Variational Bridge Regression

Artin Armagan
Department of Statistical Science
Duke University
Durham, NC 27708
artin@stat.duke.edu

Abstract

Here we obtain approximate Bayes inferences through variational methods when an exponential power family type prior is specified for the regression coefficients to mimic the characteristics of the Bridge regression. We accomplish this through hierarchical modeling of such priors. Although the mixing distribution is not explicitly stated for scale normal mixtures, we obtain the required moments only to attain the variational distributions for the regression coefficients. By choosing specific values of hyper-parameters (tuning parameters) present in the model, we can mimic the model selection performance of best subset selection in sparse underlying settings. The fundamental difference between MAP, *maximum a posteriori*, estimation and the proposed method is that, here we can obtain approximate inferences besides a point estimator. We also empirically analyze the frequentist properties of the estimator obtained. Results suggest that the proposed method yields an estimator that performs significantly better in sparse underlying setups than the existing state-of-the-art procedures in both $n > p$ and $p > n$ scenarios.

1 INTRODUCTION

Consider the familiar linear regression model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{y}$ is an n-dimensional vector of responses, $\mathbf{X}$ is the $n \times p$ dimensional design matrix and $\boldsymbol{\varepsilon}$ is an n-dimensional vector of independent noise variables which are normally distributed, $\mathcal{N}_p\left(\mathbf{0}, \sigma^2\mathbf{I}_p\right)$ with variance σ^2. In what follows, assume that a subset of the regression coefficients, $\boldsymbol{\beta}$, is zero indicating that the corresponding regressors do not contribute to the response in the underlying model.

Consider the bridge regression (Frank and Friedman 1993) which results from the following regularization problem:

$$\min_{\boldsymbol{\beta}} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)' \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} |\beta_j|^{\gamma} \qquad (1)$$

where $\gamma = \{0, 1, 2\}$ result in familiar and well-studied solutions, i.e. best subset selection, lasso and ridge estimators respectively.

Our particular focus is on the set $\gamma = (0, 1)$. It is also well-known that the values within this set will lead to nonconcave minimization problems which may render hard to deal with. On the other hand, such penalty regions provide much sparser solutions than the lasso where $\gamma = 1$ (Tibshirani 1996).

A Bayesian solution can be obtained by placing appropriate priors on the regression coefficients that will mimic the effects of the Bridge penalty. As is very well-known, this choice of prior would be an independent exponential power density on each of the coefficients. A closed form solution is not possible with a normal likelihood since we lose the quadratic structure of the prior in the kernel of the density. However, as it has been studied, this family of distributions for $0 < \gamma < 2$ can be represented as scale mixtures of normals which makes a prior of such form possible with analytical ease through some hierarchical modeling (Andrews and Mallows 1974, West 1984, 1987). A very well-known example of this is the double-exponential distribution (yielding the lasso solution) which can be modeled as a mixture of normals with an exponential distribution as the mixing distribution. Although not related to this family, it is again well-known that a Student's t distribution can be obtained as a mixture

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

of normals where the mixing distribution is gamma. Although a normal mixture representation is possible for the exponential power family for $0 < \gamma < 2$, a recognizable mixing distribution may not be obtained for $\gamma \neq 1$ in a readily available form which will present a problem in Bayesian hierarchical modeling. The hierarchical structures studied by (Figueiredo 2003) and (Park and Casella 2008) to achieve the lasso solution in the Bayesian paradigm make use of such an explicitly stated mixing distribution to mimic the effects of a double-exponential prior. Despite the unavailability of such explicitly stated mixing distributions, we may be able to exploit the mixture formulation under the variational Bayes framework by only extracting the required moments.

2 THE MODEL

Here a normal likelihood is assumed, $\mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, with independent priors on the regression coefficients of the form $p(\beta_j|\lambda) \propto \exp\{-\lambda|\beta_j|^\gamma\}$, $\lambda > 0$, $0 < \gamma < 1$ and a typical conjugate prior on the error precision $\sigma^{-2} \sim \mathcal{G}(c_0, d_0)$. It is obvious to those who are familiar with the simplest Bayesian models that the exponential power family prior is not a conjugate prior with the normal likelihood for $\gamma \neq 2$.

Following (Andrews and Mallows 1974, West 1984), the above family of distributions can be written as

$$p(\beta_j|\lambda, \gamma) = \int_0^\infty \mathcal{N}(0, \tau_j^{-1}\lambda^{-2/\gamma}) f(\tau_j) d\tau_j, \quad (2)$$

where $\mathcal{N}(\alpha, \varrho)$ denotes a normal pdf with mean α and variance ϱ, $f(\tau_j) \propto \tau^{-1/2} q(\tau_j)$, and $q(\tau_j)$ is the density of the stable distribution of index $\gamma/2$.

If we proceed by placing independent normal priors on the regression coefficients $p(\beta_j|\tau_j, \lambda, \gamma) = \lambda^{1/\gamma}\sqrt{\tau_j/2\pi}\exp(-\tau_j\lambda^{2/\gamma}\beta_j^2/2)$ and regard $f(\tau_j)$ as a hyper-prior on τ_j, we obtain

$$\frac{\lambda^{1/\gamma}}{2\Gamma(1+1/\gamma)}\exp\left(-\lambda|\beta_j|^\gamma\right) = \int_0^\infty p(\beta_j|\tau_j, \lambda, \gamma) f(\tau_j) d\tau_j, \quad (3)$$

where $\Gamma(.)$ denotes the Gamma function.

Let us first introduce the variational framework and then derive the approximate marginal distributions for the parameters.

2.1 VARIATIONAL INFERENCE

The marginal likelihood of the observed data in (1) or in many other non-trivial models cannot be obtained analytically. Yet the integral can easily be approximated via the variational methods (Jordan et al. 1999). We can decompose the marginal likelihood conditional on λ and γ following (Bishop 2006). Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau})$. Given a λ and a γ value, we have

$$\begin{aligned}\log p(\mathbf{y}|\lambda, \gamma) \quad &= \quad \underbrace{\int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}|\lambda, \gamma)\log\frac{p(\boldsymbol{\theta}, \mathbf{y}|\lambda, \gamma)}{q(\boldsymbol{\theta}|\lambda, \gamma)}d\boldsymbol{\theta}}_{\mathcal{L}_{\lambda,\gamma}} \\ &\quad \underbrace{-\int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}|\lambda, \gamma)\log\frac{p(\boldsymbol{\theta}|\mathbf{y}, \lambda, \gamma)}{q(\boldsymbol{\theta}|\lambda, \gamma)}d\boldsymbol{\theta}}_{KL(q\|p)}\end{aligned} \quad (4)$$

where $\mathcal{L}_{\lambda,\gamma}$ is referred to as *the lower-bound* on the marginal likelihood and $KL(.\|.)$ denotes the Kullback-Leibler divergence between two distributions. Since this quantity is a strictly non-negative one and is equal to 0 only when $p(\boldsymbol{\theta}|\mathbf{y}, \lambda, \gamma) = q(\boldsymbol{\theta}|\lambda, \gamma)$, the first term in (4) constitutes a lower-bound on $\log p(\mathbf{y}|\lambda, \gamma)$. It is evident that maximizing the first term in the right hand side of (4) is equivalent to minimizing the second term in the right hand side, suggesting that $q(\boldsymbol{\theta}|\lambda, \gamma)$ is an approximation to the posterior density $p(\boldsymbol{\theta}|\mathbf{y}, \lambda, \gamma)$.

Following (Bishop and Tipping 2000) we consider a factorized form

$$q(\boldsymbol{\theta}|\lambda, \gamma) = \prod_i q_i(\boldsymbol{\theta}_i|\lambda, \gamma), \quad (5)$$

where $\boldsymbol{\theta}_i$ is a sub-vector of $\boldsymbol{\theta}$. Maximizing the lower bound with respect to $q_i(\boldsymbol{\theta}_i|\lambda, \gamma)$ yields

$$q_i(\boldsymbol{\theta}_i|\lambda, \gamma) = \frac{\exp\langle\log p(\mathbf{y}, \boldsymbol{\theta}|\lambda, \gamma)\rangle_{j\neq i}}{\int_{\boldsymbol{\Theta}_i}\exp\langle\log p(\mathbf{y}, \boldsymbol{\theta}|\lambda, \gamma)\rangle_{j\neq i}d\boldsymbol{\theta}_i}, \quad (6)$$

where $\langle . \rangle_{j\neq i}$ denotes the expectation with respect to the distributions $q_j(\boldsymbol{\theta}_j)$ for $j \neq i$. As we will see, due to the conjugate structure we will obtain for our model via the scale normal mixture representation of the exponential power prior, these expectations will be easily evaluated. Thus the procedure will consist of initializing the required moments and cycling through them by updating the distributions given by (6).

2.2 APPROXIMATE POSTERIORS

Following the solution given in (6) and the aforementioned normal mixture representation of the exponential power distribution, we will obtain the approximate marginal posterior distributions for the regression coefficients and the error variance. Although, due to the unknown mixing distribution $f(\tau_j)$, we cannot get an explicit expression for $q(\tau_j)$, fortunately we can evaluate $\langle\tau_j\rangle$ which is required for $q(\boldsymbol{\beta}|\lambda, \gamma)$.

The approximate marginal posterior distributions of the regression coefficients and the error precision are given by

$$
\begin{aligned}
q(\boldsymbol{\beta}|\lambda,\gamma) &\overset{d}{=} \mathcal{N}\left(\widehat{\boldsymbol{\beta}},\Sigma_{\boldsymbol{\beta}}\right) && (7)\\
q(\sigma^{-2}|\lambda,\gamma) &\overset{d}{=} \mathcal{G}\left(\widehat{c},\widehat{d}\right) && (8)
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \langle\sigma^{-2}\rangle\Sigma_{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y}\\
\Sigma_{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\langle\sigma^{-2}\rangle+\mathbf{T}\right)^{-1}\\
\mathbf{T} &= \lambda^{2/\gamma}diag(\langle\tau_j\rangle)\\
\widehat{c} &= n/2+c_0\\
\widehat{d} &= \frac{1}{2}\mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{X}\langle\boldsymbol{\beta}\rangle+\frac{1}{2}\sum_{i=1}^{n}\mathbf{x}_i\langle\boldsymbol{\beta}\boldsymbol{\beta}'\rangle\mathbf{x}_i'+d.
\end{aligned}
$$

The above-mentioned moments are obvious except for $\langle\tau_j\rangle$ since we do not have an explicit form available for its approximate distribution.

$$
\begin{aligned}
\langle\boldsymbol{\beta}\rangle &= \widehat{\boldsymbol{\beta}} && (9)\\
\langle\boldsymbol{\beta}\boldsymbol{\beta}'\rangle &= \boldsymbol{\Sigma}_{\boldsymbol{\beta}}+\langle\boldsymbol{\beta}\rangle\langle\boldsymbol{\beta}'\rangle && (10)\\
\langle\sigma^{-2}\rangle &= \widehat{c}/\widehat{d} && (11)
\end{aligned}
$$

Following (6), the approximate marginal distribution of τ_j can be written as

$$
q(\tau_j|\lambda,\gamma)=\frac{\exp\left(\langle\log p(\beta_j|\tau_j,\lambda,\gamma)\rangle+\langle\log f(\tau_j)\rangle\right)}{\int_0^\infty\exp\left(\langle\log p(\beta_j|\tau_j,\lambda,\gamma)\rangle+\langle\log f(\tau_j)\rangle\right)d\tau_j}. \quad (12)
$$

Notice from (3) that, the left hand side evaluated at $\beta_j^2=\langle\beta_j^2\rangle$ is the normalizing constant for $q(\tau_j|\lambda,\gamma)$. Furthermore, if we differentiate both sides of (3) with respect to β_j^2 and evaluate it again at $\beta_j^2=\langle\beta_j^2\rangle$, we obtain

$$
\begin{aligned}
&\frac{\lambda^{1-1/\gamma}\gamma\langle\beta_j^2\rangle^{\gamma/2}}{2\Gamma(1+1/\gamma)}\exp\left(-\lambda\langle\beta_j^2\rangle^{\gamma/2}\right)\\
&=\int_0^\infty\tau_j\underbrace{\frac{\lambda^{1/\gamma}\tau_j}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\tau_j\lambda^{2/\gamma}\langle\beta_j^2\rangle\right)f(\tau_j)}_{\propto q(\tau_j)}d\tau_j. \quad (13)
\end{aligned}
$$

This is the expectation of τ_j with respect to some unnormalized density which is proportional to $q(\tau_j)$. After normalization we obtain,

$$
\langle\tau_j\rangle=\lambda^{1-2/\gamma}\gamma\langle\beta_j^2\rangle^{\gamma/2-1}. \quad (14)
$$

Now we have all the required moments to carry out the iterative procedure explained in Section 2.1.

The lower bound $\mathcal{L}_{\lambda,\gamma}$ can be calculated very straightforwardly both for tracking the monotonic increase and for possibly setting a convergence criterion.

$$
\begin{aligned}
\mathcal{L}_{\lambda,\gamma} = {} & \langle\log p\left(\mathbf{y}|\boldsymbol{\beta},\sigma^2\right)\rangle+\langle\log p\left(\boldsymbol{\beta}|\boldsymbol{\tau},\lambda,\gamma\right)\rangle\\
&+\langle\log p\left(\sigma^{-2}\right)\rangle+\langle\log f\left(\boldsymbol{\tau}\right)\rangle\\
&-\langle\log q\left(\boldsymbol{\beta}|\lambda,\gamma\right)\rangle-\langle\log q\left(\sigma^{-2}|\lambda,\gamma\right)\rangle\\
&-\langle\log q\left(\boldsymbol{\tau}|\lambda,\gamma\right)\rangle \quad (15)
\end{aligned}
$$

where

$$
\begin{aligned}
\langle\log p\left(\mathbf{y}|\boldsymbol{\beta},\sigma^2\right)\rangle = {} & -\frac{1}{2}\sum_{i=1}^{n}n_i[\log 2\pi-\langle\log\sigma^{-2}\rangle]\\
&-\langle\sigma^{-2}\rangle(\widehat{d}-d_0) \quad (16)
\end{aligned}
$$

$$
\begin{aligned}
\langle\log p\left(\boldsymbol{\beta}|\boldsymbol{\tau},\lambda,\gamma\right)\rangle = {} & -\frac{p}{2}\log 2\pi+\frac{1}{2}\sum_{j=1}^{p}\langle\log\tau_j\rangle\\
&+\gamma^{-1}\log\lambda\\
&-\lambda^{2/\gamma}\sum_{j=1}^{p}\langle\tau_j\rangle\langle\beta_j^2\rangle \quad (17)
\end{aligned}
$$

$$
\begin{aligned}
\langle\log p\left(\sigma^{-2}\right)\rangle = {} & c_0\log d_0+(c_0-1)\langle\log\sigma^{-2}\rangle\\
&-d_0\langle\sigma^{-2}\rangle-\log\Gamma(c_0) \quad (18)
\end{aligned}
$$

$$
\langle\log f\left(\boldsymbol{\tau}\right)\rangle=\sum_{j=1}^{p}\langle\log f(\tau_j)\rangle \quad (19)
$$

$$
\langle\log q\left(\boldsymbol{\beta}|\lambda,\gamma\right)\rangle=-\frac{p}{2}\left(\log 2\pi+1\right)-\frac{1}{2}\log|\Sigma_{\boldsymbol{\beta}}| \quad (20)
$$

$$
\begin{aligned}
\langle\log q\left(\sigma^{-2}|\lambda,\gamma\right)\rangle = {} & \widehat{c}\log\widehat{d}+(\widehat{c}-1)\langle\log\sigma^{-2}\rangle\\
&-\widehat{d}\langle\sigma^{-2}\rangle-\log\Gamma\left(\widehat{c}\right) \quad (21)
\end{aligned}
$$

$$
\begin{aligned}
\langle\log q\left(\boldsymbol{\tau}|\lambda,\gamma\right)\rangle = {} & \sum_{j=1}^{p}\langle\log p(\beta_j|\tau_j,\lambda,\gamma)\rangle+p\log 2\\
&+\sum_{j=1}^{p}\langle\log f(\tau_j)\rangle-p\gamma^{-1}\log\lambda\\
&+p\log\Gamma(1+1/\gamma)+\lambda\sum_{j=1}^{p}\langle\beta_j^2\rangle^{\gamma/2}. \quad (22)
\end{aligned}
$$

After simplifications the lower-bound reduces to

$$
\begin{aligned}
\mathcal{L}_{\lambda,\gamma} = {} & -\frac{n}{2}\log(2\pi)+\frac{p}{2}+\frac{1}{2}\log|\Sigma_{\boldsymbol{\beta}}|+c_0\log d_0\\
&-\widehat{c}\log\widehat{d}-\log\Gamma(c_0)+\log\Gamma(\widehat{c})+p\gamma^{-1}\log\lambda\\
&-p\log 2-p\log\Gamma(1+1/\gamma)-\lambda\sum_{j=1}^{p}\langle\beta_j^2\rangle^{\gamma/2}. \quad (23)
\end{aligned}
$$

2.3 SOME REMARKS

Since our primary goal is sparsity, the lower end of the interval $\gamma = (0, 1)$ will be our preference. Recall that when $\gamma = 0$, (1) results in subset selection which performs the best in sparse underlying setups (Tibshirani 1996). Thus we will expect that our procedure yields sparser solutions as $\gamma \to 0$.

There is a striking similarity between solutions obtained by the variational bridge regression (VBR) as $\gamma \to 0$ and the variational relevance vector machines (VRVM) which arises from Student's t prior (through scale normal mixtures where the mixing distribution is gamma) on the regression coefficients (Bishop and Tipping 2000). Student's t distribution is a natural prior in robust Bayesian analysis due to its heavy tails. It is also used frequently as a natural shrinkage prior assigning larger densities in the neighborhood of 0.

The first moment of the regression coefficients for the VBR procedure can be obtained by the iterative procedure

$$\begin{aligned}\langle\boldsymbol{\beta}\rangle_{(k+1)} &= arg\min_{\boldsymbol{\beta}} (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &\quad +\lambda\gamma\langle\sigma^{-2}\rangle_{(k)}^{-1}\sum_{j=1}^{p}\beta_j^2\langle\beta_j^2\rangle_{(k)}^{\gamma/2-1}, \quad (24)\end{aligned}$$

and for the VRVM, obtained by

$$\begin{aligned}\langle\boldsymbol{\beta}\rangle_{(k+1)} &= arg\min_{\boldsymbol{\beta}} (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &\quad +(2\eta+1)\langle\sigma^{-2}\rangle_{(k)}^{-1}\sum_{j=1}^{p}\frac{\beta_j^2}{\langle\beta_j^2\rangle+2\mu}, \end{aligned} \quad (25)$$

where $\langle.\rangle_{(k)}$ denotes the expectation computed at the kth iteration, and η and μ are some hyperparameters arising from normal-gamma type priors used in VRVM; see (Bishop and Tipping 2000) for details. Notice from (24) and (25), as $\gamma \to 0$, $\mu \to 0$ and $\lambda\gamma \to 2\eta + 1$, these solutions will be identical. Of course, in these procedures, one would not choose to set $\mu = 0$, $\gamma = 0$ as these would lead to the impropriety of the posteriors remembering that in such cases the decomposition given in (4) would lose its meaning, i.e. the normalizing constant diverges. It would be sensible to choose small values as done by (Bishop and Tipping 2000).

Also note that if we replace $\langle\beta_j^2\rangle_{(k)}$ and $\langle\beta_j\rangle_{(k+1)}$ by $\beta_{j(k)}^2$ and $\beta_{j(k+1)}$ respectively where $\beta_{j(k+1)}$ is the minimizer of the above-mentioned optimization problems at the $(k+1)$th iteration, after convergence, the solution obtained is the MAP estimate of $\boldsymbol{\beta}$ which emerges through an expectation-maximization (EM) procedure maximizing $\log p(\boldsymbol{\beta}|\mathbf{y})$.

2.3.1 Computational Considerations

For smaller values of γ, and larger values of λ, as the algorithm proceeds to the solution, some $\langle\beta_j^2\rangle$ will approach zero corresponding to irrelevant predictors. Recall that $\langle\beta_j^2\rangle = \Sigma_{\boldsymbol{\beta},jj} + \langle\beta_j\rangle^2$ where $\Sigma_{\boldsymbol{\beta},jj}$ is the jth diagonal of the variance-covariance matrix of $\boldsymbol{\beta}$. Thus, as $\gamma \downarrow 0$ and $\lambda\gamma \to c$, where c is some sufficiently large positive finite constant, the approximate posteriors will become nearly singular along the dimensions of irrelevant predictors. To avoid numerical break-downs, we suggest that those β_js are removed from the model as the algorithm proceeds. More formally, we set $\beta_j = 0$, if $\langle\beta_j^2\rangle < \epsilon_{Mach}$ where ϵ_{Mach} denotes the machine epsilon. This will also reduce the computational burden since in a reduced model we will have to invert a smaller dimensional matrix. Thus the computational cost of the algorithm at each iteration is $O(p_k^3)$ where p_k is the the dimension of the problem at the kth iteration. When the lower-bound is tracked as the algorithm proceeds, some breaks can be observed where the dimension of the problem is being reduced. Note that as we remove predictors from the model, we are in fact creating a new problem with a different size which means we are computing a brand-new lower-bound. Thus this procedure can be seen as a series of lower-bound maximization problems where the problem with reduced number of dimensions uses the non-zero elements of $\langle\boldsymbol{\beta}\rangle$ in the larger model as its initial values.

2.4 EXPERIMENTS

In this section we will investigate the frequentist properties of the point estimator given by the proposed method ($\langle\boldsymbol{\beta}\rangle$) and contrast it with some of the state-of-the-art methods. Here we will consider $\gamma = 0.001$ and $\lambda\gamma = \{2, 3, 4, 5, 6, 7\}$ and will also set $c_0 = 0.1$ and $d_0 = 0.001$ for the experiments. These choices of hyper-parameters will lead to rather vague priors on $\boldsymbol{\beta}$ and σ^{-2}. Notice that the power exponential prior resulting from chosen γ and λ values will be very heavy-tailed and strongly peaked at zero.

We consider a model of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$. For the lasso and adaptive lasso solution paths we use the `lars` package in `R`. In the following experiments, $\mathbf{y}$ is centered and the columns of $\mathbf{X}$ are scaled to have unit 2-norm, i.e. $\|\mathbf{x}_j\|_2 = 1$ for $j = 1, ..., p$.

2.4.1 Simulation 1

This section reports the results of a simulation study comparing the model selection performance of the proposed method with that of lasso and adaptive lasso solution paths.

We set the number of predictors to $p = 64$, number of samples to $n = \{2.5p, 7.5p\}$, error standard deviation to $\sigma = 1$, and sparsity level to $q = \{0.25p, 0.5p, 0.75p\}$ where q is the number of nonzero β_js in the model and the regression coefficients to $\boldsymbol{\beta}_{1:q} = (1, ..., 1)'$, $\boldsymbol{\beta}_{q+1:p} = \mathbf{0}$. Then the model is generated as follows:

1. $\mathbf{C} \sim \mathcal{W}(p, \mathbf{I}_p)$, where $\mathcal{W}$ denotes a Wishart density
2. $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, where $\mathbf{x}_i$ is the ith row of $\mathbf{X}$
3. $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$

For each (n, q) combination we generate 100 covariance matrices, $\mathbf{C}$, and for each $\mathbf{C}$ we generate 100 design matrices, $\mathbf{X}$. In Simulation 2 and 3, as in some earlier studies (Tibshirani 1996, Zou 2006, Figueiredo 2003), we fixed the correlation structure at $cor(x_{ij}, x_{ik}) = 0.5^{|j-k|}$ where j and k denote the jth and the kth covariates. The results of (Zhao and Yu 2006) show that under such a design (power decay correlation) the lasso is model selection consistent. By randomizing the correlation of the design we hope to provide a more realistic assessment of effectiveness of such procedures in terms of correct model specification. The results are presented in Figure 1.

From top to bottom in Figure 1 we have six pairs of sample size and sparsity levels, (n, q)=$\{(2.5p, 0.25p)$, $(7.5p, 0.25p)$, $(2.5p, 0.5p)$, $(7.5p, 0.5p)$, $(2.5p, 0.75p)$, $(7.5p, 0.75p)\}$. Not surprisingly as we move in the direction (↓), the performance of the proposed method diminishes. Nonetheless, in all cases we can see that the proposed method outperforms the entire solution paths of the competing procedures. Recall that here we are comparing individual estimates arising from the proposed method with the entire solution paths of the lasso and the adaptive lasso (thus one would still need to choose the tuning parameters). This performance is most pronounced, again not surprisingly, in sparse underlying cases (first two subfigures in Figure 1). Due to the structure of the prior in VBR (for small γ), we are observing a sparse estimation behavior similar to that of subset selection. Note that for $(n, q) = (2.5p, 0.75p)$ all methods fail to detect the correct underlying model. Of course this is not to say necessarily that they would also provide bad predictions.

When it comes to the choice of a particular $\lambda\gamma$ value, $\lambda\gamma = \{2, 3\}$ values may be chosen as a good compromise among all cases considered. However, recall that the method is tested only with a value of $\gamma = 0.001$ as we wanted to mimic a subset selection behavior and obtain sparse estimation which requires a prior assumption of a sparse underlying model. Further discussion is given in Conclusions.

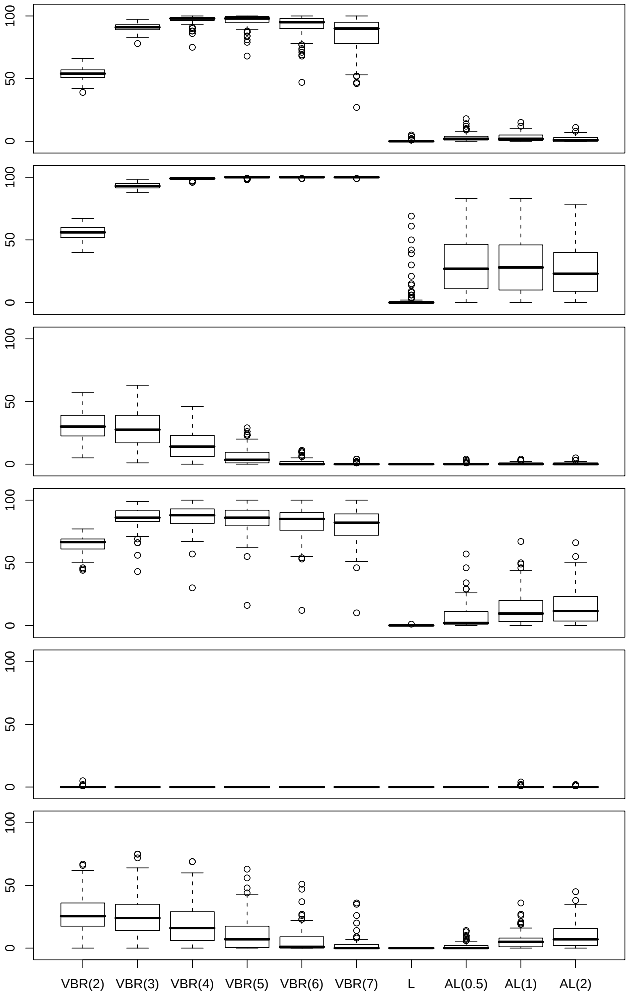

Figure 1: Model selection performances of VBR vs. the lasso and adaptive lasso solution paths. In each subfigure the boxplots are for VBR ($\lambda\gamma = \{2, 3, 4, 5, 6, 7\}$), lasso and adaptive lasso ($\gamma = \{0.5, 1, 2\}$) respectively. The vertical axis represents the number of correctly specified models (out of 100).

2.4.2 Simulation 2

We now compare the prediction accuracy and model selection consistency using following two models which are drawn from (Tibshirani 1996). The proposed method is contrasted with the lasso (Tibshirani 1996), the adaptive lasso (Zou 2006), the non-negative garrote (Breiman 1995) and the ordinary least squares (OLS) estimate.

Model 1: In this example, we let $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ with iid normal predictors $\mathbf{x}_i$ ($i = 1, ..., n$) where the pairwise correlation between the jth and the kth columns of $\mathbf{X}$ is adjusted to be $(.5)^{|j-k|}$.

Model 2: We use the same setup as model 1 with $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)'$.

To choose the tuning parameters for the lasso and the adaptive lasso we use 10-fold cross-validation. We also use the the method of (Yuan and Lin 2005), CML, to select the tuning parameter for the lasso. For both models we set $\sigma = 3$ and experiment with two levels of sample size, $n = \{50, 100\}$. 100 data sets are generated for each case. In Table 1, we report the median prediction error (MSE) on a test set of 10,000 observation for each of the 100 cases. The values in the parentheses give the bootstrap standard error of the median MSE values obtained. C, I and CM respectively stand for the number of correct predictors chosen, number of incorrect predictors chosen and the proportion of cases (out of 100) where the correct model was specified by the method. The bootstrap standard error was calculated by generating 500 bootstrap samples from 100 MSE values, finding the median MSE for each case, and then calculating the standard error of the median MSE.

It is clear from the results that the proposed method outperforms the others by a large margin under Model 2 both in terms of prediction and model selection accuracy. Under Model 1, it still performs the best for certain choices of $\lambda\gamma$. It is evident and expected that the smaller values of $\lambda\gamma$ perform better in the denser model (Model 1). Having evaluated these results, $\lambda\gamma = 3$ value can be chosen as a good compromise.

2.4.3 Simulation 3

Here we consider the case where $p > n$. We let $\boldsymbol{\beta}_{1:q} = (5, ..., 5)'$, $\boldsymbol{\beta}_{q+1:p} = \mathbf{0}$, $p = 250$, $q = 10$, $\sigma = \{1, 3\}$ and $n = 50$ with iid normal predictors $\mathbf{x}_i$ $(i = 1, ..., n)$ where the pairwise correlation between the jth and the kth columns of $\mathbf{X}$ is again adjusted to be $(.5)^{|j-k|}$. We compare the proposed method with the lasso and the adaptive lasso.

Again it is evident from the results presented in Table 2 that for certain choices of $\lambda\gamma$, the proposed method significantly outperforms the others. For $p > n$ case, VBR yields better results for larger $\lambda\gamma$ values than it did for $n > p$ case. $\lambda\gamma = 5$ can be suggested as a good compromise in this simulation study.

Table 1: Prediction and model selection performances of VBR and the competing methods in Simulation 2.

Method	MSE(se)	C	I	CM
Model 1, $n = 50$				
VBR ($\lambda\gamma = 2$)	10.2255(0.1477)	2.69	0.17	60
VBR ($\lambda\gamma = 3$)	11.0053(0.1981)	2.39	0.06	43
VBR ($\lambda\gamma = 4$)	11.4702(0.1972)	2.08	0.02	27
VBR ($\lambda\gamma = 5$)	11.9725(0.4651)	1.86	0.01	19
VBR ($\lambda\gamma = 6$)	15.2724(0.3710)	1.55	0	9
VBR ($\lambda\gamma = 7$)	15.6777(0.1171)	1.31	0	8
Lasso (CV)	10.2814(0.0999)	3	2.11	11
Lasso (CML)	10.2530(0.1266)	2.95	1.09	26
A. Lasso (CV)	10.5387(0.1576)	2.79	1.19	27
Non. Garrote	10.2000(0.0996)	2.98	2.41	2
OLS	15.7119(0.7054)	3	5	0
Model 1, $n = 100$				
VBR ($\lambda\gamma = 2$)	9.4047(0.0495)	2.98	0.08	90
VBR ($\lambda\gamma = 3$)	9.4322(0.0492)	2.87	0.02	87
VBR ($\lambda\gamma = 4$)	9.6489(0.1230)	2.70	0.01	71
VBR ($\lambda\gamma = 5$)	10.4674(0.3727)	2.49	0	52
VBR ($\lambda\gamma = 6$)	10.9238(0.0899)	2.30	0	39
VBR ($\lambda\gamma = 7$)	11.0405(0.0726)	2.16	0	29
Lasso (CV)	9.6846(0.0677)	3	2.04	15
Lasso (CML)	9.5897(0.0529)	3	0.99	36
A. Lasso (CV)	9.7485(0.0874)	2.87	0.92	40
Non. Garrote	9.5880(0.0587)	3	2.39	7
OLS	9.6985(0.0868)	3	5	0
Model 2, $n = 50$				
VBR ($\lambda\gamma = 2$)	9.3366(0.0434)	1	0.14	86
VBR ($\lambda\gamma = 3$)	9.2668(0.0342)	1	0.01	99
VBR ($\lambda\gamma = 4$)	9.2610(0.0350)	1	0.01	99
VBR ($\lambda\gamma = 5$)	9.2545(0.0367)	1	0	100
VBR ($\lambda\gamma = 6$)	9.2592(0.0335)	0.99	0	99
VBR ($\lambda\gamma = 7$)	9.2650(0.0466)	0.99	0	99
Lasso (CV)	9.7558(0.0782)	1	1.53	32
Lasso (CML)	9.5396(0.0662)	1	1	44
A. Lasso	9.6441(0.1430)	1	0.97	41
Non. Garrote	9.9496(0.0735)	1	3.1	1
OLS	15.1516(0.5148)	1	7	0
Model 2, $n = 100$				
VBR ($\lambda\gamma = 2$)	9.1768(0.0171)	1	0.11	90
VBR ($\lambda\gamma = 3$)	9.1579(0.0205)	1	0	100
VBR ($\lambda\gamma = 4$)	9.1625(0.0180)	1	0	100
VBR ($\lambda\gamma = 5$)	9.1613(0.0181)	1	0	100
VBR ($\lambda\gamma = 6$)	9.1620(0.0188)	1	0	100
VBR ($\lambda\gamma = 7$)	9.1656(0.0154)	1	0	100
Lasso (CV)	9.4131(0.0354)	1	2.06	20
Lasso (CML)	9.2780(0.0300)	1	1.20	42
A. Lasso	9.4289(0.0757)	1	1.12	34
Non. Garrote	9.4253(0.0418)	1	3.31	1
OLS	9.6921(0.0632)	1	7	0

3 EXTENSIONS

The discussed method can easily be extended to the binary response case. Suppose we make n binary observations, $y_1, ..., y_n$. Following (Albert and Chib 1993), we can create a probit model by introducing n latent variables, $z_1, ..., z_n$, where z_i are independent $\mathcal{N}(\mathbf{x}_i\boldsymbol{\beta}, 1)$, and defining $y_i = 1$ if $z_i > 0$ and $y_i = 0$ otherwise. The joint posterior density for our problem can then be written as

$$p(\boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\tau}, \mathbf{y}) = p(\boldsymbol{\beta}|\boldsymbol{\tau})p(\boldsymbol{\tau}) \prod_{i=1}^{n} \mathcal{N}\left(z_i; \mathbf{x}_i\boldsymbol{\beta}, 1\right), \qquad (26)$$

such that, $z_i > 0$ if $i \in \mathcal{I}$ and $z_i \leq 0$ otherwise, where $\mathcal{I} = \{i|y_i = 1\}$.

Similarly to the derivations in the linear regression case, it is easy to show that the variational approximations to the distributions of $\boldsymbol{\beta}$ and z_i for $i = 1, ..., n$

Table 2: Prediction and model selection performances of VBR, lasso and adaptive lasso in Simulation 3.

Method	MSE(se)	C	I	CM
	$\sigma = 1$			
VBR ($\lambda\gamma = 2$)	1.6257(0.0730)	10	26.81	0
VBR ($\lambda\gamma = 3$)	0.8126(0.0753)	10	11.01	38
VBR ($\lambda\gamma = 4$)	0.2677(0.0163)	10	1.09	94
VBR ($\lambda\gamma = 5$)	0.2575(0.0173)	10	0	100
VBR ($\lambda\gamma = 6$)	0.2608(0.0172)	10	0	100
VBR ($\lambda\gamma = 7$)	0.2638(0.0167)	10	0	100
Lasso (CV)	0.8233(0.0809)	10	20.35	0
Lasso (CML)	1.0549(0.0804)	10	20.37	0
A. Lasso (CV)	1.7222(0.0958)	10	38.97	0
	$\sigma = 3$			
VBR ($\lambda\gamma = 2$)	18.8568(0.6706)	10	32.91	0
VBR ($\lambda\gamma = 3$)	16.5078(1.2552)	10	21.67	15
VBR ($\lambda\gamma = 4$)	3.6775(0.5267)	9.99	7.18	70
VBR ($\lambda\gamma = 5$)	3.3592(0.2555)	9.96	2.39	92
VBR ($\lambda\gamma = 6$)	3.5160(0.3382)	9.95	1.04	95
VBR ($\lambda\gamma = 7$)	3.9673(0.4997)	9.64	0.36	91
Lasso (CV)	10.8621(0.7361)	10	21.97	0
Lasso (CML)	15.2470(0.4904)	9.99	39.01	0
A. Lasso (CV)	7.8980(0.1114)	10	18.15	0

are

$$q(\boldsymbol{\beta}|\lambda,\gamma) \quad \overset{d}{=} \quad \mathcal{N}\left(\widehat{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}\right) \tag{27}$$

$$q(z_i|\lambda,\gamma) \quad \overset{d}{=} \quad \begin{cases} \mathcal{N}^{+}\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}, 1\right), & y_i = 1 \\ \mathcal{N}^{-}\left(\mathbf{x}_i\widehat{\boldsymbol{\beta}}, 1\right), & y_i = 0 \end{cases} \tag{28}$$

where

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \Sigma_{\boldsymbol{\beta}}\mathbf{X}'\langle\mathbf{z}\rangle \\ \Sigma_{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X} + \mathbf{T}\right)^{-1} \\ \mathbf{T} &= \lambda^{2/\gamma} diag(\langle\tau_j\rangle) \\ \langle z_i\rangle &= \begin{cases} \mathbf{x}_i\langle\boldsymbol{\beta}\rangle + \frac{\phi(\mathbf{x}_i\langle\boldsymbol{\beta}\rangle)}{1-\Phi(-\mathbf{x}_i\langle\boldsymbol{\beta}\rangle)}, & y_i = 1 \\ \mathbf{x}_i\langle\boldsymbol{\beta}\rangle - \frac{\phi(\mathbf{x}_i\langle\boldsymbol{\beta}\rangle)}{\Phi(-\mathbf{x}_i\langle\boldsymbol{\beta}\rangle)}, & y_i = 0 \end{cases} \\ \langle\tau_j\rangle &= \lambda^{1-2/\gamma}\gamma\langle\beta_j^2\rangle^{\gamma/2-1} \\ \langle\boldsymbol{\beta}\rangle &= \widehat{\boldsymbol{\beta}} \\ \langle\boldsymbol{\beta}\boldsymbol{\beta}'\rangle &= \boldsymbol{\Sigma}_{\boldsymbol{\beta}} + \langle\boldsymbol{\beta}\rangle\langle\boldsymbol{\beta}'\rangle, \end{aligned}$$

$\mathcal{N}^{+}$ and $\mathcal{N}^{-}$ denote normal densities truncated from left and right respectively and $\phi(.)$ and $\Phi(.)$ denote the standard normal density and distribution functions respectively.

The lower bound in this case can be calculated again very straight-forwardly as

$$\begin{aligned} \mathcal{L}_{\lambda,\gamma} \quad = \quad & \langle\log p\left(\mathbf{y},\mathbf{z}|\boldsymbol{\beta}\right)\rangle + \langle\log p\left(\boldsymbol{\beta}|\boldsymbol{\tau},\lambda,\gamma\right)\rangle \\ & + \langle\log p\left(\boldsymbol{\tau}\right)\rangle - \langle\log q\left(\boldsymbol{\beta}|\lambda,\gamma\right)\rangle \\ & - \langle\log q\left(\mathbf{z}|\lambda,\gamma\right) - \langle\log q\left(\boldsymbol{\tau}|\lambda,\gamma\right)\rangle \end{aligned} \tag{29}$$

where

$$\begin{aligned} \langle\log p\left(\mathbf{y},\mathbf{z}|\boldsymbol{\beta}\right)\rangle \quad = \quad & -\frac{n}{2}\log 2\pi - \frac{1}{2}\langle\mathbf{z}'\mathbf{z}\rangle - \langle\mathbf{z}'\rangle\mathbf{X}\langle\boldsymbol{\beta}\rangle \\ & + \frac{1}{2}\sum_{i=1}^{n}\mathbf{x}_i\langle\boldsymbol{\beta}\boldsymbol{\beta}'\rangle\mathbf{x}_i' \end{aligned} \tag{30}$$

$$\begin{aligned} \langle\log p\left(\boldsymbol{\beta}|\boldsymbol{\tau},\lambda,\gamma\right)\rangle \quad = \quad & -\frac{p}{2}\log 2\pi + \frac{1}{2}\sum_{j=1}^{p}\langle\log\tau_j\rangle \\ & + \gamma^{-1}\log\lambda \\ & - \lambda^{2/\gamma}\sum_{j=1}^{p}\langle\tau_j\rangle\langle\beta_j^2\rangle \end{aligned} \tag{31}$$

$$\langle\log p\left(\boldsymbol{\tau}\right)\rangle = \sum_{j=1}^{p}\langle\log\pi(\tau_j)\rangle \tag{32}$$

$$\langle\log q\left(\boldsymbol{\beta}|\lambda,\gamma\right)\rangle = -\frac{p}{2}\left(\log 2\pi + 1\right) - \frac{1}{2}\log|\Sigma_{\boldsymbol{\beta}}| \tag{33}$$

$$\begin{aligned} \langle\log q\left(\mathbf{z}|\lambda,\gamma\right)\rangle \quad = \quad & -\frac{n}{2}\log 2\pi - \frac{1}{2}\langle\mathbf{z}'\mathbf{z}\rangle - \langle\mathbf{z}'\rangle\mathbf{X}\langle\boldsymbol{\beta}\rangle \\ & + \frac{1}{2}\langle\boldsymbol{\beta}\rangle'\mathbf{X}'\mathbf{X}\langle\boldsymbol{\beta}\rangle \\ & - \sum_{i\in\mathcal{I}}\log\left(1 - \Phi(-\mathbf{x}_i\langle\boldsymbol{\beta}\rangle)\right) \\ & - \sum_{i\in\mathcal{I}'}\log\left(\Phi(-\mathbf{x}_i\langle\boldsymbol{\beta}\rangle)\right) \end{aligned} \tag{34}$$

$$\begin{aligned} \langle\log q\left(\boldsymbol{\tau}|\lambda,\gamma\right)\rangle \quad = \quad & \sum_{j=1}^{p}\langle\log p(\beta_j|\tau_j,\lambda,\gamma)\rangle + p\log 2 \\ & + \sum_{j=1}^{p}\langle\log f(\tau_j)\rangle - p\gamma^{-1}\log\lambda \\ & + p\log\Gamma(1+1/\gamma) + \lambda\sum_{j=1}^{p}\langle\beta_j^2\rangle^{\gamma/2}. \end{aligned} \tag{35}$$

After simplifications we obtain

$$\begin{aligned} \mathcal{L}_{\lambda,\gamma} \quad = \quad & -\frac{n}{2}\log(2\pi) + p + \log|\Sigma_{\boldsymbol{\beta}}| + \gamma^{-1}\log\lambda \\ & - \log 2 - \log\Gamma(1+1/\gamma) - \lambda\sum_{j=1}^{p}\langle\beta_j^2\rangle^{\gamma/2} \\ & + \sum_{i\in\mathcal{I}}\log\left(1 - \Phi(-\mathbf{x}_i\langle\boldsymbol{\beta}\rangle)\right) \\ & + \sum_{i\in\mathcal{I}'}\log\left(\Phi(-\mathbf{x}_i\langle\boldsymbol{\beta}\rangle)\right) \\ & + \frac{1}{2}\left(\sum_{i=1}^{n}\mathbf{x}_i\langle\boldsymbol{\beta}\boldsymbol{\beta}'\rangle\mathbf{x}_i' - \langle\boldsymbol{\beta}\rangle'\mathbf{X}'\mathbf{X}\langle\boldsymbol{\beta}\rangle\right). \end{aligned} \tag{36}$$

4 CONCLUSIONS

We have considered a variational approximation approach to Bayesian inference in regression models where the prior on the regression coefficients is of the exponential power form. Nonconjugacy of such priors with a normal likelihood results in posteriors that we cannot efficiently sample from. Exploiting the scale mixture normal representation of such distributions we form a hierarchical Bayesian model which mimics the behavior of such priors. Although the mixing distribution is not explicitly obtained, under the mean field variational approximations we merely need to acquire the required moments which is very straight-forward. We showed that this yields a computationally rather attractive procedure providing us with a sparse estimator as well as approximate inferences on the parameters of interest. The choice of mentioned hyperparameters is of course a concerning issue. However, as our primary goal in this study is sparse estimation, we worked with small values of γ which very well mimic the sparse estimation characteristic of best subset selection. Then we experimented with a set of $\lambda\gamma$ values to empirically evaluate their performances to come up with a reasonable compromise for fast and sparse estimation/prediction.

As mentioned earlier, we used a small value of γ to obtain sparse estimation which required a prior assumption of a sparse underlying model. This is not a radical assumption in many of today's problems with large number of available predictors and potentially merely a few of them explaining the response. If we do not have such a prior assumption of a sparse underlying model, it may be better to conduct a search over γ and λ values as the exploited hierarchy covers a wide spectrum of possible priors from the ones that strongly enforce sparsity to the ones leading to ridge-regression-like solutions.

We also provided a straight-forward extension to the analysis of binary responses yet limited the experiments to the linear regression case due to space constraints.

Acknowledgements

The author would like to thank Sayan Mukherjee and the reviewers for their helpful comments.

References

Albert, J. H. and Chib, S. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88 (1993).

Andrews, D. F. and Mallows, C. L. "Scale Mixtures of Normal Distributions." *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):99–102 (1974).

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer (2006).

Bishop, C. M. and Tipping, M. E. "Variational Relevance Vector Machines." In *UAI '00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, 46–53. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. (2000).

Breiman, L. "Better Subset Regression Using the Nonnegative Garrote." *Technometrics*, 37(4):373–384 (1995).

Figueiredo, M. A. T. "Adaptive sparseness for supervised learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1150–1159 (2003).

Frank, I. E. and Friedman, J. H. "A Statistical View of Some Chemometrics Regression Tools." *Technometrics*, 35(2):109–135 (1993).

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. *An introduction to variational methods for graphical models*. Cambridge, MA, USA: MIT Press (1999).

Park, T. and Casella, G. "The Bayesian Lasso." *Journal of the American Statistical Association*, 103:681–686(6) (2008).

Tibshirani, R. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288 (1996).

West, M. "Outlier Models and Prior Distributions in Bayesian Linear Regression." *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):431–439 (1984).

—. "On Scale Mixtures of Normal Distributions." *Biometrika*, 74(3):646–648 (1987).

Yuan, M. and Lin, Y. "Efficient Empirical Bayes Variable Selection and Estimation in Linear Models." *Journal of the American Statistical Association*, 100(472):1215–1225 (2005).

Zhao, P. and Yu, B. "On Model Selection Consistency of Lasso." *J. Mach. Learn. Res.*, 7 (2006).

Zou, H. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association*, 101:1418–1429 (2006).