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

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

July 9, 2017

## 1 Introduction

Regularized methods have become widely popular as a inferential tool in high-dimensional data, owing to the popularity of Lasso [Tibshirani, 1996] and many of its variants [Tibshirani, 2014]. Regularized methods prevent overfitting by 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:

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

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

The gold-standard for regularized method is Lasso that simultaneously performs estimation and model selection by constraining the $\ell_1$ norm of the underlying parameter vector, i.e. $\omega(\beta_j) = |\beta_j|$.

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

[*]*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.

Lasso enjoys both computational efficiency, due to LARS [Efron et al., 2004] and coordinate descent [Friedman et al., 2007], as well as theoretical optimality properties [Bühlmann and van de Geer, 2011]. Bickel et al. [2009] have shown that the Lasso estimator achieves near-orcale property in recovering the true $\boldsymbol{\beta}_0$, under Gaussianity and certain design matrix conditions, up to a factor of $\sqrt{\log(2p)}$: yielding a $\sqrt{\log n}$ rate when $p$ grown polynomially as $n$. However, Lasso's performance in high-dimensional data is critically dependent on estimating the standard deviation $\sigma$ of the noise $\epsilon$, 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 $\sigma$, 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 $\ell_1$ penalty is not scale-invariant in the sense that the invariance relation $\hat{\boldsymbol{\beta}}(\sigma Y, X) = \sigma\hat{\boldsymbol{\beta}}(Y, X)$ does not hold for all $\sigma > 0$. Since the standard deviation of noise $\epsilon$ is $\sigma$, one way of obtaining a scale-invariant penalized estimator is to set $\lambda^* = \lambda\sigma$ in (1), yielding:

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

Estimating $\sigma$ by $||Y - X\boldsymbol{\beta}|| / \sqrt{n}$ and using the $\ell_1$ penalty $\Omega(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ leads to the $\sqrt{\text{Lasso}}$ estimator:

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

Clearly, the square-root Lasso estimator is scale-invariant and hence independent of the knowledge of $\sigma$, and still enjoys computational efficiency as the objective function is convex. The resulting estimator also enjoys near-oracle convergence rate, similar to Lasso, when $\text{supp}(\boldsymbol{\beta}_0)$ 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 $\hat{\sigma} = ||Y - X\boldsymbol{\beta}|| / \sqrt{n}$ can be written as solution to the convex system:

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

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

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

This implies that the square-root Lasso (or, scaled Lasso) can be efficiently calculated by a scheme that alternately finds a Lasso estimate $\hat{\boldsymbol{\beta}}$ 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 $\lambda$. 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 $\sigma$, 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.

Our main contribution in this paper is 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 behaviour by investigating the resulting posterior distribution and the shrinkage weights. The rest of the paper is organized as follows: §2 describes the Bayesian square-root Lasso, §3 provides some numerical examples and performance on a real data set, §4 provides some new theoretical results for the sparse regression problem and §5 concludes with future directions.

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

### 2.1 Hierarchical Model

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 step in the Bayesian hierarchy for $\sqrt{\text{Lasso}}$ follows from the well-known identity due to

Lévy [1940] given by:

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

For deriving the full hierarchical model for $\sqrt{\text{Lasso}}$, we use $a = 1$, and $2\lambda = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ to get

$$\exp\left[-\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\}^{1/2}\right] = \int_0^\infty \frac{1}{(2\pi)^{1/2}t^{3/2}} \mathrm{e}^{-1/(2t)} \mathrm{e}^{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})t} dt \tag{8}$$

To complete the hierarchy we use the normal scale mixture of Laplace prior on $\boldsymbol{\beta}$ [Andrews and Mallows, 1974] as follows:

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

The hyper-parameter $\tau$ serves the role of the tuning parameter in square-root Lasso. We can either estimate it via an empirical Bayes marginal maximum likelihood or use a Gamma hyperprior on $\tau$ to learn via full Bayes.

$$\pi(\tau^2) = \frac{\delta^r}{\Gamma(r)} (\tau^2)^{r-1} e^{-\delta\tau^2}, \; \tau^2 > 0, \; (r > 0, \delta > 0). \tag{9}$$

Under the Gamma hyper-prior, the joint distribution of $y_i$ and all the hyperparameters in the model is :

$$f(\mathbf{y}, \boldsymbol{\beta}, t, \boldsymbol{\lambda}, \tau^2 \mid \mathbf{r}, \delta) \propto \frac{1}{t^{3/2}} \exp\{-1/(2t)\} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})t\right\} \prod_{i=1}^{p} (\lambda_i^2)^{-\frac{1}{2}} \exp\{-\beta_i^2/(2\lambda_i^2)\} \frac{\tau^2}{2} e^{-\lambda_i^2\tau^2/2} (\tau^2)^{r-1} e^{-\delta\tau^2} \tag{10}$$

The joint distribution in (11) provides the full hierarchical model for a Bayesian treatment. To maintain notational similarity, we first transform $t/2 = \sigma^{-2}$ and write the reprametrized joint distribution as:

$$f(\mathbf{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda}, \tau^2 \mid \mathbf{r}, \delta) \propto \frac{1}{\sqrt{\sigma^2}} \exp\{-\sigma^2/(4)\} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/\sigma^2\right\} \prod_{i=1}^{p} (\lambda_i^2)^{-\frac{1}{2}} \exp\{-\beta_i^2/(2\lambda_i^2)\} \frac{\tau^2}{2} e^{-\lambda_i^2\tau^2/2} (\tau^2)^{r-1} e^{-\delta\tau^2} \tag{11}$$

Hence the hierarchical representation of the full model is as follows:

$$\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}) \tag{12}$$

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, D_\lambda),\ D_\lambda = \text{Diag}(\lambda_1^2, \ldots, \lambda_p^2) \tag{13}$$

$$\lambda_1^2, \ldots, \lambda_p^2 \sim \prod_{j=1}^{p} \sim \frac{\tau^2}{2} e^{-\lambda_i^2 \tau^2/2},\ \lambda_i^2 > 0, \tag{14}$$

$$\tau^2, \sigma^2 \sim (\tau^2)^{r-1} e^{-\delta\tau^2} \exp\{-\sigma^2/(4)\},\ \tau^2, \sigma^2 > 0. \tag{15}$$

## 2.2 Gibbs Sampler

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

$$f(\mathbf{y}, \boldsymbol{\beta}, t, \boldsymbol{\lambda}, \tau^2 \mid \mathbf{r}, \delta) \propto \frac{1}{t^{3/2}} \exp\{-1/(2t)\} \exp\left[-\frac{1}{2}\{\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}t + \mathbf{D}_\lambda^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}t\}\right] \prod_{i=1}^{p} (\lambda_i^2)^{-\frac{1}{2}} \frac{\tau^2}{2} e^{-\lambda_i^2\tau^2/2} (\tau^2)^{r-1} e^{-\delta\tau^2} \tag{16}$$

The full conditional distributions of $\boldsymbol{\beta}$ and $\tau$ are easy to derive: The full conditional of $\boldsymbol{\beta}$ is multivariate normal and $\tau$ is Gamma, exploiting the conjugacy. The parameters $t$ and $\lambda_i^2$ 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:

$$\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\lambda}, t \sim \mathcal{N}\left(\mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}t, \mathbf{A}^{-1}\right), i = 1, \ldots, p, \tag{17}$$

$$\text{where } \mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{X}t + \mathbf{D}_\lambda^{-1} \tag{18}$$

$$t \mid \mathbf{y}, \boldsymbol{\beta} \sim \text{Inv-Gauss}\left(\mu' = {}^1/_{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|}, \lambda' = 1\right) \tag{19}$$

$$\lambda_i^2 \mid \beta_i, \tau \sim \text{Inv-Gauss}(\mu' = |\frac{\beta_i}{\tau}|, \lambda' = \beta_i^2) \tag{20}$$

$$\tau^2 \mid \boldsymbol{\lambda}, r, \delta \sim \text{Gamma}(n + r, \delta + \sum_{i=1}^{p} \lambda_i^2/2) \tag{21}$$

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 $\beta_i$'s are univariate Gaussian, and hence more efficient than the multivariate sampling.

$$\beta_i \mid y_i, \lambda_i, t \sim \mathcal{N}\left(y_i \frac{\lambda_i^2 t}{1 + t\lambda_i^2}, \frac{\lambda_i^2}{1 + t\lambda_i^2}\right), i = 1, \ldots, p. \tag{22}$$

## 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.

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

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

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

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.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. URL http://projecteuclid.org/euclid.aos/1083178935.

Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, and others. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007. URL http://projecteuclid.org/euclid.aoas/1196438020.

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.

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

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

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.

Robert J Tibshirani. In praise of sparsity and convexity. *Past, Present, and Future of Statistical Science*, pages 497–505, 2014.