Adapting to Heavy Tails and Sparsity Bayesian $\sqrt{\text{Lasso}}$ and $\sqrt{\text{DL}}$

Mohamed Abdelkader Abba *
North Carolina State University

Anindya Bhadra†
Purdue University

Jyotishka Datta‡
University of Arkansas

Nicholas G. Polson §
University of Chicago
Booth School of Business

Brandon T. Willard¶
University of Chicago
Booth School of Business

September 30, 2018

Abstract

The Lasso and the Horseshoe, gold-standards in the frequentist and Bayesian paradigms, critically depend on learning the error variance. This causes a lack of scale invariance and adaptability to heavy-tailed data. The $\sqrt{\text{Lasso}}$ [Belloni et al., 2011] attempt to correct this by using the ℓ_1 norm on both the likelihood and the penalty for the objective function. In contrast, there is essentially no methods for uncertainty quantification or automatic parameter tuning via a formal Bayesian treatment of an unknown error distribution. Here we provide a fully Bayesian solution to these problems, called Bayesian $\sqrt{\text{Lasso}}$ and its extension $\sqrt{\text{DL}}$, that achieve scale invariance and robustness to heavy tails, while maintaining computational efficiency. The $\sqrt{\text{DL}}$ leads to uncertainty quantification by yielding standard error estimates and credible sets for the underlying parameters. Furthermore, the hierarchical model leads to an automatic tuning of the penalty parameter using a full Bayes or empirical Bayes approach, avoiding any ad-hoc choice over a grid. A surprising and useful result is that these methods adapt to the level of sparsity for the entire range of $\pi \in [0, 1]$ for nearly black objects, unlike the other global-local shrinkage priors. We provide an efficient Gibbs sampling scheme based on Normal scale mixture representation of Laplace densities. Performance on real and simulated data exhibit excellent small sample properties and we establish some theoretical guarantees.

1 Introduction

Global-local shrinkage priors have been established as the current state-of-the art inferential tool for sparse signal detection and recovery as well as the default choice for handling non-linearity in what have hitherto been paradoxical problems without a Bayesian answer. Despite these success stories, certain aspects of their behavior, such as performance in presence of correlated errors or adapting to unknown error distribution, remain unexplored. The aim of this paper is to offer insightful

*Address: N5109 SAS Hall, 2311 Stinson Dr.,Raleigh, NC 27695, email: mabba@ncsu.edu
†Address: 250 N. University St., West Lafayette, IN 47907, email: bhadra@purdue.edu.
‡Address: SCEN 309, 1 University of Arkansas, Fayetteville, AR, 72701, email: jd033@uark.edu.
§Address: 5807 S. Woodlawn Ave., Chicago, IL 60637, email: ngp@chicagobooth.edu.
¶Address: 5807 S. Woodlawn Ave., Chicago, IL 60637, email: bwillard@uchicago.edu.

solutions to these open problems motivated by the changing landscape of modern applications. The tools developed here aim at achieving scalability and strong theoretical support while focusing on their usefulness in current applications.

Feature selection, or selecting a subset of covariates, is pervasive in countless modern applications, specially those involving a 'wide' data set, where number of features (p) far exceed the number of samples (n). The most popular inferential problems are the sparse normal means problem: $(Y_i \mid \beta_i) \overset{\text{ind}}{\sim} \mathcal{N}(\beta_i, 1), i = 1, \ldots, n$, and the sparse linear regression: $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, $p \gg n$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where β is a 'nearly black object', that is, $\beta \in l_0[p_n] \equiv \{\beta : \#(\beta_i \neq 0) \leq p_n\}$, where $p_n = o(n)$. Current literature provides a rich variety of methodologies for high-dimensional inference based on regularization which implicitly or explicitly penalizes models based on their dimensionality. Convex penalties, such as the Lasso [Tibshirani, 1996], the elastic net [Zou and Hastie, 2005], or their variants, enjoy a number of advantages, such as uniqueness of solution, efficient computation and relatively straightforward theoretical analysis. The gold standard is Lasso (Least Absolute Shrinkage and Selection Operator) that produces a sparse point estimate by constraining the ℓ_1 norm of the parameter vector.

Regularized methods prevent over-fitting by implicitly or explicitly penalizing model dimensionality. 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 $Y = X\beta + \epsilon$, a regularized estimate of β is obtained by minimizing the penalized likelihood:

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

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

The gold-standard for regularized method is Lasso that simultaneously performs estimation and model selection by constraining the ℓ_1 norm of the underlying parameter vector, i.e. $\omega(\beta_j) = |\beta_j|$.

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

Bayesian Duality

The penalization approaches can be also explained from a Bayesian framework by interpreting the penalty as the logarithm of a suitable prior as follows:

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

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

The Bayesian correspondence leads to uncertainty quantification by yielding standard error estimates and credible sets for the underlying parameters and automatic tuning of the penalty parameter using a full Bayes or empirical Bayes approach, avoiding any ad-hoc choice over a grid. However, convex penalties such as Lasso yield a posterior that is 'useless for uncertainty quantification' [Castillo et al., 2015] and equivalent Bayesian hierarchical models are notably absent for the non-convex methods.

The popularity of global-local (G-L) shrinkage priors in the 'nearly-black' or 'ultra-sparse' regime, marked by parameter $\beta \in \ell_0[p_n]$, with $p_n \to 0$, is largely due their optimal theoretical and empirical performance. The key idea behind G-L priors is to use global shrinkage to adjust to the overall sparsity and local shrinkage to identify the strong signals. These priors avoid the computational bottle-neck of searching over an exponentially growing model space, which obstructs the spike-

and-slab prior [Mitchell and Beauchamp, 1988] on ultra-high dimensions. For the sparse normal means model $(y_i \mid \beta_i) \overset{\text{ind}}{\sim} \mathcal{N}(\beta_i, 1)$ for $i = 1, \ldots, n$, the horseshoe prior [Carvalho et al., 2010] is given by the hierarchical model:

$$(y_i \mid \beta_i) \sim \mathcal{N}(\beta_i, \sigma^2),\ (\beta_i \mid u_i, \tau) \sim \mathcal{N}(0, u_i^2\tau^2),\ u_i^2 \sim \mathrm{C}^+(0,1), \quad i = 1, \ldots, n.$$

Horseshoe prior operates by directly modeling the posterior inclusion probability $P(\beta_i \neq 0 \mid y_i)$ such that the probability concentrates near 0 or 1 for noise and signals, respectively. This follows from the linearity of posterior mean under the horseshoe prior that mimics a spike-and-slab model:

$$\mathbf{E}(\beta_i \mid y_i) = \{1 - \mathbf{E}(\kappa_i \mid y_i)\} y_i \text{ where } \kappa_i = 1/(1 + u_i^2\tau^2) \tag{3}$$

The U-shaped posterior is a direct outcome of putting a $\mathrm{Be}(1/2, 1/2)$ prior on the shrinkage coefficient κ_i, lending horseshoe its name. Since the inception of the horseshoe prior, many G-L priors have been proposed, focusing on the sparse normal means and regression problem. Some of the popular G-L priors include the Normal Exponential Gamma [Griffin and Brown, 2010], generalized double Pareto (GDP) [Armagan et al., 2013], the three-parameter beta [Armagan et al., 2011], the Dirichlet-Laplace [Bhattacharya et al., 2015] and the more recent spike-and-slab Lasso [Rovcková and George, 2016], horseshoe+ [Bhadra et al., 2016] and the R2-D2 [Zhang et al., 2016] priors.

Square-root Lasso

Despite the attractive features of Lasso, its performance in high-dimensional data is critically dependent on estimating the standard deviation σ of the noise ϵ, which remains a non-trivial problem in $p \gg n$ situation. The square-root Lasso, proposed by Belloni et al. [2011], is a modification of Lasso that eliminates the need for knowing σ, or pre-estimating it. The square-root Lasso is also independent of the Gaussianity or sub-gaussianity of noise. In fact, as Giraud [2014] points out, the Lasso estimate with ℓ_1 penalty is not scale-invariant in the sense that the invariance relation $\hat{\beta}(\sigma Y, X) = \sigma\hat{\beta}(Y, X)$ does not hold for all $\sigma > 0$. Since the standard deviation of noise ϵ is σ, one way of obtaining a scale-invariant penalized estimator is to set $\lambda^* = \lambda\sigma$ in (1), yielding:

$$\hat{\beta}^{\text{inv}} = \sigma^{-1} ||Y - X\beta||^2 + \lambda\Omega(\beta), \text{where, } \sigma = \text{sdev}(\epsilon) \tag{4}$$

Estimating σ by $||Y - X\beta|| / \sqrt{n}$ and using the ℓ_1 penalty $\Omega(\beta) = \|\beta\|_1$ leads to the $\sqrt{\text{Lasso}}$ estimator:

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

Clearly, the square-root Lasso estimator is scale-invariant and hence independent of the knowledge of σ, and still enjoys computational efficiency as the objective function is convex. The resulting estimator also enjoys near-oracle convergence rate, similar to Lasso, when $\text{supp}(\beta_0)$ has only s elements, $s < n$ [Belloni et al., 2011].

The square-root Lasso admits an alternative representation / algorithm, as another variant of Lasso called Scaled Lasso [Sun and Zhang, 2012], that establishes the connection between the original Lasso and the square-root Lasso. Following Giraud [2014], the square-root Lasso estimator in (5) and $\hat{\sigma} = ||Y - X\beta|| / \sqrt{n}$ can be written as solution to the convex system:

$$(\hat{\beta}, \hat{\sigma}) = \underset{\beta \in \Re^p, \sigma \in \Re^+}{\text{argmin}} \left\{ \frac{n\sigma}{2} + \frac{||Y - X\beta||_2^2}{2\sigma} + \lambda\|\beta\|_1 \right\} \tag{6}$$

Hence, we have the following relationship between Lasso and the square-root Lasso estimators:

$$\hat{\beta}_{\lambda}^{\sqrt{\text{lasso}}} = \hat{\beta}_{2\lambda\hat{\sigma}}^{\text{lasso}}, \quad \text{where} \quad \hat{\sigma} = ||Y - X\beta|| / \sqrt{n}$$

This implies that the square-root Lasso (or, scaled Lasso) can be efficiently calculated by a scheme that alternately finds a Lasso estimate $\hat{\beta}$ and $\hat{\sigma}$, resulting in the scaled-Lasso algorithm [Sun and Zhang, 2012].

Despite the attractive properties of these methods, there is a common caveat: the choice of tuning parameter λ. For Lasso, the tuning can be done either via a k-fold cross-validation or a complexity selection technique [Giraud et al., 2012]. However, these methods come with some concerns: while the k-fold CV works well empirically, it lacks theoretical support and the complexity selection is only guaranteed to work under Gaussianity of the data. The scale-invariant methods improve this situation slightly by making the tuning parameter free of σ, but it still requires tuning by adapting to the data.

Furthermore, it has been noted by some authors [Chatterjee and Lahiri, 2011] that the Lasso-based estimates do not yield meaningful standard errors for the parameter estimates, motivating full Bayesian treatment that produces reliable uncertainty quantification without extra effort. The Bayesian treatments of penalized regression depend on the useful duality of penalty and log-prior, and (Normal) scale mixture representation of the prior (e.g. Laplace as Normal-Gamma) that leads to efficient computation via EM/ECME or MCMC algorithms.

The main contribution in this paper is twofold. First, we provide a Bayesian interpretation of the square-root Lasso estimator based on the scale mixture representation of the Laplace density. Apart from quantifying uncertainty, this representation provides at least two alternative computational tools: via MCMC and via proximal algorithm [Polson et al., 2015]. We also offer new insights into the estimators behavior by investigating the resulting posterior distribution and the shrinkage weights. Next, we extend and generalize the Bayesian $\sqrt{\text{Lasso}}$ estimator with an appropriate local shrinkage term to the Bayesin $\sqrt{\text{DL}}$ estimator. The proposed estimator achieves better robustness compared to the popular G-L priors such as horseshoe in terms of (1) adapting to strong covariate dependence and (2) adapting to the level of sparsity in the data.

The rest of the paper is organized as follows: §2 describes the Bayesian square-root Lasso and the Bayesian square-root Dirichlet–Laplace estimator, §3 provides the computational strategies, i.e. the Gibbs sampler for the fully Bayesian implementation, §4 illustrates the unique posterior properties of the new priors that distinguishes it from the earlier shrinkage priors in its class. Finally, §6 provides some numerical examples to illustrate how the proposed method outperforms the existing G-L priors, and section 7 provides concludes with future directions.

2 Proposed Methodology: Bayesian $\sqrt{\text{Lasso}}$ and $\sqrt{\text{DL}}$

→ This is a repetition of the part before eq. (5)

Penalized regression methods such as Lasso are critically dependent on estimating the error variance σ^2 , which remains a non-trivial problem in high-dimensional $p \gg n$ situation. The square-root Lasso [Belloni et al., 2011] is a variant of Lasso that eliminates the need for knowing or pre-estimating σ and adapts to sub-Gaussian noise. The $\sqrt{\text{Lasso}}$ method uses a plug-in estimate of $\hat{\sigma} = ||Y - X\beta|| / \sqrt{n}$ in the Lasso optimization (1) to obtain :

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

↓ was λ before.

The resulting estimator enjoys near-oracle convergence rate, similar to Lasso, when $\text{supp}(\beta_0)$ has only s elements, $s < n$ as well as computational speed by dint of its convexity [Belloni et al.,

2011]. Moreover, unlike Lasso, the $\sqrt{\text{Lasso}}$ estimator is also scale invariant, i.e., $\hat{\beta}(\sigma Y, X) = \sigma\hat{\beta}(Y, X), \forall\sigma > 0$ [Giraud, 2014]. It turns out that the resulting estimator is identical to another regularization method, called the scaled Lasso [Sun and Zhang, 2012], which jointly optimizes β and σ.

$$(\hat{\beta}, \hat{\sigma}) = \underset{\beta\in\Re^p, \sigma\in\Re^+}{\text{argmin}} \left\{ n/2 + ||Y - X\beta||_2^2 / (2\sigma) + \tau\|\beta\|_1 \right\}$$

This is repetition of eq. (6)

Hence, we have the following relationship between Lasso and the square-root Lasso estimators:

$$\hat{\beta}_\lambda^{\sqrt{\text{lasso}}} = \hat{\beta}_{2\lambda\hat{\sigma}}^{\text{lasso}}, \quad \text{where} \quad \hat{\sigma} = ||Y - X\beta|| / \sqrt{n}$$

This implies that the square-root Lasso (or, scaled Lasso) can be efficiently calculated by a scheme that alternately finds a Lasso estimate $\hat{\beta}$ and $\hat{\sigma}$, resulting in the scaled-Lasso algorithm [Sun and Zhang, 2012].

Despite the attractive properties of these methods, there is a common caveat: the choice of tuning parameter λ. For Lasso, the tuning can be done either via a k-fold cross-validation or a complexity selection technique [Giraud et al., 2012]. However, these methods come with some concerns: while the k-fold CV works well empirically, it lacks theoretical support and the complexity selection is only guaranteed to work under Gaussianity of the data. The scale-invariant methods improve this situation slightly by making the tuning parameter free of σ, but it still requires tuning by adapting to the data.

repetition (see page 4)

Despite their desirable characteristics, $\sqrt{\text{Lasso}}$ has two major concerns. First, the choice of tuning parameter τ: one can use either a k-fold cross-validation or a complexity selection, but the former lacks theoretical support and the latter is restricted to Gaussian data [Giraud et al., 2012]. Second, inability to yield meaningful error estimates for the parameters by Lasso-based methods [Chatterjee and Lahiri, 2011]. To solve these issues, we propose a Bayesian $\sqrt{\text{Lasso}}$ that fully quantifies uncertainty and leads to efficient computation via MCMC.

repetition

(for all practical purposes, & the above part of τ_λ can be removed)

2.1 Hierarchical Model for Bayes $\sqrt{\text{Lasso}}$

Here we derive the Bayesian hierarchical model corresponding to the $\sqrt{\text{Lasso}}$ in (5). Since the likelihood-prior decompsoition of (5) yield a Laplace density for both the observation and the prior model, we use a Gaussian scale mixture representation of Laplace to write the Bayesian hierarchy. The key steps in the Bayesian hierarchy for $\sqrt{\text{Lasso}}$ follows from the well-known identity due to Lévy [1940] given by:

$$\int_0^\infty \frac{a}{(2\pi)^{1/2}t^{3/2}} \exp\{-a^2/(2t)\} \exp\{-\lambda t\} dt = \exp\{-a(2\lambda)^{1/2}\}. \tag{7}$$

The Levy identity (7) leads to the well-known normal scale mixture representation of Laplace density [Andrews and Mallows, 1974]. Let $Q(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2$. Using $a = 1$, and $2\lambda = Q(\beta)$ yields:

$$\exp\left[-||Y - X\beta||_2\right] = \int_0^\infty \frac{1}{(2\pi)^{1/2}t^{3/2}} \exp\{-1/(2t)\} \exp\{-Q(\beta)t/2\} dt \tag{8}$$

Alternatively, we can use $a = Q(\beta)$ and $\lambda = 1/2$ to obtain an equivalent decomposition:

$$\exp\left[-||Y - X\beta||_2\right] = \int_0^\infty \frac{1}{(2\pi)^{1/2}v} \exp\{-v^2/2\} \exp\{-Q(\beta)/2v^2\} dv^2 \tag{9}$$

To complete the hierarchy we use the normal scale mixture of Laplace prior on β as follows:

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

The hyper-parameter τ serves the role of the tuning parameter in square-root Lasso. There are several different ways of treating τ. We can treat it as a fixed tuning parameter and use pre-specified values on a grid to choose one. We can also either estimate τ via an empirical Bayes marginal maximum likelihood or use a suitable hyperprior on τ to learn via full Bayes. For the Bayesian Lasso, Park and Casella [2008] used a Gamma hyper-prior to make the tuning parameter a part of the Gibbs sampler.

$$\pi(\tau^2) = \frac{\delta^r}{\Gamma(r)} (\tau^2)^{r-1} e^{-\delta\tau^2}, \ \tau^2 > 0, \ (r > 0, \delta > 0). \tag{10}$$

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

$$f(\mathbf{y}, \boldsymbol{\beta}, v^2, \boldsymbol{\lambda}, \tau^2 \mid \mathbf{r}, \delta) \propto \frac{1}{(2\pi)^{1/2} v} e^{-v^2/(2)} \exp\{-\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2/v^2\}$$
$$\prod_{i=1}^{p} (\lambda_i^2)^{-\frac{1}{2}} \exp\{-\beta_i^2/(2\lambda_i^2)\} \frac{\tau^2}{2} e^{-\lambda_i^2\tau^2/2} (\tau^2)^{r-1} e^{-\delta\tau^2} \tag{11}$$

The joint distribution in (11) provides the full hierarchical model for a Bayesian treatment.

$$[\mathbf{y} \mid \boldsymbol{\beta}, v^2] \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, v^2\mathbf{I}) \tag{12}$$
$$[\boldsymbol{\beta} \mid \boldsymbol{\lambda}] \sim \mathcal{N}(\mathbf{0}, D_\lambda), \quad D_\lambda = \mathrm{Diag}(\lambda_1^2, \ldots, \lambda_p^2) \tag{13}$$
$$[\lambda_1^2, \ldots, \lambda_p^2 \mid \tau^2] \sim \prod_{j=1}^{p} \sim \frac{\tau^2}{2} e^{-\lambda_j^2\tau^2/2} d\lambda_j^2, \quad \lambda_j^2 > 0, \tag{14}$$
$$[v^2] \sim \mathit{Gamma}((n+1)/2, 1/2), \tag{15}$$
$$[\tau^2] \sim p(\tau^2) d\tau^2, \ \tau^2 > 0. \quad [\tau^2 \sim \mathcal{G}(r, \delta), \text{ or } \tau^2 \sim \mathcal{C}(0, 1).] \tag{16}$$

2.2 Adding a Global Component: $\sqrt{\text{Dirichlet–Laplace}}$

The use of local shrinkage priors in sparse models and high dimensional data settings has been investigated thoroughly by several authors. For example, [Castillo et al., 2015] have proved that local shrinkage priors do not achieve posterior contraction around the true model. Moreover, in Castillo et al. [2015], the authors explained that from a Bayesian perspective, this lack of concentration property, renders these priors useless. They defend their point of view by saying that poor concentration around true model values yields dishonest credible intervals, leading to poor uncertainty quantification. In this section, we will try to incorporate a global component into our model in order to improve its performance. The hierarchical model for Bayesian $\sqrt{\text{Lasso}}$ is given by (12):

$$\left.\begin{array}{ll} \beta_j & \overset{iid}{\sim} \mathcal{N}(0, \tau_j^2) \\ \tau_j^2 & \overset{iid}{\sim} \mathcal{E}\mathrm{xp}(\lambda^2/2) \end{array}\right\} \Rightarrow \boldsymbol{\beta} \sim DE(\lambda) \text{ and } \lambda^2 \sim \pi(\lambda). \tag{17}$$

The above parametrization lacks a global parameter that could adjust with signal strength. In fact, global-local shrinkage priors usually have the following Gaussian scale mixture representation:

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

where τ is a global scale parameter, controlling how large the β_j parameters are in general (i.e. a global shrinkage parameter), while the local standard deviation parameters ψ_j control how big the parameter is allowed to be locally. In a fully Bayesian approach, the priors for τ and the ψ_j are typically set to be independent. Recent works of van der Pas et al. [2017] shows that treating τ as fixed or using an empirical Bayes approach also leads to near-minimax rates of posterior contraction.

A better parametrization would be to constrain the ψ_j to lie on a simplex. This would then give us the interpretation that τ is the overall standard deviation if the covariates are properly scaled and the local parameters control how the individual parameters contribute to this variability. The standard parameterization leads to some confounding between the scales of the local and global parameters, which can lead to both an interpretational and computational problems. Interestingly, Bhattacharya et al. [2015] showed that in some specific cases you can go from a model where the local parameters are constrained to the simplex to the unconstrained case.

Moreover, it is easy to see from (17) that the joint prior distribution of the parameter vector, while easily tractable due to independence, does not place sufficient prior mass on sparse regions, since the double exponential density is bounded at zero. Recent choices of priors were motivated by this basic assessment. For example, the horseshoe prior was carefully formulated to yield a spike at zero accounting for sparsity as well as heavy tail property in order to recover strong signals. Here however, we chose to follow the ideas of [Bhattacharya et al., 2015], and model the full joint prior distribution of $\boldsymbol{\beta}$ on $\mathbf{R}^p$.

In what follows, let $\mathrm{DE}(\tau)$ denote a double exponential distribution where τ is the scale parameter, i.e. with density $f(x) = (2\tau)^{-1}\mathrm{e}^{-|x|/\tau}$. Also, we use the following form for the giG generalized inverse gamma distribution: $Y \sim \mathrm{giG}(\lambda, \rho, \chi)$ if $f(y) \propto y^{\lambda-1}\mathrm{e}^{-0.5(\rho y + \chi/y)}$ for $y > 0$.

Bhattacharya et al. [2015] proposed a completely different class of shrinkage priors. Instead of modeling the marginal distribution of the regression coefficients, they looked at the joint distribution. Recall that in 17, the joint prior distribution is p-dimensional DE with a single global scale τ. Bhattacharya et al. [2015] instead, introduced a vector of scales $(\phi_1\tau, \ldots, \phi_p\tau)$, where $(\phi_1, \ldots, \phi_p)$ is constrained to lie in the $(p-1)$ dimensional simplex $\mathcal{S}^{n-1} = \{\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p) : \phi_j \geq 0, \sum_{j=1}^{p} \phi_j = 1\}$ and is assigned a $\mathrm{Dir}(a, \ldots, a)$ prior. This prior choice under adequate values of a helps force a large subset of $\boldsymbol{\beta}$ to be simultaneously close to zero with high probability. The corresponding prior is hence :

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

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

In [Bhattacharya et al., 2015], the authors extensively studied the marginal properties of $\beta_j \mid \tau$, integrating out $\boldsymbol{\phi}$. The following proposition summarizes their findings.

Proposition 1. *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 $a < 1$.*

This property ensures that the Dirichlet-Laplace prior places enough mass around sparse vectors. Furthermore Bhattacharya et al. claimed that τ plays a critical role in determining the tails of the marginal distribution of β_j's. In a full Bayesian framework they recommend placing $\mathrm{Gamma}(pa, 1/2)$ prior on τ. Furthermore, using the representation of the DE distribution as a scale mixture of Gaussians:

$$\beta_j \mid \phi, \tau \sim \mathrm{DE}(\phi_j\tau) \Rightarrow \begin{cases} \beta_j & \sim \mathcal{N}(0, \psi_j\phi_j^2\tau^2); \\ \psi_j & \sim \mathcal{E}\mathrm{xp}(1/2), \end{cases}$$

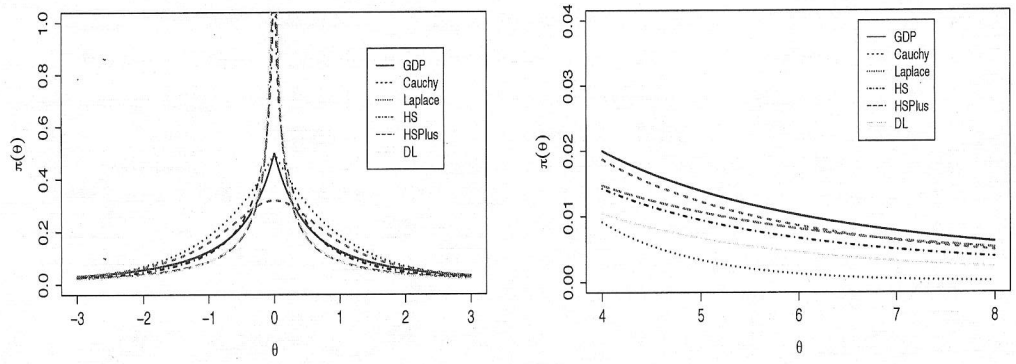


Figure 1: Marginal prior densities near the origin (left) and in the tail regions (right). The legends denote the horseshoe+ (HSPlus), horseshoe (HS), Dirichlet-Laplace (DL), generalized double Pareto (GDP), Cauchy and Laplace priors.

We get the augmented full hierarchical model :

$$[\mathbf{y} \mid \boldsymbol{\beta}, v^2] \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, v^2\mathbf{I}_n), \tag{18}$$

$$[\boldsymbol{\beta} \mid \boldsymbol{\phi}, \tau, \boldsymbol{\psi}] \sim \mathcal{N}(\mathbf{0}, D_{\boldsymbol{\psi}\boldsymbol{\phi}\tau}), \quad D_{\boldsymbol{\psi}\boldsymbol{\phi}\tau} = \text{Diag}(\psi_1\phi_1^2\tau^2, \ldots, \psi_p\phi_p^2\tau^2), \tag{19}$$

$$\psi_j \overset{iid}{\sim} \mathcal{E}\text{xp}(1/2), \tag{20}$$

$$\boldsymbol{\phi} \sim \text{Dir}(a, \ldots, a), \tag{21}$$

$$\frac{\tau}{v^2} \sim \mathcal{G}\text{amma}(pa, 1/2), \tag{22}$$

$$[v^2] \sim \mathcal{G}\text{amma}(\frac{n+1}{2}, 1/2). \tag{23}$$

As we can see from Fig. 1, both the Horseshoe and the DL_a exhibit a singularity near zero. This marginal behavior at the origin guarantees sufficient prior mass near zero in order to accommodate for nearly black vectors. Furthermore, in the lower panel of Figure 1, we see a comparison of the tails of the different shrinkage priors. Unlike horseshoe and DL_a, the Laplace prior does not have heavy tails that leave room for prior mass on possible high signal values. Hence we would expect the two former shrinkage priors to outperform the latter in both signal recovery and noise shrinkage.

3 Computation

3.1 Gibbs Sampler for $\sqrt{\text{Lasso}}$

Let $D_\lambda = \text{Diag}(\lambda_1^2, \ldots, \lambda_p^2)$ be the diagonal matrix of local shrinkage parameters. Using the equivalent decomposition (8), and collecting the terms for $\boldsymbol{\beta}$, the joint distribution can be re-written as

follows with $t = 1/v^2$:

$$f(\mathbf{y}, \boldsymbol{\beta}, t, \boldsymbol{\lambda}, \tau^2 \mid \mathbf{r}, \delta) \propto \frac{1}{t^{3/2}} \exp\{-1/(2t)\} \exp\left[-\frac{1}{2}\{\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}t + \mathbf{D}_{\lambda}^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}t\}\right] \prod_{i=1}^{p} (\lambda_i^2)^{-\frac{1}{2}} \frac{\tau^2}{2} e^{-\lambda_i^2\tau^2/2} (\tau^2)^{r-1} e^{-\delta\tau^2} \quad (24)$$

The full conditional distributions of $\boldsymbol{\beta}$ and τ are easy to derive: The full conditional of $\boldsymbol{\beta}$ is multivariate normal and τ is Gamma, exploiting the conjugacy. The parameters t and λ_i^2 follow inverse Gaussian distribution, where we assume the following parametric form of the inverse Gaussian density:

$$f(x \mid \lambda', \mu') = \sqrt{\frac{\lambda'}{2\pi}} x^{-3/2} \exp\left\{-\frac{\lambda'(x-\mu')^2}{2(\mu')^2 x^2}\right\}, \quad x > 0$$

The full conditional distributions needed for implementing a Gibbs sampler are:

$$\begin{aligned}
&\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\lambda}, t \sim \mathcal{N}\left(\mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}t, \mathbf{A}^{-1}\right), i = 1, \ldots, p, \\
&\text{where } \mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{X}t + \mathbf{D}_{\lambda}^{-1} \\
&t \mid \mathbf{y}, \boldsymbol{\beta} \sim \text{Inv-Gauss}\left(\mu' = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^{-1}, \lambda' = 1\right) \\
&\lambda_i^{-2} \mid \beta_i, \tau \sim \text{Inv-Gauss}(\mu' = |\frac{\tau}{\beta_i}|, \lambda' = \tau^2) \\
&\tau^2 \mid \boldsymbol{\lambda}, r, \delta \sim \text{Gamma}(p + r, \delta + \sum_{i=1}^{p} \lambda_i^2/2)
\end{aligned}$$

direct
Your sampler costs $O(p^3)$. Use Bhattacharya Biometrika sampler and cost is $O(pn^2)$.

A special case of the linear regression model is the sparse normal means model: $y_i = \beta_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, which results when the design matrix is equal to the identity matrix of appropriate dimension. The Gibbs sampler for the normal means model is identical to that for the linear regression, but faster as the full conditional distribution of β_i's are univariate Gaussian, and hence more efficient than the multivariate sampling.

$$\beta_i \mid y_i, \lambda_i, t \sim \mathcal{N}\left(y_i \frac{\lambda_i^2 t}{1 + t\lambda_i^2}, \frac{\lambda_i^2}{1 + t\lambda_i^2}\right), i = 1, \ldots, p. \quad (25)$$

3.2 Gibbs Sampler for $\sqrt{\text{DL}}$

The hierarchical model for $\sqrt{\text{DL}}$, given in (18) exploits the Laplace Gaussian scale mixture and leads to straightforward posterior computations. To reduce autocorrelation, we rely on a blocked Gibbs sampler scheme. The sampler moves from the following blocks (i) $[\boldsymbol{\beta} \mid \boldsymbol{\psi}, \boldsymbol{\phi}, \tau, \mathbf{v}^2, \mathbf{y}]$, (ii) $[\boldsymbol{\psi} \mid \boldsymbol{\phi}, \tau, \boldsymbol{\beta}]$, (iii) $[\boldsymbol{\phi} \mid \boldsymbol{\beta}]$, (iv) $[\tau \mid \boldsymbol{\phi}, \boldsymbol{\beta}, v^2]$, and (v) $[v^2 \mid \boldsymbol{\beta}, \tau, \mathbf{y}]$. Computing the full conditional distribution of the above blocks is standard and straightforward due to conjugacy except for the third block $[\boldsymbol{\phi} \mid \boldsymbol{\beta}]$. In their paper Bhattacharya et al. [2015], developed a very efficient sampling scheme for this non-trivial step. We state the following result from their paper, for a complete proof see [Bhattacharya et al., 2015].

Theorem 2. *The joint posterior of* $[\boldsymbol{\phi} \mid \boldsymbol{\beta}]$ *has the same distribution as* $(T_1/T, \ldots T_p/T)$, *where* T_j*'s are independently distributed according to a* gIG$a - 1, 1, 2|\beta_j|$, and $T = \sum_{j=1}^{p} T_j$.

← something looks strange

Using the representation in Theorem 2, we get the following blocked Gibbs sampler:

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

$$\left[\Sigma^{-1} = \frac{XX^T}{v^2} + \frac{D^{-1}_{\psi\phi^2}}{\tau^2}\right]$$

Again, you should use Bhattacharya sampler.

(ii) Conditional posterior of $[\psi \mid \phi, \tau, \beta]$ 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 $[\phi \mid \beta]$ by drawing $T_1, \ldots T_p$ 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 $[\tau \mid \phi, \beta, v^2]$ from a gIG$(pa - p, 1, 2\sum_{j=1}^{p} |\beta_j|/\phi_j)$ distribution.

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

4 Posterior Properties

Just one subsection? Then why is it necessary? Not sure.

4.1 Dependence on Error Variance σ

The key advantage of $\sqrt{\text{Lasso}}$, as pointed out by its authors [Belloni et al., 2011], is its ambivalence towards the error variance σ^2, resulting in an invariant estimator. It seems that these advantages would carry over to the Bayesian hierarchy as well. We illustrate this feature with a toy example borrowed from Polson and Scott [2010], created to warn against ignoring the dependence between τ and σ^2. The original example in Polson and Scott [2010] generated two observations with true mean 20, and considered the posterior under two different prior choices $\tau \sim \mathcal{C}^+(0, 1)$ (absolute scaling) and $\tau \sim \mathcal{C}^+(0, \sigma)$ (relative scaling) and showed that the posterior becomes bimodal under the absolute scaling prior. The authors argued that "the issue is one of averaging over uncertainty about σ in estimating the signal-to-noise ratio" – precisely what the $\sqrt{\text{Lasso}}$ aims to protect from.

We recreate this example in Fig. 2, with four different choices for handling the hyper-parameters σ^2 and τ^2:

1. $\tau \sim \mathcal{C}^+(0, 1)$ (absolute scaling), $\sigma \sim 1/\sigma^2$ (Jeffreys's).
2. $\tau \sim \mathcal{C}^+(0, \sigma)$ (relative scaling), $\sigma \sim 1/\sigma^2$ (Jeffreys's).
3. τ fixed, $\sigma \sim 1/\sigma^2$ (Jeffreys's), and
4. τ, σ fixed,

where, 'fixed' hyperparameters are estimated using an Empirical Bayes approach. The final candidate is the Bayesian $\sqrt{\text{Lasso}}$, which is free of σ, and we put a standard half-Cauchy prior on its global shrinkage term τ^2. Fig. 2 shows the posterior mode of $p(\beta \mid y)$ for the five different candidates. As expected, the horseshoe posterior concentrates near the true value for both the empirical Bayes approach and the relative scaling prior on τ, but shows bimodality for other choices. The Bayesian $\sqrt{\text{Lasso}}$ does not have a scale parameter σ to worry about, and it concentrates near $\beta = 20$ for the half-Cauchy prior on τ.

It should be emphasized that the argument in the above example is not to establish the superiority of $\sqrt{\text{Lasso}}$ over horseshoe, but rather to point out the importance of hyper-parameters in a Bayesian hierarchical model to scale to the unknown error variance. Admittedly, one can simply use an empirical Bayes approach to get rid of such undesired situations. However, the striking difference in the behaviour of the posterior densities in Fig. 2 suggests that the scaling of global parameters

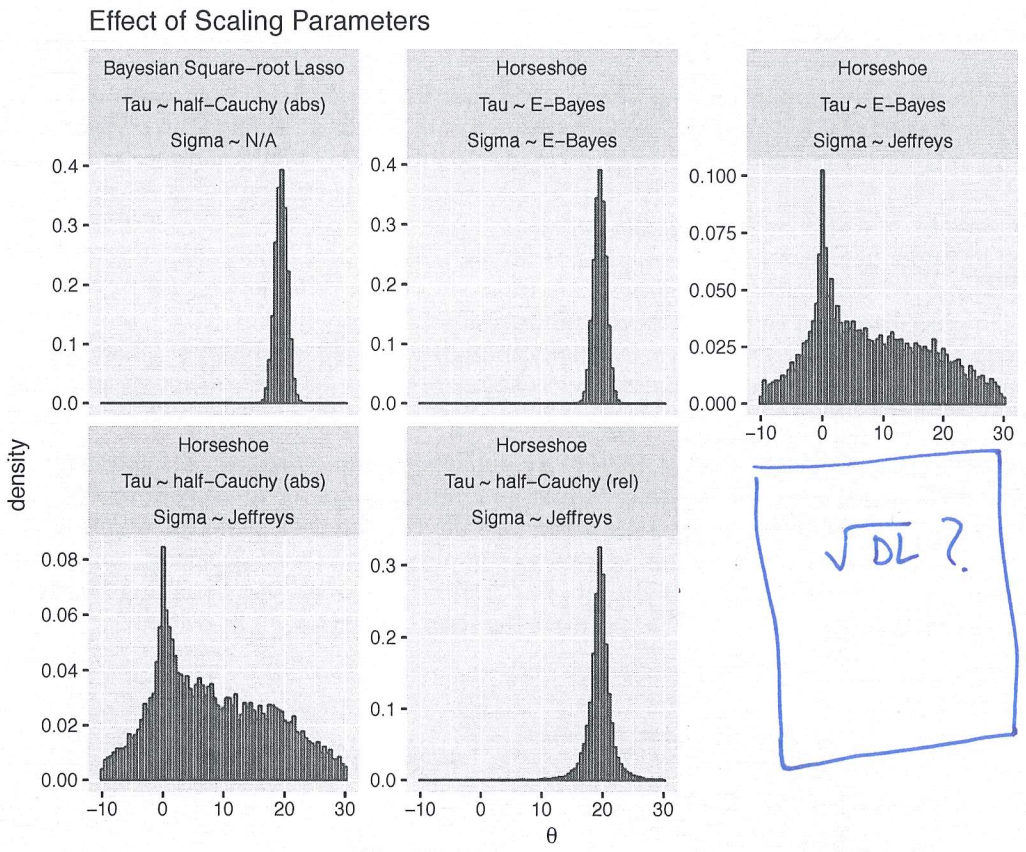


Figure 2: Behavior of the posterior density under different methods of handling the hyper-parameters σ^2 and τ for the Horseshoe prior as well as the Bayesian $\sqrt{\text{Lasso}}$ for a half-Cauchy prior on its global shrinkage parameter.

is a delicate issue, likely to be pervasive in all global-local shrinakge prior. The Bayesian $\sqrt{\text{Lasso}}$ escapes unharmed by its design to ignore σ.

5 Hyper-parameters

Handling the treatment of hyper-parameters can prove to dramatically affect the performance of Bayesian methods. For example in Figure 2, we showed how in a very simple setting the horseshoe estimator behaves very differently based on the method used to handle and estimate the global shrinkage parameter τ. Whether to use Empirical Bayes, Full Bayes and relative scaling or not are question we should address and discuss.

In addition, there has been a great amount of interest in the theoretical properties of the blocked Gibbs sampler and their convergence properties. In fact, Bayesian shrinkage methods almost all rely on a blocked Gibbs sampler scheme to explore the parameter spaces. Rajaratnam et al. [2017] and Pal and Khare [2014] studied the performance of and properties of Gibbs samplers in the context of Bayesian shrinkage for regression. While they proved geometric ergodicity, they pointed out that more often than not the samples obtained from these samplers usually present high auto-correlation and the chain suffers from slow convergence, and proposed ways to overcome these problems.

5.1 Computational Issues

As we have seen in the previous sections, the use of scale mixtures of normals to represent otherwise non-conjugate priors on the regression coefficients is a common feature of Bayesian shrinkage models. Usually, this data augmentation procedure leads to a three step Gibbs sampler to sample from the intractable joint posterior. A first step for the regression coefficients β, a second for the variance parameter σ, and a last step for the augmented parameter (here we regroup the augmented as well as hyper-parameters of the model). Although, Khare and Hobert [2013] and Pal and Khare [2014] proved geometric ergodicity of the three step Gibbs sampler for the Bayesian Lasso and the Dirichlet-Laplace prior. It has been pointed out in Rajaratnam et al. [2017], that convergence of these sampler can be rather slow specially in high-dimensional settings. Given that the *"large p small n"*, is precisely the setting where these methods are used to overcome model complexity, computational issues in such settings would present a problematic drawback.

To address this bottleneck, Rajaratnam et al. [2017] rely on *blocking* and *collapsing*. A Gibbs sampler is said to be collapsed if the joint posterior is marginalized over one or more parameters to reduce sampling steps. This often increases convergence rate, but the new posterior might not be tractable and any gain would then be lost in a more complicated scheme. Blocking, requires grouping multiple parameters together and jointly sampling them in one step. Grouping highly correlated parameters, is generally expected to improve the convergence rate of the MCMC.

In their paper, Rajaratnam et al. [2017] consider the case of the Bayesian Lasso, where the prior distribution on the parameters is exactly the same as in (17), except for the prior placed on the precision parameter. The hierarchical model is given by :

$$\begin{aligned}
[\mathbf{y} \mid \beta, \sigma^2] &\sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}) \\
[\beta \mid \tau, \sigma^2] &\sim \mathcal{N}(\mathbf{0}, \sigma^2 D_\tau), \quad D_\tau = \text{Diag}(\tau_1^2, \ldots, \tau_p^2) \\
[\tau_1^2, \ldots, \tau_p^2 \mid \lambda^2] &\sim \prod_{j=1}^{p} \sim \frac{\lambda^2}{2} e^{-\tau_j^2\lambda^2/2} \mathrm{d}\tau_j^2, \quad \tau_j^2 > 0, \\
[\sigma^2] &\sim \frac{1}{\sigma^2}, \quad \sigma^2 > 0, \\
[\lambda^2] &\sim p(\lambda^2)d\lambda^2, \ \lambda^2 > 0. \quad [\lambda^2 \sim \mathcal{G}(r, \delta), \text{ or } \lambda^2 \sim \mathcal{C}(0,1)].
\end{aligned} \tag{26}$$

The corresponding Gibbs sampler is :

$$\begin{aligned}
[\beta \mid \tau, \mathbf{y}, \sigma^2] &\sim \mathcal{N}(\mathbf{A}_\tau^{-1}\mathbf{X}^t\mathbf{y}, \sigma^2\mathbf{A}_\tau^{-1}), \quad \text{where } \mathbf{A}_\tau = \mathbf{X}^t\mathbf{X} + \mathbf{D}_\tau^{-1} \\
[\frac{1}{\tau_j^2} \mid \beta, \sigma, \lambda^2] &\sim \text{Inv} - \text{Gaussian}\left(\sqrt{\frac{\nu^2\sigma^2}{\text{fi}_j^2}}, \nu^2\right) \leftarrow \text{check} \\
[\sigma^2 \mid \mathbf{y}, \beta, \tau] &\sim \mathcal{IG}\left(\frac{n+p-1}{2}, \frac{||\mathbf{y} - \mathbf{X}\beta||_2^2 + \beta^t\mathbf{D}_\tau^{-1}\beta}{2}\right)
\end{aligned} \tag{27}$$

The above three-step Gibbs sampler, while straight-forward and easy to implement, converges very slowly in high-dimensional settings. Rajaratnam et al. [2017] demonstrated that this problem arises mainly due to the high a posteriori dependence between β and σ^2. And following this, they were able to group these parameters in one step through the following result. For the proof see Rajaratnam et al. [2017].

Lemma 3. *In model* (26) $[\sigma^2 \mid \mathbf{y}, \tau]$ *has the inverse gamma distribution with shape* $(n-1)/2$ *and scale parameter* $\mathbf{y}^t\left(\mathbf{I}_n - \mathbf{X}\mathbf{A}_\tau^{-1}\mathbf{X}^t\right)\mathbf{y}/2$.

Using the above result they constructed a sampler in only two steps, first $(\beta, \sigma^2) \mid \tau$ and then

$\tau \mid (\beta, \sigma^2)$. The new collapsed Gibbs sampler is ergodic and as tractable as the original one. Convergence is considerably faster and they also observe low samples auto-correlation in their numerical comparisons.

5.2 Handling the global shrinkage parameter

As we have discussed in subsection 2.3.2 and particularly through the example borrowed from Polson and Scott [2010] the dependence between τ and σ^2 if not addressed properly might lead to unsatisfactory results. This problem, is expected to prevail in all global-local shrinkage priors, and in our case adding a global component to the model we observed the same behavior with absolute scaling.The golden rule here is to always scale global precision parameters. This is only one of the many questions that are often ignored, although greatly affect the performance of Bayesian hierarchical models. van der Pas et al. [2017] studied in depth the performance of the horseshoe prior and how different treatments for τ affect the theoretical properties of the estimators in the sparse normal means problem. They determined that the global shrinkage parameter τ is very important towards the minimax contraction rate. Also, van der Pas et al. [2014] showed that τ can be interpreted as the proportion of non-zero parameters up to a logarithmic factor.

In the full Bayes approach case, van der Pas et al. [2017] specified conditions under which the prior choice on τ results in near minimax contraction rate. Under their conditions, the prior must be truncated to the left by $1/p$ among other conditions. This led to a wide use of a truncated Cauchy distribution on this hyper-parameter. They also show that the posterior credible set are honest, in the sense that they concentrate around nearly black balls in case of a sparse normal means problem. One immediate application of this later result, is to use these sets not only as a tool of uncertainty quantification, but also an ad-hoc variable selection or hypothesis testing procedure. In fact, one could just look at the $(1-\alpha)$ CI for each parameter and decide whether it's a signal or noise. They also point out that any non zero parameter has to exceed a certain threshold magnitude in order to be recovered. That is, for any $\beta \leq \sqrt{2\log(n/p_n)}$, where $\mathbf{p}_n$ is the number of true non-zero parameters, the CI are not useful in a Bayesian sense.

Moreover, one could argue that with an emipirical Bayes procedure for the global shrinkage parameter, there would be no need to worry about scaling or hyper-prior distribution choice. However, as van der Pas et al. [2017] and Datta and Ghosh [2013] point out, an empirical Bayes estimate of τ might possibly degenerate to zero, yielding improper parameter posterior distributions. This happens mostly when the model fails to identify the level of sparsity. With some conditions on sparsity level and signal magnitude, van der Pas et al. [2014] and van der Pas et al. [2016] showed the plug-in MMLE (Marginal Maximum Likelihood Estimate) of τ guarantees near minimax concentration rate.

On the other hand, Datta and Ghosh [2013] studied, the oracle properties of another decision rule. In their paper they considered the shrinkage weight $1-\kappa_i(\tau) = \hat{\beta}_i(\tau)/\mathbf{y}_i$ and proved that this multiple testing rule is Bayes Optimal, under similar conditions to van der Pas et al. [2014] in both a full Bayes or an empirical Bayes procedure on τ. They also emphasize the risk of possible degeneracy of the empirical Bayes estimate.

6 Simulation Studies

(The organization of section 6 is slightly strange. 6.1 & 6.2 are setting you consider. Not sure if they require separate subsections.)

In this section, we investigate the performance of the methods developed in section 2. The goal is to test the finite sample properties of our methods and compare them with the common procedures available. First, we will start with the Normal Means problem, then we will look at the more complex scenario of high dimensional regression. In the later case, we study the effect of certain conditions on the design matrix that have been proven to affect model selection consistency in some methods.

6.1 High Dimensional regression

Another important area of application of shrinkage priors, is high dimensional regression and particularly model selection. In this section we generate our data following the linear regression model:

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

where β is a $p \times 1$ vector of model coefficients and is assumed to be sparse, $\mathbf{y}$ is an $n \times 1$ response vector, and $\mathbf{X}$ is an $n \times p$ design matrix. Hence some of the regression coefficients are exactly zero and they correspond to irrelevant predictors (columns of $\mathbf{X}$). In our simulations, we compare the performance of the Bayesian $\sqrt{\text{Lasso}}$, to the frequentist Lasso and the Horseshoe. In terms of penalized regression Lasso has been extensively studied and proven consistent under some conditions, also it is the most widely method in penalized regression. Likewise, the horseshoe prior has received much attention from Bayesian practitioners and has also been proven to yield consistent results under mild assumptions. In comparing our methods with these two known and commonly used procedures, we will be able to judge their performances as well as notice their particular shortcomings and advantages. Moreover, simulation studies give a rather deep insight about the behavior of new methods, and allow us to investigate both favorable settings and scenarios where poor and unsatisfactory results arise. Often, it is with the study of these simulation that the first theoretical aspect are noticed, hence they give valuable directions and information both for theoretical and practical purposes.

Recall that the advantage of $\sqrt{\text{Lasso}}$ is its ambivalence to the error variance σ^2, and since our hierarchical model given by (5) is but a representation of the $\sqrt{\text{Lasso}}$ penalty, we would expect our proposed Bayesian representation to work under large values of σ^2, hence accommodating sub-Gaussian errors and heavy tailed data.

6.2 Variable Selection for Shrinkage Priors

The problem of variable selection and particularly in a *"small n, large p"* setting has received quite some attention both from the frequentist an Bayesian perspective. The Lasse, $\sqrt{\text{Lasso}}$, horseshoe, Dirichlet-Laplace and numerous other methods were in part developed to tackle this problem. While the frequentist methods usually yield a sparse estimate, that is the estimated $\hat{\beta}$ vector has entries that are exactly zero, their Bayesian counterparts always require a decision rule to classify an estimated coefficient as either zero or not. As we discussed in section **2.6.2**, decision rules for the horseshoe have already been studied, and shown to be optimal under some conditions for the normal means problem and regression settings where the design matrix is orthogonal. Likewise, for the Bayesian shrinkage priors considered in this work, we need a method to decide whether a coefficient should be classified as signal or noise. Such decision rule, can also be viewed as a variable selection step, given that any covariate for which the coefficient has been classified as zero is thrown out of the model. In this work, we decided to look at the posterior sample means of the $\hat{\beta}$ vector, and apply a k-means clustering on $|\hat{\beta}_j|$ with only two cluster centers. We expect two clusters centers, one concentrated around zero for the noise signals and one away from zero. This method is motivated by the assumption that the true parameter vector is generated according to a two groups model. That is, each β_i is generated from :

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

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 $\hat{\beta}^{dec} = \left\{\hat{\beta}_j^{dec}\right\}_{1 \leq j \leq p}$.

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

In this work, we stick to the case of the two groups model. However, the k-means method can be extended to a wider class of models. In fact, Li and Pati [2017] suggested a sequential 2-means clustering algorithm in case the model presents signals of varying strength level.

One of the many interesting questions that arise with variable selection, is *model selection consistency*. This property essentially means that the method used consistently selects the true model. It should be emphasized that model selection consistency and estimator consistency are entirely two different properties. Recall that estimator consistency holds if and only if:

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

while model selection consistency requires:

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

Some authors have also considered sign consistency which is a stronger version of the later requirement, where not only the zeros have to be matched but also the sign of each component estimate. Also, an estimate with wrong signs could be misleading.

6.3 Effect of Irrepresentability Condition

The optimality properties of Lasso are well-known and they depend on "neighbourhood stability" or "irrepresentability" condition and "beta-min" condition. Informally, these conditions guarantee against ill-posed design matrix and separability of signal and noise parameters. We show here a small simulation study inspired from Zhao et al.[2006] to show that the effect of 'irrepresentability condition' is not as strong on our methods as it is on the Lasso.

We describe the "irrepresentable" condition below:

Suppose, the sample covariance matrix is denoted by $\hat{\Sigma} = nX^TX$ 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

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

where $\hat{\Sigma}_{s_0,s_0}$ is a $s_0 \times s_0$ matrix corresponding to the active variables and so on. The irrepresentable condition for variable selection consistency of Lasso is:

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

This condition is sufficient and almost necessary in the sense that the necessary condition is only slightly weaker than the sufficient condition. The necssary condition requires '$\leq$ 1', while the

sufficient condition involves $\leq \theta$ for some $0 < \theta < 1$. The irrepresentable condition fails to hold if the design matrix is too ill-posed, i.e. has multi-collinearity.

[Bühlmann and van de Geer, 2011] warn the readers that the irrepresentable condition may fail even though the design matrix is not ill-posed and it might restrict what can be done in high-dimensional problems. Zhao et al. (2006) provide numerical example to show the effect of the irrepresentable condition on the variable selection performance of Lasso. They showed that the probability of selecting the true sparse model is an increasing function of the irrepresentability condition number, defined as

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

In particular, the probability of Lasso selecting the true model is almost 1 when $n_\infty > 0.2$ and it is almost zero when $\eta_\infty < -0.3$.

We simulated data with $n = 100, p = 60$ and $q = 7$ with the sparse coefficient vector $\beta_q^* = (7,5,5,4,4,3,3)^T$, σ^2 was set to 5 to allow for heavy tailed data. Like Zhao et al. (2006) we first draw the covariance matrix Σ from $Wishart(p, I_p)$ and then generate design matrix X from $N(0, \Sigma)$. This design is repeated a 100 times, 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. For the three Bayesian methods we run the Markov Chain for 9000 samples, discarding the first 1000 thousand as a burn-in step and finally thinning every two samples. We 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 in Subsection 6.2. The Lasso automatically yields sparse vector estimates, we only need to select the tuning parameter λ, which represents the penalty level, we set λ to the value that minimizes the MSE based on a 10 fold cross validation.

The goal of this simulation study is to observe the effect of the irrepresentability condition on our proposed methods and compare them to the Lasso and horseshoe. We are particularly interested in model selection consistency, so we look at the proportion of correctly selected models out of the 100 replicates for each design.

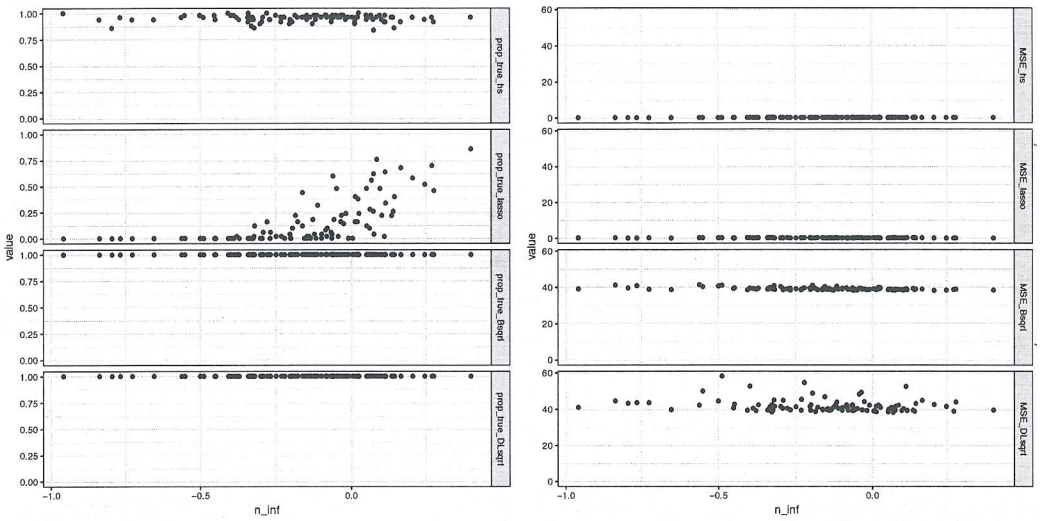
Zhao and Yu [2006] showed that the irrepresentability condition may not hold for such a design matrix. In fact, in our simulation studies the η_∞'s for the 100 simulated designs were between $[-1.02, 0.36]$. We expect the Lasso to perform well when $\eta_\infty > 0$ and poorly when $\eta_\infty < 0$. We generate $n = 100$ design matrices and for each design, 100 simulations were conducted by generating the noise vector from $\mathcal{N}(0, \sigma^2 I)$.

Figure 3a below shows the percentage of correctly selected model as a function of the irrepresentable condition number, η_∞ for Lasso,the Horseshoe prior, the Bayesian-$\sqrt{\text{Lasso}}$ and the $\sqrt{\text{DL}}$.

As expected, Lasso's variable selection performance is crucially dependent on the irrepresentability condition but the Horseshoe prior almost always recovers the true sparse β vector irrespective of η_∞. Strikingly, both our methods succeed in always recovering the true model. This strong performance independently of η_∞, clearly presents an advantage and is worth studying from a theoretical view point.

Given that the values of the non-zero entries of the true β, do not differ much in magnitude, we think that this excellent performance in terms of variable selection, is in part due to the 2-means clustering procedure.

In Fig. 3b, we see the evolution of the MSE calculated over the replicated samples for each design matrix. A surprising results, that we observe, is the pronounced difference from the model selection summary in figure 3a. Note, however that the MSE computed here is for the posterior median before any decision rule has been applied. Since the Bayesian methods do not provide exact sparse estimates, there will be an added error component wise across the whole β vector. Having a mod-



(a) Effect of η_∞ on model selection

(b) Effect of η_∞ on the MSE

Figure 3: Effect of Irrepresentability Condition on model selection and estimation error for different shrinkage estimators

erately large p, thus increases this error proportionally. Interestingly, the lasso despite not selecting the true model in approximately all designs, shows a very low MSE. This can be explained in part by the exact zero estimator yielded by this method, coupled with the sparse nature of the underlying true vector. Horseshoe, remarkably performs very well in terms of both true model selection and low MSE. Given that it is a Bayesian method, hence returning no exact zero entries, its impressive performance indicates both a very low bias for all components as well as small empirical variance of the posterior median.

It is also important to note how this simulation study provides a clear separation and points out the difference between model selection consistency and estimation consistency. As pointed out in the beginning of this section, the two properties are different and somehow counter-intuitively none of them implies the other. In Fig. 3b we see how lasso has very low MSE, which suggest estimator consistency, while Fig. 3a clearly shows that lasso does not enjoy model selection consistency. Conversely, both $\sqrt{\text{Lasso}}$and $\sqrt{\text{DL}}$ successfully capture the true model independently from η_∞, but at the same time shows high values of MSE.

6.4 Adapting to Sparsity levels

Most penalized regression methods, and shrinkage priors operate under the assumption that the parameter of interest is sparse in some sense. In addition, the widespread attention that these methods have received in the past decade was mostly focused on theoretical properties in the case of sparse models. Although, sparsity or parsimony of statistical models is crucial for their proper interpretations, as in sciences and social sciences, we should address the cases where true coefficient vectors, have zero entries but are not completely sparse. Furthermore, the case of nearly black vectors has been investigated thoroughly, yet little attention has been given to adaptability to varying degrees of sparsity. In this section, we try to address this issue by running simulations on model designs with varying underlying levels of sparsity. Like the previous section, we will compare our methods to Lasso the gold standard for best subset selection of predictors, and the horseshoe prior which is a state-of-the-art Bayesian estimator for sparse signals. We will focus on misclassification probability and MSE as indicators of method performance. We limit ourselves to the case of two group generating model for model parameters.

We simulated data with $n = p = 100$, the entries of the design matrix $\mathbf{X}$ were simulated from a $\mathcal{N}(0,2)$ distribution, with the error variance set to $\sigma^2 = 5$. We sampled 100 different design matrices, and for each of these design matrices, we 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:

$$\mathbf{y}^k = \mathbf{X}\beta^k + \epsilon, \text{ where } \beta^k = (\underbrace{5,\ldots,5}_{q=kp/10}, \overbrace{0,\ldots 0}^{p-q}) \text{ for } k = 1,\ldots,9. \tag{28}$$

Hence for each sparsity level, we have a 100 replicates, from which we compute the misclassification proportion, that is the number of times a given method does not select the true model, and the MSE. Here we also compute for the Bayesian methods, the MSE after the decision rule was performed MSE($\hat{\beta}^{dec}$). Table 1, 2 and 3 summarize the numerical results for varying levels of sparsity.

Table 1: Misclassification proportion according to sparsity

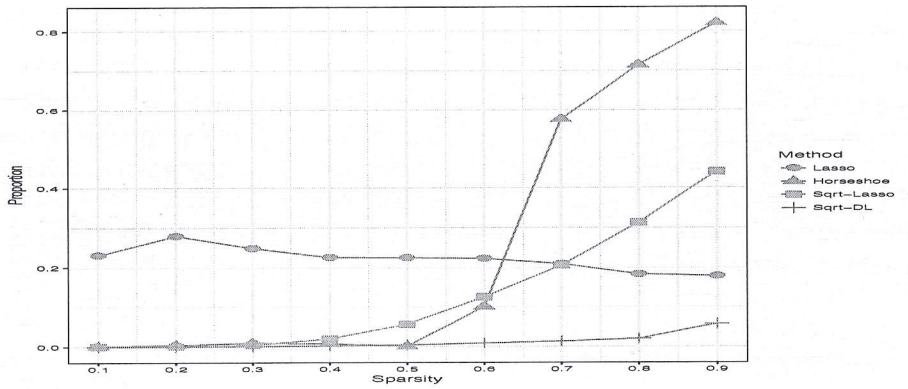
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Lasso	0.2306	0.2786	0.2481	0.2250	0.2241	0.2225	0.2087	0.1829	0.1782
Horseshoe	0.0013	0.0041	0.0094	0.0077	0.0011	0.1015	0.5747	0.7140	0.8210
B-$\sqrt{\text{Lasso}}$	0.0000	0.0006	0.0040	0.0201	0.0569	0.1261	0.2054	0.3124	0.4414
$\sqrt{\text{DL}}$	0.0000	0.0004	0.0008	0.0022	0.0043	0.0091	0.0134	0.0201	0.0577

From Table 1, we see that Lasso never selects the true model, and on average misses 20% of the coefficients. The horseshoe does well when the sparsity level is very low, this is well in accordance with the theoretical results for horseshoe's performance in the case of nearly black vectors. The Bayesian $\sqrt{\text{Lasso}}$, does almost as well as the horseshoe in terms of misclassification probability in the case of sparse parameters. However, both methods seem to break down when the proportion of non-zero parameters increases, i.e. when we shift from a sparse regime to a dense regime. The $\sqrt{\text{DL}}$ escapes this problem and seems totally oblivious to sparsity level. This method almost pinpoints the true model in all cases. Figure 4a, gives a better comparison than the above table, we can see clearly how the misclassification proportion for the horseshoe and Bayesian $\sqrt{\text{Lasso}}$are affected by sparsity levels, and how $\sqrt{\text{DL}}$ adapts easily to that level. This suggests that the added global shrinkage parameter in $\sqrt{\text{DL}}$ successfully adapts to the sparsity level of the β vector.

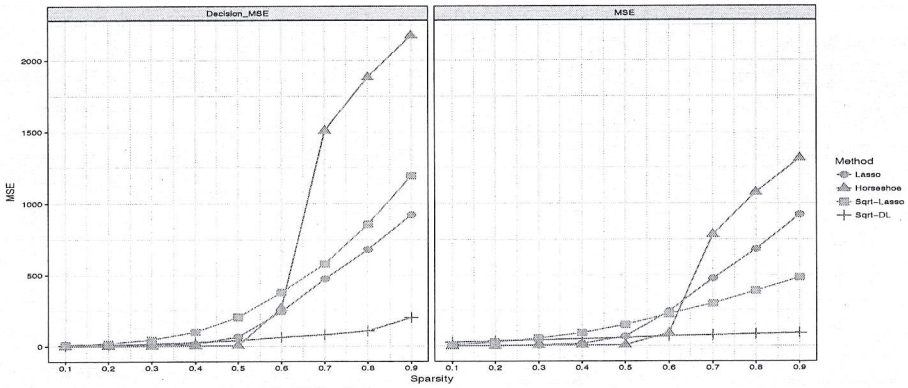
Likewise, from Tables 2 and 3 , we can see how the MSE for all four methods is affected by sparsity level. Note that for Table 2, the MSE was computed after a classification step was applied to the original Bayesian estimates, that is the estimates will have zero values for parameters that were not classified as signals. This is similar to the half-thresholding estimates of ?, but here we apply the idea to a general shrinkage estimator. Lasso and horseshoe exhibit a better MSE in sparse settings, but their MSE increases drastically, whenever the number of non-zero parameters increases and the sparsity level crosses a certain threshold. The Bayesian-$\sqrt{\text{Lasso}}$, and the $\sqrt{\text{DL}}$ have higher MSE values, however after classifying the parameters the MSE decreases significantly.

Table 2: MSE according to sparsity after applying a decision rule

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Lasso	1.29	2.69	6.18	17.07	66.76	241.71	472.8	679.67	922.11
Horseshoe	0.33	0.95	2.31	3.83	5.27	268.69	1509.02	1883.04	2177.61
B-$\sqrt{\text{Lasso}}$	8.68	20.95	47.49	101.59	203.84	378.51	579.83	856.5	1193.86
$\sqrt{\text{DL}}$	4.1	9.97	17.5	28.2	40.83	63.65	81.59	107.83	201.22



(a) Misclassification proportion as a function of sparsity level.



(b) A comparison of the MSE as a function of sparsity level

Figure 4: Effect of changing sparsity levels on model selection and estimation

Table 3: MSE according to sparsity

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Lasso	1.29	2.69	6.18	17.07	66.76	241.71	472.8	679.67	922.11
Horseshoe	0.5	1.49	3.55	5.56	6.96	90.73	782.25	1076.87	1313.92
B-$\sqrt{\text{Lasso}}$	10.74	26	54.59	92.16	149.51	223.96	298.3	387.07	481.41
$\sqrt{\text{DL}}$	27.63	35.65	42.76	50.06	58.48	69.78	76.6	84.56	91.42

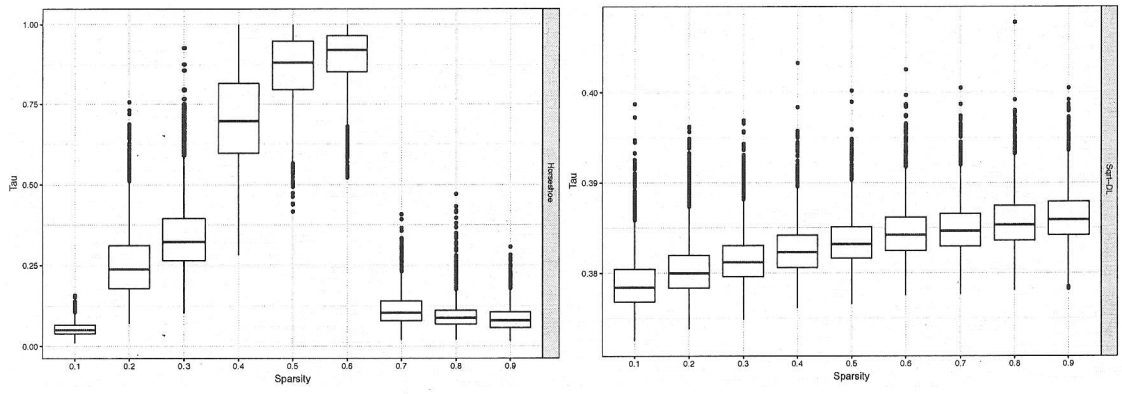
7 Discussion and Future Directions

In this work, we first developed a Bayesian representation of the $\sqrt{\text{Lasso}}$, taking advantage of normal scale mixture representation of the Laplace prior, we were able to develop a Gibbs sampler for the parameters of the model. Numerical results showed satisfactory performance in the case of sparse normal means and high dimensional regression. Furthermore, unlike the horseshoe and other global local shrinkage priors, this method obviates the need to learn, scale or estimate the precision parameter σ. We also found that a k-means classification step on the posterior estimates of the parameter vector outperforms other decision rules like the use of credible intervals for horseshoe.

Motivated by the strong properties of global local shrinkage priors, specifically, their singularities at zero and their ability to concentrate at near minimax rate, we 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. To show how the global parameter adapts to sparsity level, 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. Here we are only interested in the effect of different sparsity levels on τ. In Fig. 5a, we see how in the case of the horseshoe the boxplots for the τ samples continue to increase until we reach level of approximately .5 where a dramatic breakdown happens. Clearly, in the left side of the figure, we can say that τ follows the monotone increase in the proportion of non-zero parameters, but when this proportion approaches and exceeds the .5 threshold, the method is no longer able to follow and adapt the sparsity level. This behavior also explains why the MSE and the proportion of misclassified β's exploded whenever the proportion of non-zero β's exceeded the threshold of 0.5, as shown in Fig 4a and Fig. 4b.

On the other hand, we also saw in Fig 4a and Fig. 4b, how the $\sqrt{\text{DL}}$ performance remained satisfactory and was in no way affected by changes in the sparsity level. The boxplots of τ samples in Fig. 5b, back up our conjecture, unlike the horseshoe, here the global shrinkage parameter follows and learns correctly the degree of sparsity. In the future, we would like to theoretically investigate this claim and try to prove it.



(a) Evolution of τ on terms of sparsity level for the Horseshoe method

(b) Evolution of τ on terms of sparsity level for the $\sqrt{\text{DL}}$ method

Figure 5: Adapativity of τ for different shrinkage priors.

The methods developed in this work, only addressed the case where data can be modeled through a Gaussian likelihood. While any continuous type response variable can be somehow transformed to fit this class, the same cannot be said of count or categorical type data. In [Datta and Dunson, 2016], the authors developed a new class of continuous local-global shrinkage priors tailored for sparse counts. One of the future aims of this work, is to extend our methods in order to accommodate discrete data structures.

References

D.F. Andrews and C.L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 36(1):99–102, 1974. ISSN 00359246. doi: 10.2307/2984774.

Artin Armagan, Merlise Clyde, and David B Dunson. Generalized beta mixtures of Gaussians. In *Advances in Neural Information Processing Systems*, pages 523–531, 2011.

Artin Armagan, David B Dunson, and Jaeyong Lee. Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119–143, 2013.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon Willard. The Horseshoe+ Estimator of Ultra-Sparse Signals. *Bayesian Analysis*, to appear, 2016.

Anirban Bhattacharya, Debdeep Pati, Natesh S. Pillai, and David B. Dunson. Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110:1479–1490, 2015.

Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data*. Springer-Verlag Berlin Heidelberg, 2011.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480, 2010.

Ismail Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, October 2015. URL `http://dx.doi.org/10.1214/15-AOS1334`.

A. Chatterjee and S. N. Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011. doi: 10.1198/jasa.2011.tm10159. URL `http://dx.doi.org/10.1198/jasa.2011.tm10159`.

Jyotishka Datta and David B Dunson. Bayesian inference on quasi-sparse count data. *Biometrika*, 103(4):971–983, 2016.

Jyotishka Datta and Jayanta K Ghosh. Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis*, 8(1):111–132, 2013.

Christophe Giraud. *Introduction to high-dimensional statistics*, volume 138. CRC Press, 2014.

Christophe Giraud, Sylvie Huet, Nicolas Verzelen, et al. High-dimensional regression with unknown variance. *Statistical Science*, 27(4):500–518, 2012.

Jim E Griffin and Philip J Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

Kshitij Khare and James P. Hobert. Geometric ergodicity of the bayesian lasso. *Electronic Journal of Statistics*, 7:2150–2163, 2013.

Paul Lévy. Sur certains processus stochastiques homogènes. *Compositio mathematica*, 7:283–339, 1940.

Hanning Li and Debdeep Pati. Variable selection using shrinkage priors. *Computational Statistics & Data Analysis*, 107:107–119, 2017.

T. J. Mitchell and J. J. Beauchamp. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023–1032, December 1988. ISSN 01621459. URL `http://dx.doi.org/10.2307/2290129`.

Subhadip Pal and Kshitij Khare. Geometric ergodicity for bayesian shrinkage models. *Electronic Journal of Statistics*, 8(1):640–645, 2014.

Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. URL `http://amstat.tandfonline.com/doi/abs/10.1198/016214508000000337`.

Nicholas G Polson and James G Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538, 2010.

Nicholas G Polson, James G Scott, and Brandon T Willard. Proximal algorithms in statistics and machine learning. *Statistical Science*, 30(4):559–581, 2015.

Bala Rajaratnam, Doug Sparks, Kshitij Khare, and Liyuan Zhang. Scalable bayesian shrinkage and uncertainty quantification in high-dimensional regression. 2017.

Veronika Rovcková and Edward I George. The spike-and-slab lasso. *Journal of the American Statistical Association*, (just-accepted), 2016.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58:267–288, 1996.

SL van der Pas, BJK Kleijn, and AW van der Vaart. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8:2585–2618, 2014.

Stéphanie van der Pas, Jean-Bernard Salomond, and Johannes Schmidt-Hieber. Conditions for Posterior Contraction in the Sparse Normal Means Problem. *Electronic Journal of Statistics*, 10: 976–1000, 2016.

Stéphanie van der Pas, Botond Szabó, van der Vaart, and Aad. Adaptive posterior contraction rates for the horseshoe. *arXiv:1702.03698*, 2017.

Yan Zhang, Brian J Reich, and Howard D Bondell. High Dimensional Linear Regression via the R2-D2 Shrinkage Prior. *arXiv preprint arXiv:1609.00046*, 2016.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.