quad \nu_l \overset{\text{iid}}{\sim} \mathcal{B}(\alpha, 1).$$

- Set $\lambda_1 << \lambda_0$ to adaptively threshold smaller β_{jk}.
- Prespecification of K and identifiability constraints are avoided.
- Implementation with a parameter expanded likelihood EM algorithm yields automatic rotations which converge rapidly to excellent sparse modal estimates.

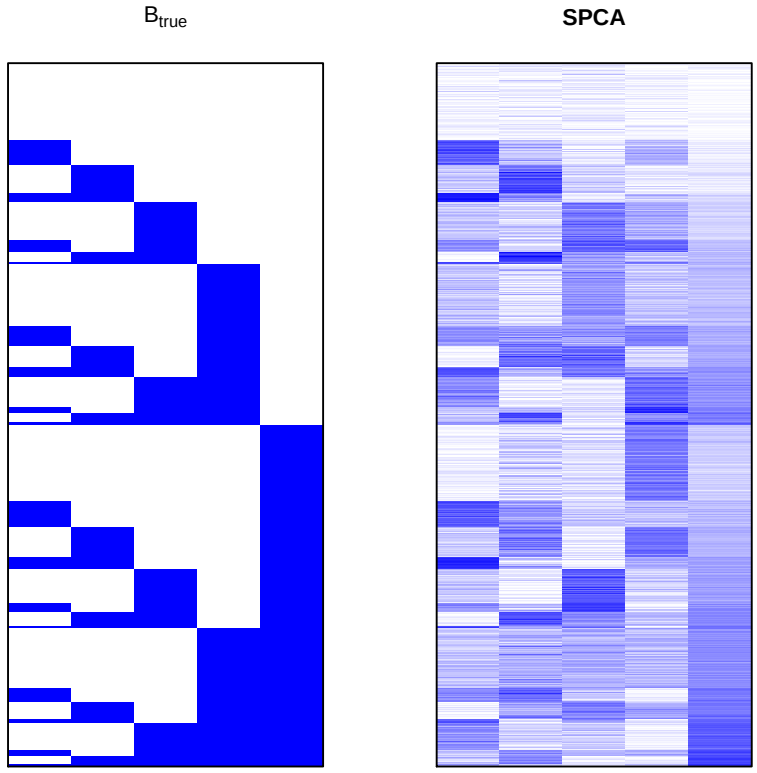
The SSL-IBP Prior with Automatic Rotations

A challenging problem with $n = 100, G = 2000, K = 5$:



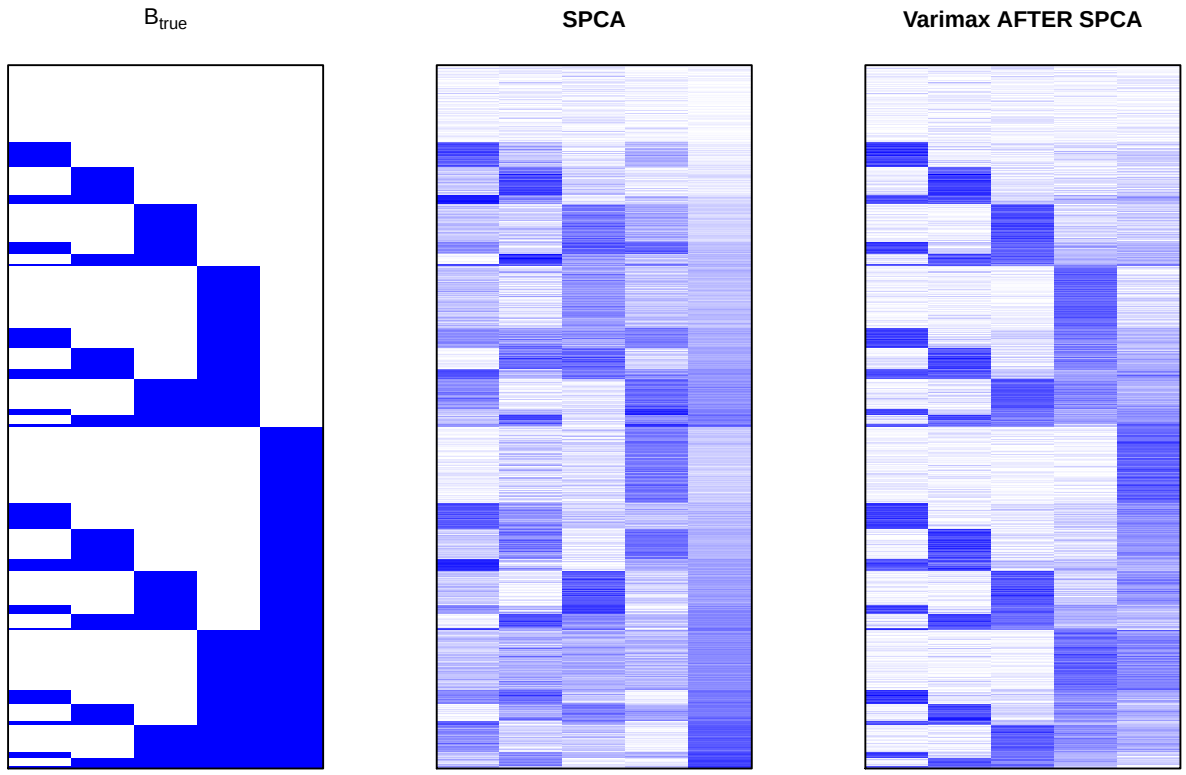
The SSL-IBP Prior with Automatic Rotations

A challenging problem with $n = 100, G = 2000, K = 5$:



The SSL-IBP Prior with Automatic Rotations

A challenging problem with $n = 100, G = 2000, K = 5$:



The SSL-IBP Prior with Automatic Rotations

A challenging problem with $n = 100, G = 2000, K = 5$:

Some Final Remarks

- Implementation of the Spike-and-Slab LASSO with the C-written R package SSLASSO, is available on CRAN.

- By suitable tuning of λ, the LASSO mode can achieve the near minimax rate of convergence. However, the concentration of the full posterior of the LASSO is a disaster. For the minimax choice of λ, it puts essentially no mass on balls around β_0 with a radius of a substantially larger order than the minimax rate. (Castillo et al. (2015)) .

- By suitable tuning of Δ_i's, both the global mode and the posterior concentration of the Spike-and-Slab LASSO can achieve the near-minimax rate of convergence. Unlike the LASSO, the Spike-and-Slab LASSO posterior keeps pace with the global mode!

- The SSL prior can be incorporated naturally into general Bayesian methodology. For example, R&G (2016) used an SSL prior coupled with an Indian Buffet Process $\pi(\gamma)$ for fast Bayesian Factor Analysis.

Thank you!

Some References

- Ročková, V. and George E. (2016+). The Spike-and-Slab LASSO, *Journal of the American Statistical Association*, (*in press*).
- Ročková, V. (2018), Bayesian Estimation of Sparse Signals with a Continuous Spike-and-Slab Prior, *The Annals of Statistics*, 46:401–437).
- Ročková, V. and George, E. (2016), Bayesian Penalty Mixing: The Case of a Non-separable Penalty, *Statistical Analysis for High-Dimensional Data, Abel Symposia 11*.
- Ročková, V. and George E. (2016), Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity, *Journal of the American Statistical Association, 111:1608-1622*.
- Ročková, V. and George, E. (2014), EMVS: The EM Approach to Bayesian Variable Selection, *Journal of the American Statistical Association*,109:828-846.