Adapting to Sparsity and Heavy Tailed Data

Mohamed Abdelkader Abba [1]

University of Arkansas, Department of Mathematical Sciences

July 2, 2018

[1] Under the supervision of Dr. J. Datta

Outline

Sparsity

Many modern applications of statistical inference involve wide data sets, where the number of features $p \gg n$.
Examples include, gene expression data, finance, astronomy...
The most popular of these problems are:

- sparse normal means: $(Y_i \mid \beta_i) \overset{\text{ind}}{\sim} \mathcal{N}(\beta_i, 1), i = 1, \ldots, n$;
- sparse linear regression: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$,

where $\boldsymbol{\beta}$ is a 'nearly black object', that is,

$$\boldsymbol{\beta} \in l_0[p_n] \equiv \{\boldsymbol{\beta} : \#(\beta_i \neq 0) \leq p_n\}$$

Penalized Regression

- Rich variety of methodologies for high-dimensional inference based on regularization which implicitly or explicitly penalizes model complexity.
- This amounts to controlling the bias-variance trade-off and are particularly useful for sparse learning, when the number of variables (p) exceed the number of observations (n).
- In the context of linear regression $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, a regularized estimate of $\boldsymbol{\beta}$ is obtained by minimizing the penalized likelihood:

$$\hat{\boldsymbol{\beta}}^{\text{pen}}_{\lambda^*} = \underset{\boldsymbol{\beta} \in \Re^p}{\text{argmin}} \{ ||Y - X\boldsymbol{\beta}||^2 + \lambda^* \Omega(\boldsymbol{\beta}) \}, \tag{1}$$

$$\text{where, } \Omega(\boldsymbol{\beta}) = \sum_{j=1}^{p} \omega(\beta_j) \text{ is a separable penalty} \tag{2}$$

Lasso

The gold standard for regularized methods, simultaneously performs estimation and model selection. Lasso constrains the ℓ_1 norm of the parameter vector:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}}_{\lambda^*} = \underset{\boldsymbol{\beta}\in\Re^p}{\text{argmin}}\{||Y - X\boldsymbol{\beta}||^2 + \lambda^*\|\boldsymbol{\beta}\|_1\} \tag{3}$$

Lasso enjoys both:

- computational efficiency, due to the LARS Efron et al. [2004];
- theoretical optimality properties in recovering the true $\boldsymbol{\beta}_0$ Bühlmann and van de Geer [2011].

$\sqrt{\text{Lasso}}$

Lasso presents two major drawbacks:

- dependence on knowledge / estimation of σ, this is particularly problematic when $p \gg n$;
- estimates are not scale invariant: $\hat{\boldsymbol{\beta}}(\sigma Y, X) \neq \sigma \hat{\boldsymbol{\beta}}(Y, X)$.

The later property can be achieved by setting $\lambda^* = \lambda\sigma$ yielding:

$$\hat{\boldsymbol{\beta}}^{\text{inv}} = \sigma^{-1} ||Y - X\boldsymbol{\beta}||^2 + \lambda \|\boldsymbol{\beta}\|_1. \tag{4}$$

Estimating σ by $||Y - X\boldsymbol{\beta}|| / \sqrt{n}$ leads to:

$$\hat{\boldsymbol{\beta}}_{\lambda}^{\sqrt{\text{lasso}}} = \underset{\boldsymbol{\beta} \in \Re^p}{\text{argmin}} \{ \sqrt{n} ||Y - X\boldsymbol{\beta}||_2 + \lambda \|\boldsymbol{\beta}\|_1 \} \tag{5}$$

$\sqrt{\text{Lasso}}$

The new estimator satisfies:

- scale invariance;
- independent of the knowledge of σ;
- computational efficiency due to convexity of objective function;
- Belloni et al. [2011], near oracle properties when $\#\text{supp}(\boldsymbol{\beta}_0) = s < n$.

Despite the attractive properties of the methods, they share a common caveat:

- choice of tuning parameter λ;
- lack of uncertainty quantification.

Hence the need for a Bayesian treatment.

Bayesian representation

By looking at the objective function as the logarithm of likelihood $\times$ prior, we get:

$$\min_{\beta \in \Re^d} \{l(y \mid \beta) + \text{pen}_\lambda(\beta)\} = \underset{\beta}{\text{argmax}}\, p(\beta \mid y) = p(y \mid \beta) p_\lambda(\beta)$$

$$\text{where } p(y \mid \beta) \propto \exp\{-l(y \mid \beta)\}, \quad p_\lambda(\beta) \propto \exp\{-\text{pen}_\lambda(\beta)\}. \tag{6}$$

This correspondence coupled with a full Bayesian treatment naturally leads to:

- uncertainty quantification;
- automatic tuning of λ, using full or empirical Bayes.

Outline

Bayesian $\sqrt{\text{Lasso}}$

To derive a hierarchical model for the $\sqrt{\text{Lasso}}$, we need:

$$p(Y \mid \boldsymbol{\beta}) \propto \exp\left[-||Y - X\boldsymbol{\beta}||_2\right], \text{ and using the following identity}$$

$$\exp\{-a(2\lambda)^{1/2}\} = \int_0^\infty \frac{a}{(2\pi)^{1/2}(v^2)^{1/2}} \exp\{-\frac{\lambda}{v^2}\} \exp\{-\frac{a^2 v^2}{2}\} dv^2$$

$$\text{with } a = 1, \text{ and } 2\lambda = ||Y - X\boldsymbol{\beta}||_2^2, \text{ we get:}$$

$$\exp\left[-\|Y - X\boldsymbol{\beta}\|_2\right] = \int_0^\infty \frac{1}{(2\pi)^{1/2}(v^2)^{1/2}} \exp\{-\frac{\|Y - X\boldsymbol{\beta}\|_2^2}{2v^2}\} \exp\{-\frac{v^2}{2}\} dv^2$$

In other words:

$$\exp\left[-\|Y - X\boldsymbol{\beta}\|_2\right] \propto \int \mathcal{N}(Y;\, X\boldsymbol{\beta}, v^2\mathbf{I}_n) \times \mathcal{G}(v^2;\, \frac{(n+1)}{2}, \frac{1}{2}) dv^2$$

Prior on $\boldsymbol{\beta}$

To complete the hierarchy we use the normal scale mixture representation of the Laplace distribution:

$$\pi(\beta_i) \propto e^{-\tau|\beta_i|} = \int_0^\infty \frac{1}{\sqrt{2\pi}\lambda_i} \mathrm{e}^{-\beta_i^2/(2\lambda_i^2)} \, \frac{\tau^2}{2} \mathrm{e}^{-\lambda_i^2\tau^2/2} \mathrm{d}\lambda_i^2$$

equivalently:

$$\left.\begin{array}{ll} \beta_j & \overset{iid}{\sim} \mathcal{N}(0, \lambda_j^2) \\ \lambda_j^2 & \overset{iid}{\sim} \mathcal{E}\mathrm{xp}(\tau^2/2) \end{array}\right\} \Rightarrow \boldsymbol{\beta} \sim DE(\tau) \text{ and } \tau^2 \sim \pi\left(\tau^2\right).$$

Full Hierarchical Model

Combining the scale mixture representations of the likelihood and the prior we get the full hierarchical model

$$
\begin{aligned}
[\mathbf{y} \mid \boldsymbol{\beta}, v^2] &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, v^2\mathbf{I}) \\
[\boldsymbol{\beta} \mid \boldsymbol{\lambda}] &\sim \mathcal{N}(\mathbf{0}, D_{\boldsymbol{\lambda}}), \\
\text{where } D_{\boldsymbol{\lambda}} &= \text{Diag}(\lambda_1^2, \ldots, \lambda_p^2) \\
[\lambda_1^2, \ldots, \lambda_p^2 \mid \tau^2] &\sim \prod_{j=1}^{p} \frac{\tau^2}{2} e^{-\lambda_j^2 \tau^2/2} \mathrm{d}\lambda_j^2, \quad \lambda_j^2 > 0, \\
[v^2] &\sim \mathcal{G}amma((n+1)/2, 1/2), \\
[\tau^2] &\sim p(\tau^2) d\tau^2, \ \tau^2 > 0. \\
[\tau^2 \sim \mathcal{G}(r, \delta), \text{ or } \tau^2 &\sim \mathcal{C}(0, 1).]
\end{aligned}
$$

Full conditionals for the Gibbs sampler

Under the Gamma hyper-prior on τ^2, the joint distribution of y_i and all the hyperparameters in the model is :

$$f(\mathbf{y}, \boldsymbol{\beta}, v^2, \boldsymbol{\lambda}, \tau^2 \mid \mathbf{r}, \delta) \propto \frac{1}{(2\pi)^{1/2} v} e^{-v^2/(2)} \exp\{-\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2/v^2\}$$
$$\prod_{i=1}^{p} (\lambda_i^2)^{-\frac{1}{2}} \exp\{-\beta_i^2/(2\lambda_i^2)\} \frac{\tau^2}{2} e^{-\lambda_i^2\tau^2/2} (\tau^2)^{r-1} e^{-\delta\tau^2} \quad (7)$$

So that the full conditionals are given by:

$$\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\lambda}, t \sim \mathcal{N}\left(\mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}t, \mathbf{A}^{-1}\right),$$
$$\text{where } \mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{X}t + \mathbf{D}_{\lambda}^{-1} \text{ and } t = 1/v^2$$
$$1/v^2 \mid \mathbf{y}, \boldsymbol{\beta} \sim \text{Inv-Gauss}\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^{-1}, 1\right)$$
$$\lambda_i^{-2} \mid \beta_i, \tau \sim \text{Inv-Gauss}(|\tau/\beta_i|, \tau^2)$$
$$\tau^2 \mid \boldsymbol{\lambda}, r, \delta \sim \text{Gamma}(p + r, \delta + \sum_{i=1}^{p} \lambda_i^2/2)$$

Shrinkage Profile

In the special case where $\mathbf{X} = \mathbf{I}_n$, $\beta_i \mid y_i, \lambda_i, t \sim \mathcal{N}\left(y_i \frac{\lambda_i^2 t}{1+t\lambda_i^2}, \frac{\lambda_i^2}{1+t\lambda_i^2}\right)$.

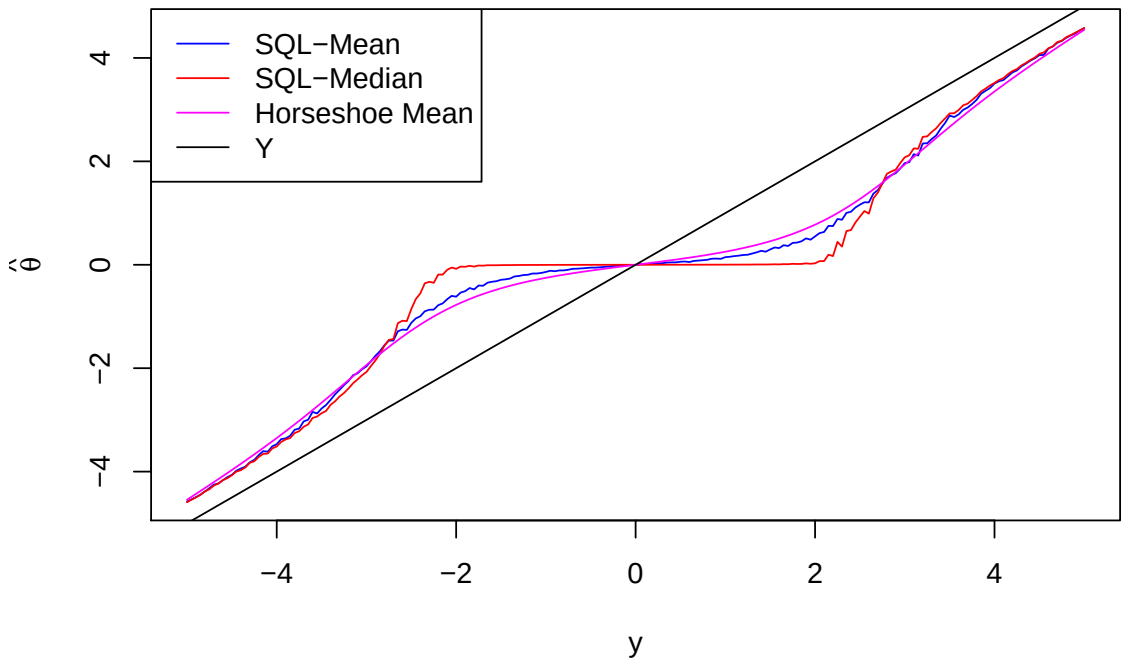


Figure: Shrinkage profile for the horseshoe posterior mean and Bayesian $\sqrt{\text{Lasso}}$ posterior mean and median estimators.

Dependence on σ^2

$\sqrt{\text{Lasso}}$'s main advantage is its ambivalence to the error variance σ^2. Does this property carry over to the Bayesian representation ?

Horseshoe

Hierarchical model:

$$(y_i \mid \beta_i) \sim \mathcal{N}(\beta_i, \sigma^2),\ (\beta_i \mid u_i, \tau) \sim \mathcal{N}(0, u_i^2\tau^2),\ u_i^2 \sim \mathsf{C}^+(0,1). \quad (8)$$

With different treatments available for τ. The following example from Polson and Scott [2010] highlights the issue of dependence between τ and σ^2.

Dependence on σ^2

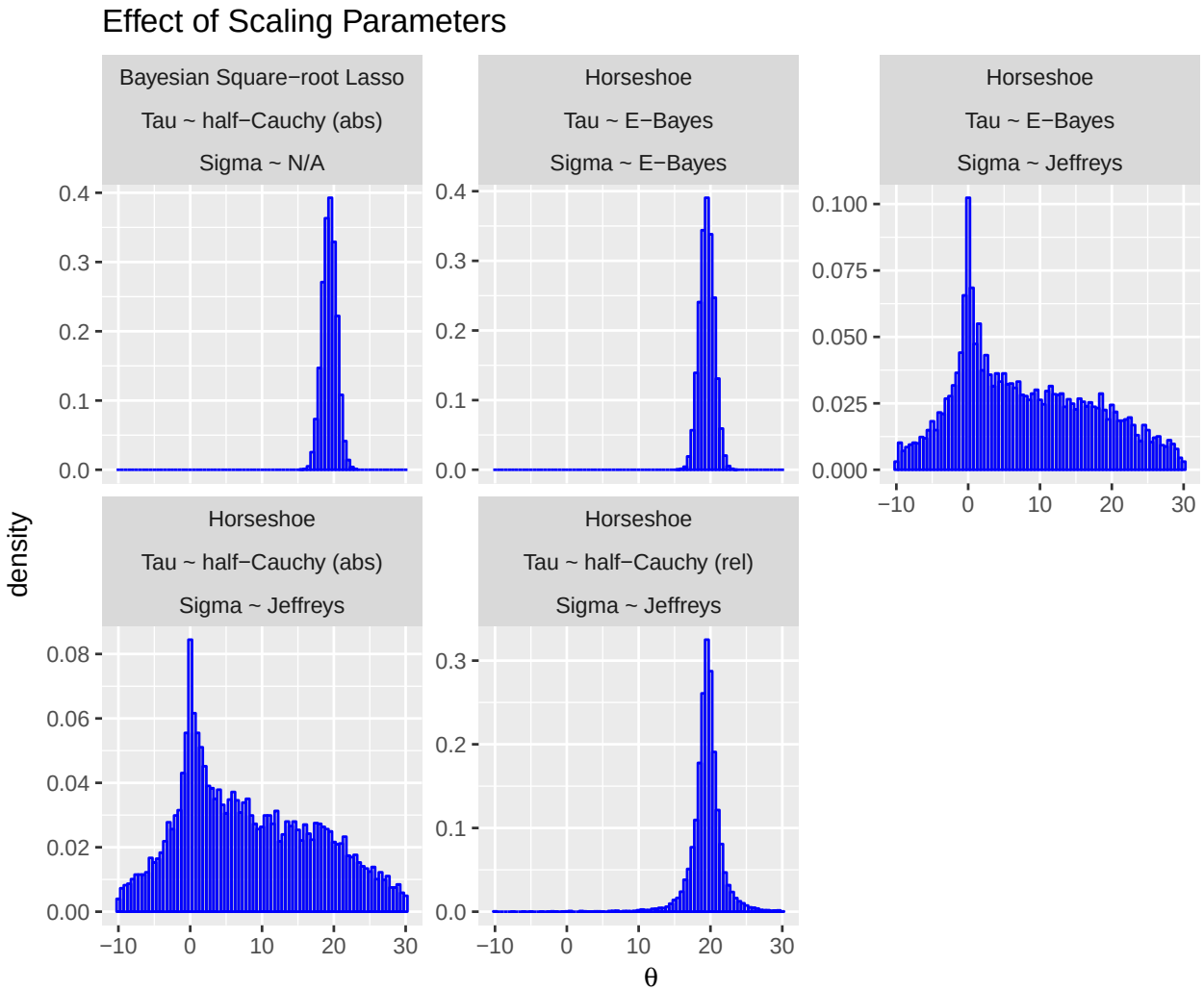


Figure: Behavior of the posterior density under different methods of handling the hyper-parameters σ^2 and τ for the Horseshoe prior as well as the Bayesian $\sqrt{\text{Lasso}}$.

Outline

Global-Local Shrinkage priors

The prior placed on $\boldsymbol{\beta}$ so far:

$$\beta_j \overset{iid}{\sim} \mathcal{N}(0, \tau_j^2),\ \tau_j^2 \overset{iid}{\sim} \mathcal{E}\text{xp}(\lambda^2/2), \text{and } \lambda^2 \sim \pi(\lambda). \tag{9}$$

- Independence of $(\beta_j \mid \lambda) \Rightarrow$ inadequate prior mass on sparse regions.
- No global shrinkage parameter, that controls or accounts for sparsity.

Instead G-L priors generally take the following form:

$$\beta_j \overset{iid}{\sim} \mathcal{N}(0, \tau^2\psi_j^2),\ \psi_j \sim f \text{ and } \tau \sim g,$$

where:

- τ, controls how large the parameters are (like a penalty level);
- ψ_j, on the other hand controls how big the parameter is allowed to be locally.

Dirichlet-Laplace prior

Instead of a global parameter in (9), Bhattacharya et al. [2015] introduced a vector of scales $(\phi_1\tau, \ldots, \phi_p\tau)$, where $(\phi_1, \ldots, \phi_p)$ is constrained to lie in the $(n-1)$ dimensional simplex.
Their prior has the following form:

$$\beta_j \mid \phi, \tau \sim \mathrm{DE}(\phi_j\tau),\ \boldsymbol{\phi} \sim \mathrm{Dir}(a, \ldots, a),\ \tau \sim g.$$

This is the Dirichlet-Laplace prior on $\boldsymbol{\beta}$, denoted as $\boldsymbol{\beta} \mid \tau \sim \mathrm{DL}_a(\tau)$.

PROPOSITION

If $\boldsymbol{\beta} \mid \tau \sim \mathrm{DL}_a(\tau)$, then the marginal distribution of β_j given τ is unbounded with a singularity at zero for any $a < 1$.

Tail and origin behavior

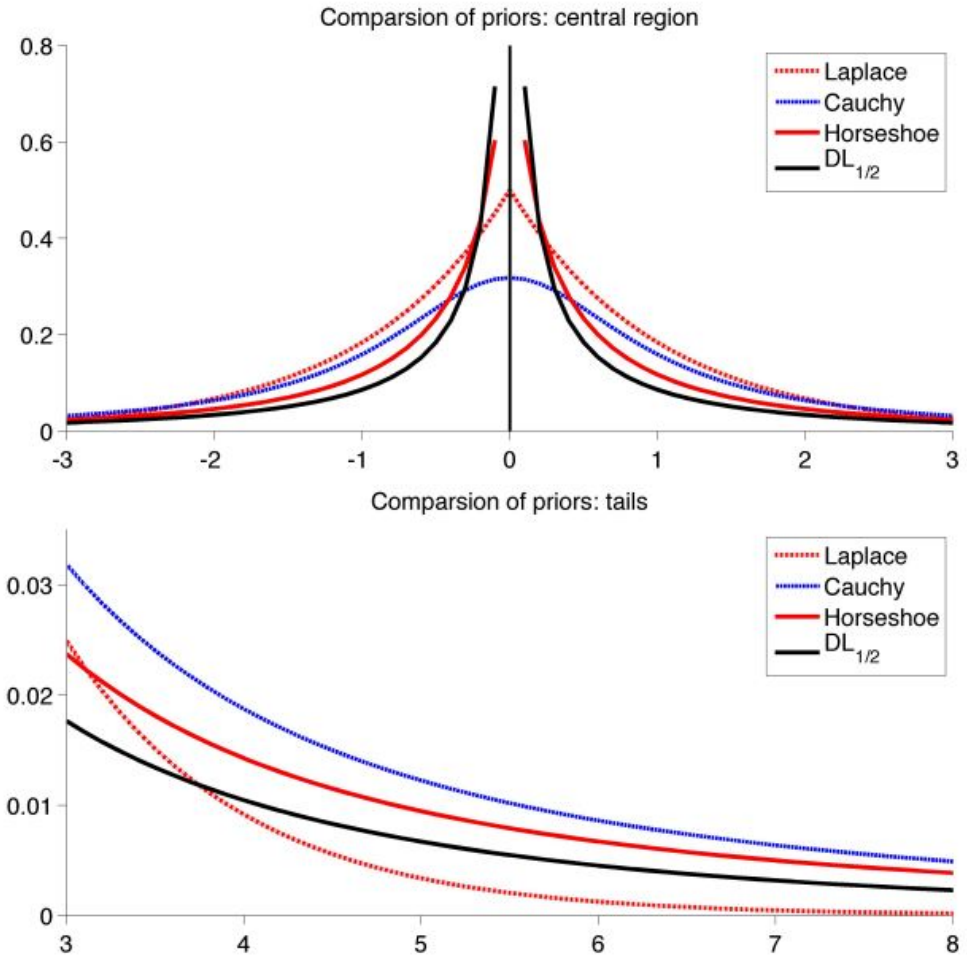


Figure: Marginal density of the DL_a with $a = 1/2$ in comparison to the Horseshoe, the Laplace prior induced by the Bayesian-$\sqrt{\text{Lasso}}$ and the Cauchy prior.

Hierarchical Model

Again using the scale mixture representation of the Laplace distribution:

$$\beta_j \mid \phi, \tau \sim \text{DE}(\phi_j \tau) \Rightarrow \left\{ \begin{array}{rcl} \beta_j & \sim & \mathcal{N}(0, \psi_j \phi_j^2 \tau^2); \\ \psi & \sim & \mathcal{E}\text{xp}(1/2), \end{array} \right.$$

we get the following full hierarchical model:

$$\begin{aligned} [\mathbf{y} \mid \boldsymbol{\beta}, v^2] &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, v^2 \mathbf{I}_n), \\ [\boldsymbol{\beta} \mid \boldsymbol{\phi}, \tau, \boldsymbol{\psi}] &\sim \mathcal{N}(\mathbf{0}, D_{\boldsymbol{\psi}\boldsymbol{\phi}\tau}), \\ D_{\boldsymbol{\psi}\boldsymbol{\phi}\tau} &= \text{Diag}(\psi\phi_1^2\tau^2, \ldots, \psi\phi_p^2\tau^2), \\ \psi_j &\overset{iid}{\sim} \mathcal{E}\text{xp}(1/2), \\ \boldsymbol{\phi} &\sim \text{Dir}(a, \ldots, a), \\ \frac{\tau}{v^2} &\sim \mathcal{G}\text{amma}(pa, 1/2), \\ [v^2] &\sim \mathcal{G}\text{amma}(\frac{n+1}{2}, 1/2). \end{aligned}$$

Posterior Computation

Full conditionals for the Gibbs sampler:

(i) Sample $[\boldsymbol{\beta} \mid \boldsymbol{\psi}, \boldsymbol{\phi}, \tau, \mathbf{v}^2, \mathbf{y}]$ from $\mathcal{N}\left(\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{y}/v^2, \boldsymbol{\Sigma}\right)$, with

$$\boldsymbol{\Sigma}^{-1} = \frac{\mathbf{X}\mathbf{X}^T}{v^2} + \frac{\mathbf{D}^{-1}_{\boldsymbol{\psi}\boldsymbol{\phi}^2}}{\tau^2}.$$

(ii) Conditional posterior of $[\boldsymbol{\psi} \mid \boldsymbol{\phi}, \tau, \boldsymbol{\beta}]$ can sampled in block by independently drawing $\psi_j \mid \phi_j, \tau, \beta_j$ from inv-Gaussian$(\frac{\phi_j \tau}{|\beta_j|}, 1)$

(iii) Sample the conditional posterior of $[\boldsymbol{\phi} \mid \boldsymbol{\beta}]$ by drawing $T_1, \ldots \mathrm{T}_p$ independently from gIG$(a-1, 1, 2|\beta_j|)$ and set $\phi_j = T_j/T$, with $T = \sum_{j=1}^p T_j$.

(iv) Sample $[\tau \mid \boldsymbol{\phi}, \boldsymbol{\beta}, v^2]$ from a gIG$(pa-p, 1, 2\sum_{j=1}^p |\beta_j|/\phi_j)$ distribution.

(v) Sample $[v^2 \mid \boldsymbol{\beta}, \tau, \mathbf{y}]$ by drawing $\frac{1}{\sigma^2}$ from inv-Gaussian$([\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\| + \tau]^{-1}, 1)$.

Outline

Sparse Normal Means

In the normal means problem, the goal is to estimate a sparse vector $\boldsymbol{\theta}$ based on a vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ generated according to the model:

$$Y_i = \theta_i + \epsilon_i, \quad i = 1, \ldots, n. \tag{10}$$

where ϵ_i's are independent standard normal variables and the means vector $\boldsymbol{\theta}$ is assumed to be sparse.
We conduct a simulation study for estimating a sparse normal mean vector for two different choices of $\boldsymbol{\beta}$:

1. $\boldsymbol{\beta} = (\underbrace{7, \ldots, 7}_{q_n=10}, \overbrace{0, \ldots, 0}^{n-q_n=90})$ and
2. $\boldsymbol{\beta} = (\underbrace{7, \ldots, 7}_{q_n=10}, \underbrace{3, \ldots, 3}_{r_n=10}, \overbrace{0, \ldots, 0}^{n-q_n-r_n=80})$.

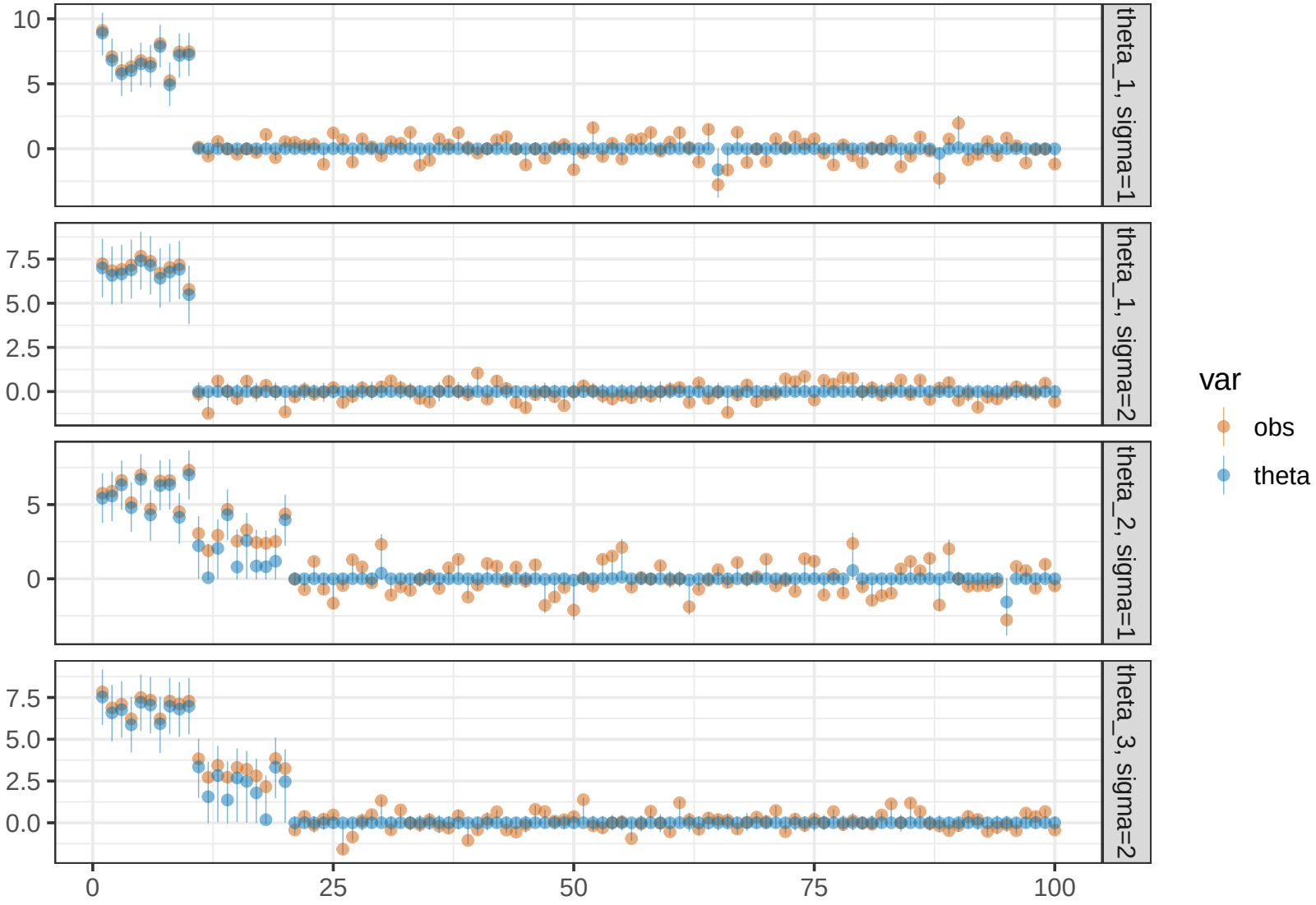


Figure: Comparison of posterior mean estimates for two different sparse normal means, $\beta_i \sim 0.8\delta_{\{7\}} + 0.1\delta_{\{3\}} + 0.1\delta_{\{0\}}$ and $\beta_i \sim 0.9\delta_{\{7\}} + 0.1\delta_{\{0\}}$ under the Bayesian $\sqrt{\text{Lasso}}$

Regression and Variable Selection

Another important area of application of shrinkage priors, is regression and particularly model selection:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{with } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

where $\boldsymbol{\beta}$ is $p \times 1$ and is assumed to be sparse. Sparsity implies that some of the regression coefficients are exactly zero and they correspond to irrelevant predictors.
Motivated by the two groups model (11) and the necessity of a variable selection step for the Bayesian methods, we apply a 2-means clustering of the posterior estimates.

$$\beta_i \sim \tfrac{q}{p}\delta_A + \tfrac{p-q}{p}\delta_0, \text{ so that } \boldsymbol{\beta} = (\underbrace{A, \ldots, A}_{q}, \overbrace{0, \ldots 0}^{p-q}) \qquad (11)$$

Regression and Variable Selection

After clustering the posterior mean vector, we classify the β's according to the following steps :

1- We look first at the two cluster centers $\{\mathbf{c}_1, \mathbf{c}_2\}$, and compare them in absolute value. Let $\mathbf{C}_s = max\{\mathbf{c}_1, \mathbf{c}_2\}$ and $\mathbf{c}_n = min\{\mathbf{c}_1, \mathbf{c}_2\}$. So that the $\mathbf{c}_s$ is the cluster center of the signals while $\mathbf{c}_n$ for the noise.

2- For all $\hat{\beta}_j$, look at the corresponding cluster, if $|\hat{\beta}_j| \in \mathbf{c}_n$, then $\hat{\beta}_j^{dec} = 0$. Otherwise, $\hat{\beta}_j^{dec} = \hat{\beta}_j$.

3- Our final estimated coefficient vector is $\hat{\boldsymbol{\beta}}^{dec} = \left\{\hat{\beta}_j^{dec}\right\}_{1 \leq j \leq p}$.

Unlike $\hat{\boldsymbol{\beta}}$, which will never have exactly zero entries, the new $\hat{\boldsymbol{\beta}}^{dec}$ given by the above described decision rule shrinks the noise coefficients to exactly zero, hence performing a variable selection.

Consistency

Estimation consistency holds if and only if:

$$\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta} \xrightarrow{\mathbf{P}} 0, \text{ as } n \longrightarrow \infty,$$

while model selection consistency requires:

$$\mathbf{P}\left[\{i : \hat{\beta}_i^n \neq 0\} = \{i : \beta_i \neq 0\}\right] \longrightarrow 1, \text{ as } n \longrightarrow \infty.$$

Characterizing a method model selection performance has proven to be a daunting task. However, in the case of the Lasso, some authors have found that there a exist one simple necessary and sufficient condition for the Lasso to select the true model.

Irrepresentability Condition

Suppose, the sample covariance matrix is denoted by $\hat{\Sigma} = nX^T X$ and the active-set $S_0 = \{j : \beta_j \neq 0\}$ consists of first s_0 elements of β. One can partition the $\hat{\Sigma}$ matrix as

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}s_0, s_0 & \hat{\Sigma}s_0, p - s_0 \\ \hat{\Sigma}p - s_0, s_0 & \hat{\Sigma}p - s_0, p - s_0 \end{pmatrix}$$

The irrepresentable condition for variable selection consistency of Lasso is:

$$||\hat{\Sigma}p - s_0, s_0\hat{\Sigma}s_0, s_0{}^{-1} sign(\beta_{S_0})||_\infty \leq \theta \quad \text{for some } 0 < \theta < 1.$$

Zhao et al.[2006], this condition is sufficient and almost necessary.
They also showed that the probability of selecting the true sparse model is an increasing function of the irrepresentability condition number, defined as

$$\eta_\infty = 1 - ||\hat{\Sigma}p - s_0, s_0\hat{\Sigma}s_0, s_0{}^{-1} sign(\beta_{S_0})||_\infty.$$

Effect of irrepresentability

- Simulation scheme: $n = 100, p = 60$ and $q = 7$, $\beta_q^* = (7, 5, 5, 4, 4, 3, 3)^T$, σ^2 was set to 5 to allow for heavy tailed data.
- First draw Σ from Wishart(p, I_p) and then generate $\mathbf{X}$ from $\mathcal{N}(0, \Sigma)$ [Zhao and Yu, 2006].
- Generate 100 different design matrices, and run 100 replicates of each one, and at each iteration we apply the Lasso, horseshoe, Bayesian-$\sqrt{\text{Lasso}}$ and the $\sqrt{\text{DL}}$ 100 times to each of the 100 generated models.
- Select the posterior median and then apply a variable selection step. For the horseshoe, we take advantage of the credible set properties and use them to classify the β_j's. For the other two methods discussed in this work, we implement the k-means clustering procedure discussed previously.

Results

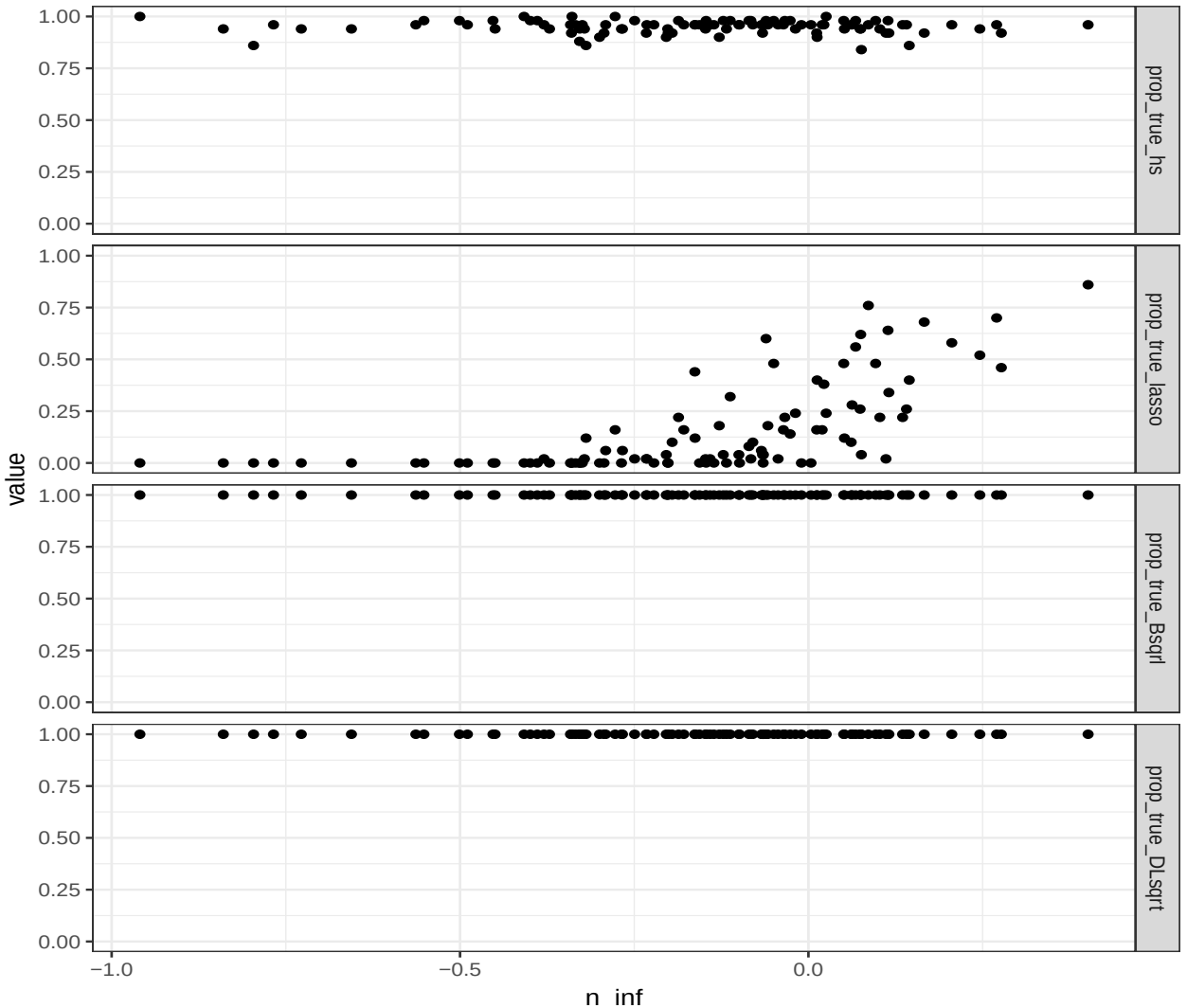


Figure: Proportion of true model selection vs. Irrepresentability Condition

Results

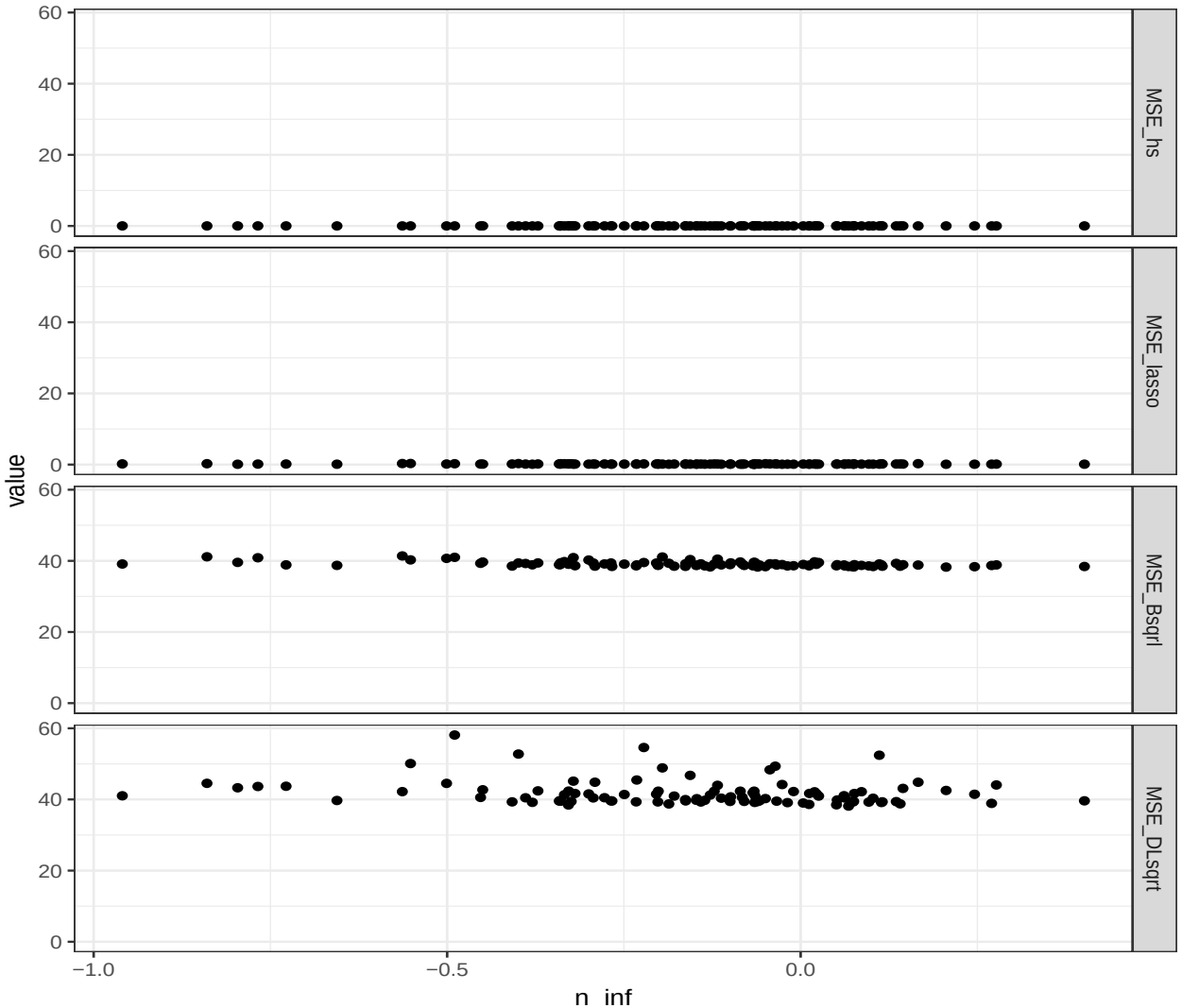


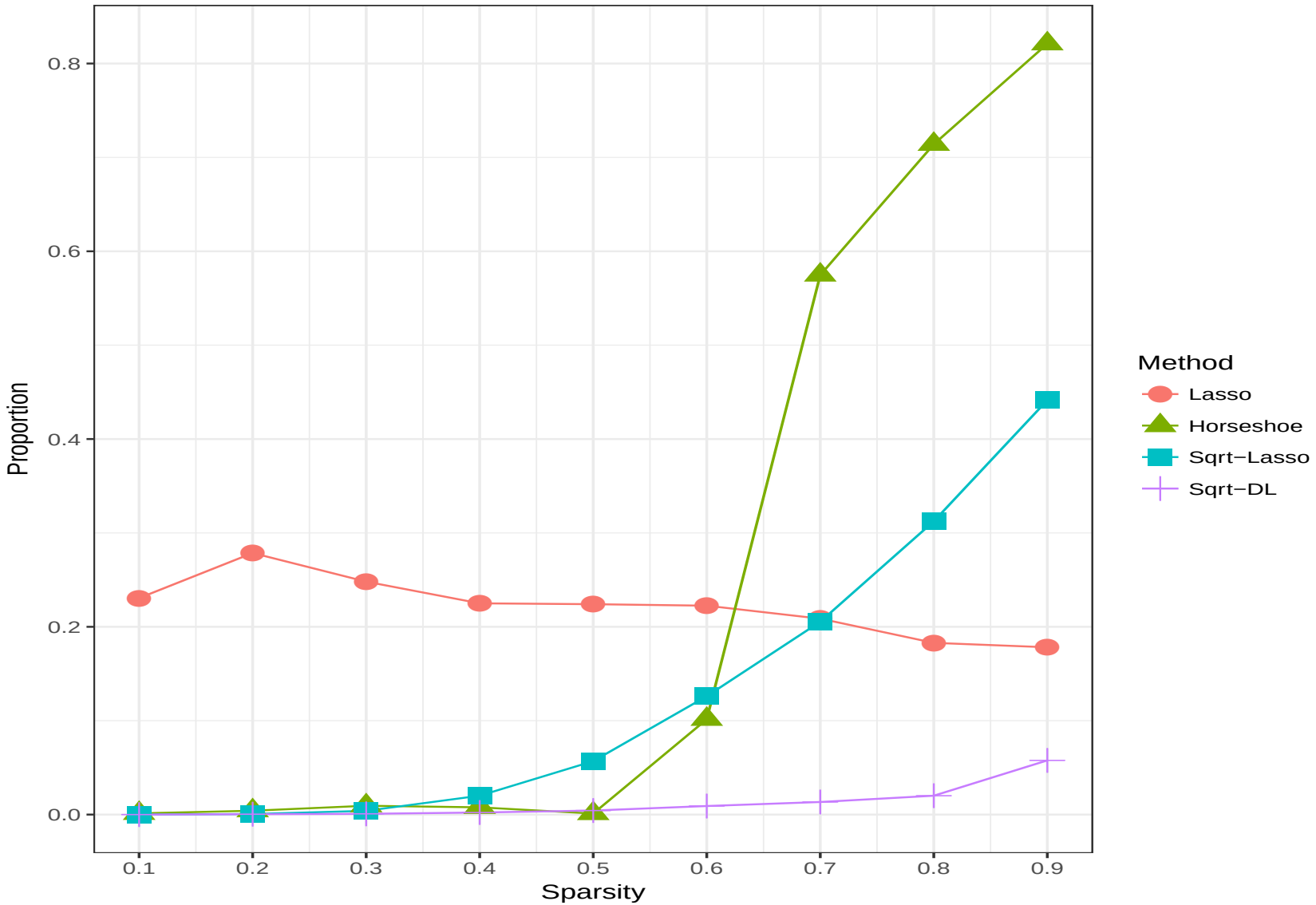
Figure: MSE vs. Irrepresentability Condition

Adapting to Sparsity ?

- Most penalized regression methods, and shrinkage priors operate under the assumption that the parameter of interest is sparse.
- **What if β_0 has zero entries but is not completely sparse ?**
- Simulated data with $n = p = 100$, the design matrix $\mathbf{X}$ rows were simulated from a univariate normal distribution $\mathcal{N}(0, 2)$, the errors variance was set to $\sigma^2 = 5$.
- Sampled 100 different design matrices, and for each of these design matrices, applied the four different methods with varying degrees of sparsity. That is for each of the 100 designs, say $\mathbf{X}$, we have nine different response vectors obeying the following equation:

$$\mathbf{y}^k = \mathbf{X}\boldsymbol{\beta}^k + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\beta}^k = (\underbrace{5, \ldots, 5}_{q = kp/10}, \overbrace{0, \ldots 0}^{p-q}) \text{ for } k = 1, \ldots, 9.$$

Effect of Sparsity



MSE as a function of sparsity

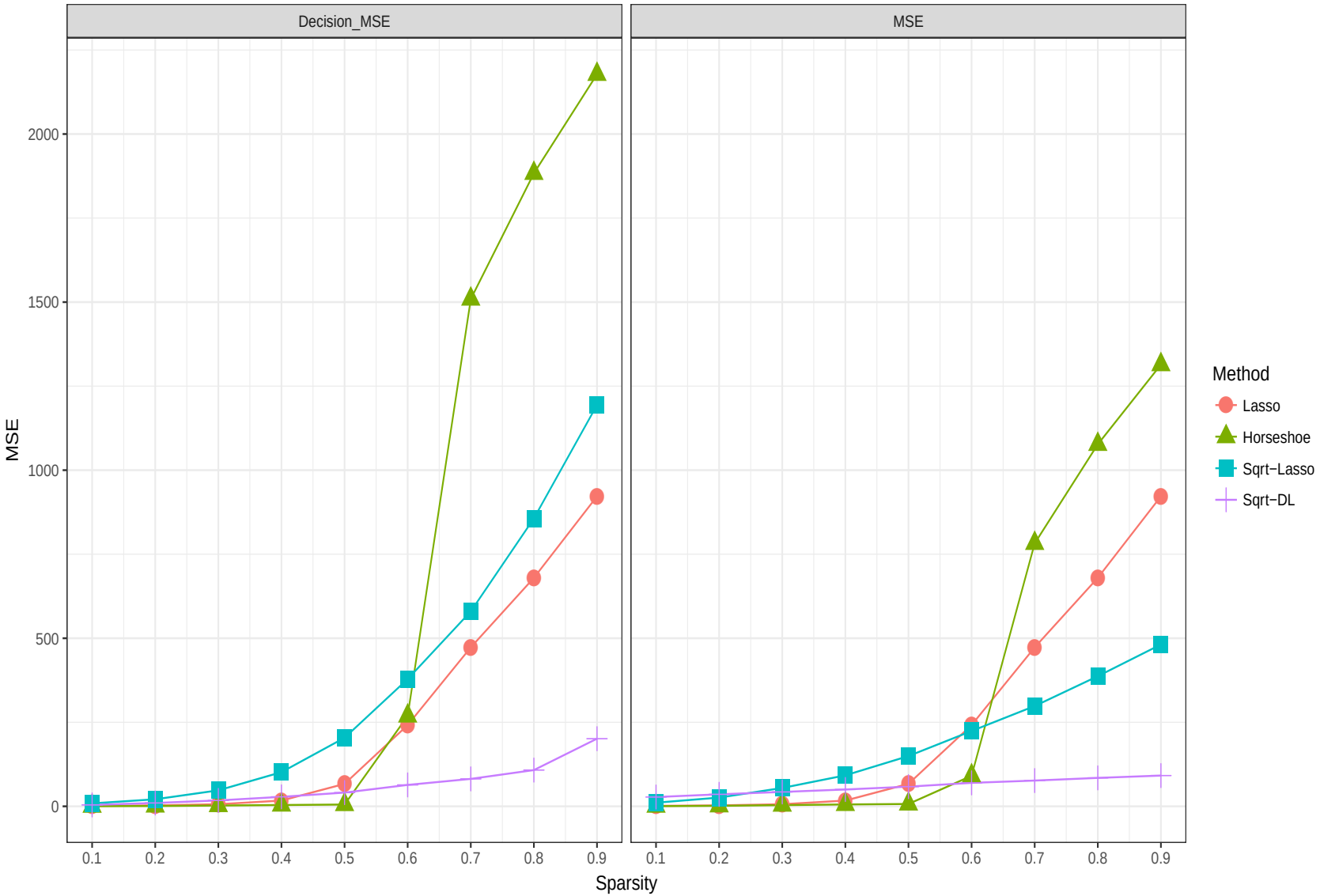


Figure: A comparison of the MSE as a function of sparsity level

Outline

Discussion

In this work, we first developed:

- a Bayesian representation of the $\sqrt{\text{Lasso}}$;
- using normal scale mixtures we developed a corresponding Gibbs sampler;
- unlike the horseshoe and G-L shrinkage priors, this method obviates the need to learn, scale or estimate the precision parameter σ;
- motivated by the strong properties G-L priors, added a global component to our model. This ensured, that the new prior placed sufficient mass around the origin, thus a priori favoring nearly black sets;
- yet we did not observe any improvement in concentration coverage, as the MSE stayed quite high in our empirical investigation.

Surprisingly, the effect of the added global parameter was a nice adaptability to sparsity levels. This new interesting property requires more theoretical investigation.

Effect of Sparsity on τ

We conducted a small experiment, where models with different proportions of non-zero parameters were constructed, and we implemented both the $\sqrt{\text{DL}}$ and the horseshoe.

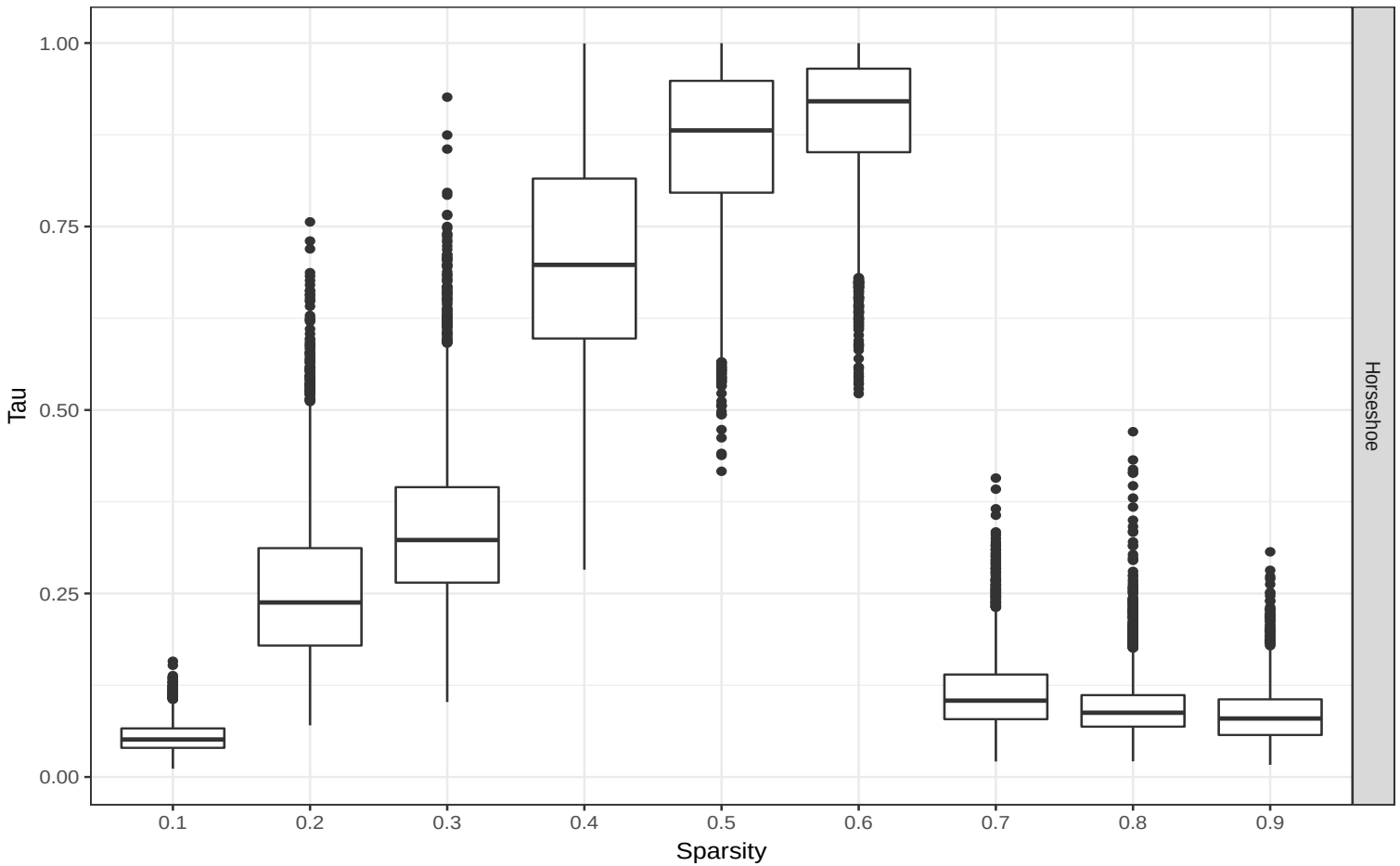


Figure: Evolution of τ in terms of sparsity level for the Horseshoe method

Effect of Sparsity on τ

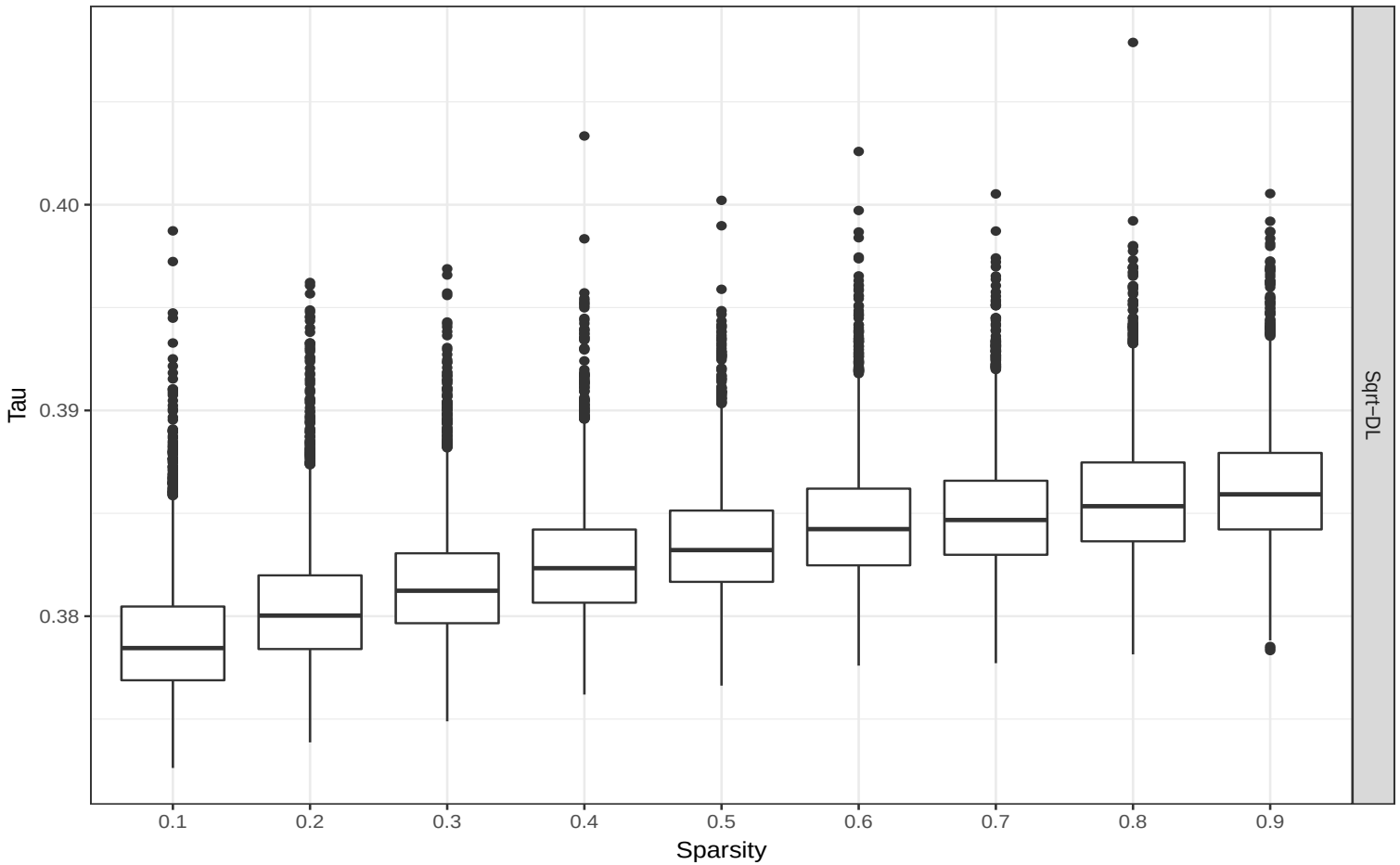


Figure: Evolution of τ in terms of sparsity level for the $\sqrt{\text{DL}}$ method

Computational Concerns

- Handling global parameters: full Bayes / MMLE / CV / m_{eff}.
- Gometric convergence of MCMC is not guaranteed.
- Rao-Blackwellization of global parameters.

Future Directions

- Theoretically investigate and prove adaptability to sparsity of $\sqrt{\text{DL}}$;
- Extend the methods developed here to the case of non-gaussian likelihood, such as count and categorical data.
- The Bayesian $\sqrt{\text{Lasso}}$ and $\sqrt{\text{DL}}$ will be investigated *vis-a-vis* other G-L priors tailored for Gaussian data when the error distribution is heavy-tailed, e.g. t-distribution.

Thanks!!