The Canadian Journal of Statistics
Vol. xx, No. yy, 20??, Pages 1–??
La revue canadienne de statistique

Bayesian inference for high-dimensional linear regression under mnet priors

Aixin Tan[1]* and Jian Huang[1]

[1]*Department of Statistics and Actuarial Science, University of Iowa*

Key words and phrases: Bayesian computing ; hyperparameters ; Markov chain Monte Carlo ; median probability model ; Penalized regression ; posterior distribution ; variable selection.

MSC 2010: Primary 62F15; secondary 60J22, 62J07

Abstract: For regression problems that involve many potential predictors, the Bayesian variable selection (BVS) method is a powerful tool, which associates each model with its posterior probabilities, and achieves superb prediction performance through Bayesian model averaging (BMA). Two challenges of using such models are, specifying a suitable prior, and computing posterior quantities for inference. We contribute to the literature of BVS modeling in the following aspects. First, we propose a new family of priors, called the mnet prior, which is indexed by a few hyperparameters that allow great flexibility in the prior density. The hyperparameters can also be treated as random, so that their values need not be tuned manually, but will instead adapt to the data. Simulation studies are used to demonstrate good prediction and variable selection performances of these models. Secondly, the analytical expression of the posterior distribution is unavailable for the BVS model under the mnet prior in general, as is the case for most BVS models. We develop an adaptive Markov chain Monte Carlo (MCMC) algorithm that facilitates the computation in high dimensional regression problems. Finally, we showcase various ways to do inference with BVS models, highlighting a new way to visualize the importance of each predictor, along with estimation of the coefficients and their uncertainties. These are demonstrated through the analysis of a breast cancer dataset.
The Canadian Journal of Statistics xx: 1–??; 20??

Résumé: Insérer votre résumé ??? *La revue canadienne de statistique* xx: 1–??; 20??

1. INTRODUCTION

Penalized regression (PR) methods such as the ridge, the lasso, the elastic net (enet), and the minimax concave penalty (mcp) methods have seen wide applications as alternatives to the least squares method in fitting regression models. Assume that the data arise from a linear model $Y = X\boldsymbol{\beta} + \epsilon$, where Y is an $n \times 1$ response, $X = (X_1, \cdots, X_q)$ denotes q potential predictors, ϵ is an $n \times 1$ vector of independent and identically distributed normal errors with mean 0 and variance σ^2, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^\top$ is the vector of coefficients. Without loss of generality, suppose $Y^\top \mathbf{1} = 0$, $X_j^\top \mathbf{1} = 0$ and $X_j^\top X_j = n$ for $j = 1, \ldots, q$, so that an intercept is not needed. Two major goals are to identify the most relevant explanatory variables, and to predict future responses. A PR method estimates $\boldsymbol{\beta}$ using the minimizer of

$$L(\boldsymbol{\beta}; \boldsymbol{\lambda}) = \frac{1}{2n}\|Y - X\boldsymbol{\beta}\|^2 + p(\boldsymbol{\beta}; \boldsymbol{\lambda}), \tag{1}$$

where $p(\cdot; \boldsymbol{\lambda}) : \mathbb{R}^q \to \mathbb{R}^+$ is a penalty function with penalty parameter $\boldsymbol{\lambda}$, and $\|\cdot\|$ denotes the L_2 norm. Relative to the least squares method, PR methods generally estimate coefficients to be

* *Author to whom correspondence may be addressed.*
E-mail: aixin-tan@uiowa.edu

closer to zero, and sometimes exactly zero, which in effect drops the corresponding predictors from the model. Assessing the uncertainty of the selected model is important yet challenging, and is the subject of an active research area (see Lockhart et al. (2014)).

It is well known that PR solutions can be interpreted as the posterior mode of Bayesian models under appropriately specified priors (Tibshirani, 1996; Zou & Hastie, 2005), and many Bayesian models have been developed upon such connections. For instance, inspired by the lasso penalty function, Park & Casella (2008) assigned an independent double exponential prior on $\boldsymbol{\beta}$, $f(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\lambda}) \propto \Pi_j \exp\left\{-\frac{\lambda_1|\beta_j|}{\sigma}\right\}$. And inspired by the enet penalty function, Li & Lin (2010) proposed an independent enet prior, $f(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\lambda}) \propto \Pi_j \exp\left\{-\frac{\lambda_1|\beta_j|+\lambda_2\beta_j^2}{\sigma^2}\right\}$. These Bayesian models are completed by assigning a prior to $(\sigma^2, \boldsymbol{\lambda})$. Compared to their PR counterparts, the Bayesian models are advantageous in having the entire posterior distributions to draw inference from. Note that the above priors for $\boldsymbol{\beta}$ are continuous, hence the posterior probability is concentrated on the full model. Various follow-up procedures are available to exclude predictors, such as dropping coefficients with close-to-zero posterior means (Ishwaran & Rao, 2005).

In this paper, we adopt an alternative framework, namely, the Bayesian variable selection (BVS) framework, given in Equations (2a) to (2d). Briefly, a binary vector $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_q)$ is introduced to denote a subset of predictors, where γ_j equals 1 if predictor j is included. A prior is assigned over the space of the 2^q models, which leads to a posterior distribution of $\boldsymbol{\gamma}$ that reflects the plausibility of each model. Conditional on a given $\boldsymbol{\gamma}$, a continuous prior is assigned to the coefficients of the selected predictors, $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = \{\beta_j : \gamma_j = 1\}$, while the other coefficients are fixed at zero. Mitchell & Beauchamp (1988) was among the first to propose a BVS model, and many variations of it have been studied, see, for e.g., George & McCulloch (1993, 1997), Geweke (1996), Smith & Kohn (1996), Liang et al. (2008), O'Hara & Sillanpää (2009). In existing literature, the priors for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ are often related to different penalty functions. For example, Johnstone & Silverman (2005) and Yuan & Lin (2005) specified an independent double-exponential prior for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, $f\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \boldsymbol{\lambda}\right) \propto \Pi_{\{j:\gamma_j=1\}} \exp\left\{-\frac{\lambda_1|\beta_j|}{\sigma^2}\right\}$, that corresponds to the lasso penalty. And Hans (2011) considered a Bayesian enet model that uses the same enet prior as that of Li & Lin (2010), but for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ instead of $\boldsymbol{\beta}$.

Our paper adds to the literature by proposing a new, versatile family of priors for the coefficients. The new prior is related to the minimax concave (mc) penalty function (Zhang, 2010), and an extension of it called the mnet penalty function (Huang et al., 2016+), to be introduced in sec. 2. We refer to the new prior as the mnet prior, and the corresponding BVS model as the Bayesian mnet (bmnet) model. An mnet prior is indexed by a vector of hyperparameters $\boldsymbol{\lambda}$, analogous to the "penalty parameters" in the mnet PR method. Certain choices of $\boldsymbol{\lambda}$ reduce the mnet prior to its special cases including the normal, the double exponential and the enet priors. Further, $\boldsymbol{\lambda}$ can be treated as random in the fully Bayesian approach. That is, instead of debating between a normal and a double exponential prior in practice, it's convenient to use the mnet prior, and let the data dictate an appropriate combination of the two through dynamic calibration of $\boldsymbol{\lambda}$.

The posterior distribution of the proposed bmnet model does not have an analytical expression. We develop a new MCMC algorithm, specifically an adaptive Metropolis-Hastings within block Gibbs sampler, that enables computing for the bmnet model in high dimensional problems.

Finally, we demonstrate different ways to present inferential results based on BVS models, that PR methods can not produce. These results include, but are not limited to, inclusion probabilities for potential predictors, credible intervals for regression coefficients, and credible intervals for predicting the response of new observations. We highlight a novel graphical display that combines the importance and the impact of predictor variables, which can be found in Figure 5.

TABLE 1: Five penalty functions.

	penalty parameter(s)	penalty function
ridge	$\boldsymbol{\lambda} = \lambda_2 \in \mathbb{R}^+$	$p_{\text{R}}(t;\boldsymbol{\lambda}) = \frac{\lambda_2}{2}t^2$
lasso	$\boldsymbol{\lambda} = \lambda_1 \in \mathbb{R}^+$	$p_L(t;\boldsymbol{\lambda}) = \lambda_1\lvert t\rvert$
enet	$\boldsymbol{\lambda} = (\lambda_1,\lambda_2) \in \mathbb{R}^+ \times \mathbb{R}^+$	$p_{\text{E}}(t;\boldsymbol{\lambda}) = \lambda_1\lvert t\rvert + \frac{\lambda_2}{2}t^2$
mc	$\boldsymbol{\lambda} = (\lambda_1,\kappa) \in \mathbb{R}^+ \times [0,1)$	$p_{\text{mc}}(t;\boldsymbol{\lambda}) = \begin{cases} \lambda_1\lvert t\rvert - \frac{\kappa}{2}t^2, & \lvert t\rvert < \lambda_1/\kappa \\ \lambda_1^2/2\kappa, & \lvert t\rvert \geq \lambda_1/\kappa \end{cases}$
mnet	$\boldsymbol{\lambda} = (\lambda_1,\lambda_2,\kappa) \in \mathbb{R}^+ \times \mathbb{R}^+ \times [0,1)$	$p_{\text{mn}}(t;\boldsymbol{\lambda}) = p_{\text{mc}}(t;\lambda_1,\kappa) + \frac{\lambda_2}{2}t^2$

The rest of the paper is organized as the following. A few PR methods are reviewed in sec. 2. The BVS framework and the mnet prior are introduced in sec. 3. Computing for the proposed bmnet model is handled by an adaptive MCMC algorithm in sec. 4. In sec. 5, simulation studies show good prediction and variable selection performances of our model in both $q < n$ and $q \geq n$ situations with both moderately and highly correlated predictors. In sec. 6, analysis of a breast cancer dataset is used to showcase various ways of making inference with our model, and with BVS models in general. The paper concludes with a discussion in sec. 7. Certain technical details and graphs are left for a supplement. All data analysis were done in R (R Core Team, 2015).

2. A BRIEF REVIEW OF PENALIZED REGRESSION METHODS

The penalty function in (1) usually takes an additive form,

$$p(\boldsymbol{\beta};\boldsymbol{\lambda}) = \sum_{j=1}^{q} p(\beta_j;\boldsymbol{\lambda})\,.$$

Table 1 lists five popular penalty functions, among many others, from a rich literature on PR methods. The ridge penalty is known for stabilizing the coefficient estimates when predictors are highly correlated. But with probability 1, all coefficients have nonzero estimates. In comparison, the lasso penalty produces sparse estimates. But the estimates are biased, and do not predict as well under multicollinearity. The enet penalty adds an L_2 term to the lasso penalty, which encourages highly correlated predictors to be added or dropped together, and often results in better variable selection performance than the lasso (Zou & Hastie, 2005).

The fourth function in Table 1 is the minimax concave penalty (mcp) proposed by Zhang (2010). The mcp has two tuning parameters, λ_1 and κ. Here, κ is the *maximum concavity parameter*, such that the mcp approaches the lasso penalty when $\kappa \to 0$, and the hard-thresholding penalty when $\kappa \to 1$. As shown in Figure 1, at $t = 0$, the mcp function applies penalization at the same rate, p', as that of lasso, but continuously relaxes the penalization until $\lvert t\rvert > \frac{\lambda}{\kappa}$, when the rate drops to 0. In effect, using the mcp results in nearly unbiased estimates for large coefficients. Here, we mention the scad penalty function (Fan & Li, 2001), which produces estimators with three desirable properties: unbiasedness, sparsity and continuity. In fact, estimators from the mcp method enjoy all these properties, while the mcp has a simper derivative than that of scad.

The fifth penalty in Table 1 is the mnet penalty proposed by Huang et al. (2016+), which adds an additional L_2 term to the mcp. Similar to how the enet improves upon the lasso, the mnet has an advantage over the mcp in variable selection, especially when the correlation matrix of the predictors has a blocking structure. The enet and the mnet methods have been shown to have selection consistency under different conditions. The conditions on the mnet are usually less restrictive, especially in cases where q is large relative to n. Further, the enet produces

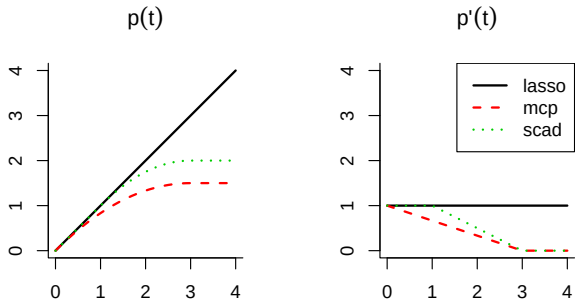


FIGURE 1: The lasso, the mcp and the scad penalty functions and their derivatives.

asymptotically biased estimators due to its L_1 component, a drawback that is overcame by the mnet. The first four penalty functions in Table 1 are indeed special cases of the mnet penalty. For example, the mnet penalty reduces to (a) the ridge penalty when $\lambda_1 = 0$, (b) the mcp when $\lambda_2 = 0$, and (c) the enet penalty when $\kappa = 0$. Further, the mcp reduces to the lasso penalty when $\kappa = 0$. The supplement contains a graphical display of the relationship among these penalties.

Specifying the penalty parameter $\boldsymbol{\lambda}$ in any penalty function is critical, and is typically done by a k-fold cross-validation, with $k = 5$ or 10. The chosen value of $\boldsymbol{\lambda}$ then leads to a final model that minimizes the penalized log-likelihood (1). Since the final model is usually sensitive to the observed dataset, as well as the random procedure used to select $\boldsymbol{\lambda}$, it is important to report the uncertainty of the selected model, as well as the estimates and predictions based on it. This is a challenging task, and the subject of much research in the PR community. In this paper, we take a Bayesian approach to regression problems, for which measuring model uncertainty is routine.

3. BAYESIAN VARIABLE SELECTION MODELS UNDER MNET PRIORS

3.1. The structure of the Bayesian mnet model

Let $\mathrm{N}_n(\mu, V)$ denote the n-dimensional multivariate normal distribution with mean μ and covariance matrix V. The general form of a BVS model is given by

$$Y|\,\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\lambda}, w \sim \mathrm{N}_n(X_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\,, \sigma^2 I_n)\,, \tag{2a}$$

$$\beta_j|\,\boldsymbol{\gamma}, \sigma^2, \boldsymbol{\lambda}, w \overset{\text{iid}}{\sim} \gamma_j f_{\boldsymbol{\lambda},\sigma^2}(\beta_j) + (1-\gamma_j)\delta_0\,, \quad \text{for } j = 1, \cdots, q, \tag{2b}$$

$$\sigma^2|\boldsymbol{\lambda}, w \sim f(\sigma^2)\,, \tag{2c}$$

$$\boldsymbol{\gamma}|\,w \sim f_w(\boldsymbol{\gamma})\,. \tag{2d}$$

Here, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q) \in \{0,1\}^q =: \Gamma$ indexes a set of selected predictors, and $\beta_{\boldsymbol{\gamma}}$ is a vector containing $\{\beta_j : \gamma_j = 1\}$. The hierarchical prior is given in (2b) – (2d). Specifically, (2c) assigns a prior to σ^2, that is independent of $\boldsymbol{\lambda}$ and w. A common non-informative prior for σ^2 is an inverse-gamma distribution with large mean and variance, as it is a conditionally conjugate prior for the Bayesian normal and the Bayesian lasso model. Alternatively, one can assign $\log(\sigma)$ a uniform prior over $(-\log(\sigma_0), \log(\sigma_0))$ for some large constant σ_0. Next, (2d) specifies a prior on $\boldsymbol{\gamma}$. A simple choice is the independent Bernoulli distribution with success probability w. Here w is a hyperparameter and we discuss ways to specify it in sec. 3.2. We mention that $\boldsymbol{\gamma}$ can also be assigned priors that incorporate information on the covariance structure of the predictors. See,

for e.g., Chipman (1996), Geweke (1996), Yuan & Lin (2005), Li & Zhang (2010).

The regression coefficients are assigned a prior in (2b), that comprises a point mass at zero, δ_0, and a continuous part, $f_{\boldsymbol{\lambda},\sigma^2}(\beta_j)$. Given a penalty function $p(\cdot;\boldsymbol{\lambda})$, we specify

$$f_{\boldsymbol{\lambda},\sigma^2}(\beta_j) = \sigma^{-1} c_{\boldsymbol{\lambda}} \exp\left\{-p\left(\frac{\beta_j}{\sigma};\boldsymbol{\lambda}\right)\right\}.$$

When the penalty function is taken to be $p_{\text{mn}}(\cdot\,;\boldsymbol{\lambda})$, the mnet penalty in Table 1, we call the model in (2) the Bayesian mnet (bmnet) model. Any mnet prior with $\kappa > 0$ is a proper prior, which has a normalizing constant $c_{\boldsymbol{\lambda}} = 1/\int_{\mathbb{R}} \exp\left\{-\frac{1}{2}p(t;\boldsymbol{\lambda})\right\}dt$, that either has an analytical form or is easy to approximate numerically. Note that proper priors must be assigned for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ in the BVS model (2), as improper priors cause indeterminacies in posterior model probabilities (Liang et al., 2008; Berger & Pericchi, 2001). So we eliminate any mnet prior with $\kappa = 0$, namely any "mc prior", from further discussion.

3.2. Choice of hyperparameters

The bmnet model provides users a lot of flexibility, as it becomes the Bayesian normal, lasso, enet models, or any particular bmnet model by varying the hyperparameter $\boldsymbol{\lambda}$. Nevertheless, as $\boldsymbol{\lambda}$ resides deep in the hierarchy of the bmnet model, it is challenging to specify the value for some or all of $(\boldsymbol{\lambda}, w) = (\lambda_1, \lambda_2, \kappa, w)$ based on intuition. Three possible solutions are, choosing the hyperparameter values using cross-validation, estimating them using the empirical Bayes (EB) method, and imposing a prior on them. Performing a k-fold cross-validation requires fitting the bmnet model k times for each combination of $(\lambda_1, \lambda_2, \kappa, w)$ over a grid in the four dimensional space, hence expensive computationally. The EB method refers to setting $(\boldsymbol{\lambda}, w)$ to its maximum likelihood estimate (mle), $(\boldsymbol{\lambda}^{\text{EB}}, w^{\text{EB}}) = \arg\max_{(\boldsymbol{\lambda},w)} f(Y|\boldsymbol{\lambda}, w)$, where $f(Y|\boldsymbol{\lambda}, w) = \sum_{\boldsymbol{\gamma}} \int\int f(Y|\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\lambda}, w) f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2|\boldsymbol{\lambda}, w)\, d\boldsymbol{\beta}\, d\sigma^2$. Although the exact mle is hard to obtain, it is possible to approximate it with a Monte Carlo EM algorithm (Casella, 2001). We do not pursue the EB solution in this paper, other than mentioning that, a simple modification to our MCMC algorithm in sec. 4 allows one to carry out the needed Monte Carlo EM algorithm.

In this paper, we focus on the third solution, namely the fully Bayesian approach, that assigns a prior on $(\boldsymbol{\lambda}, w)$. Of course, we still have to specify the prior distribution for $(\boldsymbol{\lambda}, w)$. But this is an easier task than before because the posterior is usually more robust to different choices of priors on the hyperparameters than to the hyperparameters themselves. Similar to Hans (2011), we assign the following independent priors to the hyperparameters:

$$\lambda_1 \sim \text{Gamma}\left(L, \frac{\nu}{2}\right), \lambda_2 \sim \text{Gamma}\left(R, \frac{\nu}{2}\right), \kappa \sim \text{Gamma}\left(k_a, \frac{k_b}{2}\right), \text{ and } w \sim \text{Beta}(a, b), \tag{3}$$

such that λ_1 has prior mean $2L/\nu$ and so on. The above prior amounts to assigning independent priors for $\lambda_{\text{tot}} = (\lambda_1 + \lambda_2) \sim \text{Gamma}(L + R, \nu/2)$ and $\alpha = \frac{\lambda_1}{\lambda_{\text{tot}}} \sim \text{Beta}(L, R)$. By default, we set $L = R = a = b = 1$, resulting in a uniform prior on α and w respectively. Otherwise, if we have prior knowledge, say, that the proportion of true predictors is mostly likely to be within .1 and .3, then (a, b) can be specified as the solution to $B_{(a,b)}(.1) = .025$ and $B_{(a,b)}(.3) = .975$, where $B_{(a,b)}$ is the cdf of Beta(a, b). Further, we set $\nu = 1$, so that the variance of λ_{tot} is $2/\nu = 2$ times as large as its mean. (We experimented with smaller values of ν, that is, larger prior mean and variability in λ_{tot}, and they yield similar posterior distributions of λ_1, λ_2 and other variables in our studies, and hence are not presented.) To set k_a and k_b, we aim for relatively flat priors on κ such that the maximum concavity parameter of its mnet PR counterpart, $\kappa^* = \kappa/n$, centers at the default value, $1/3$, as recommended in Huang et al. (2016+). The correspondence between κ and κ^* is explained in the supplement.

3.3. Inference based on the posterior distribution

Recall that Y denotes the observed data, with model (2), they yield the posterior $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\lambda}, w|\, Y)$, on which all inference will be based. One may use different ways to summarize the posterior distribution for different tasks such as variable selection and prediction.

For variable selection, the first method we consider is the median probability model (MPM), which consists of variables with inclusion probabilities $\pi(\gamma_j = 1|Y)$ greater than .5. Barbieri & Berger (2004) shows that the MPM is predictively optimal in some sense, and is also frequently the model that has the highest posterior probability. Another method is to obtain a credible interval for each coefficient, say at level 90%, and consider a predictor important if the corresponding credible interval excludes zero. We will compare these methods in sec. 5.2.

For predicting the response y^{new} of a new observation $x^{\text{new}} = (x_1^{\text{new}}, \ldots, x_q^{\text{new}})^\top$, the Bayesian solution that counts for model uncertainty is an average over all models (or a few top models) according to their posterior probabilities (Raftery et al., 1997), namely $(x^{\text{new}})^\top \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = \mathrm{E}(\boldsymbol{\beta}|Y) = \sum_\gamma \mathrm{E}(\beta|\gamma, Y)\pi(\gamma|Y)$. In addition to point estimates, credible intervals provide a range of most probable values for parameters or predictions. For example, a $(1-\alpha)$ credible interval for $\mathrm{E}(y|x^{\text{new}})$ consists of the upper and the lower $\alpha/2$ quantile of the distribution of $(x^{\text{new}})^\top \boldsymbol{\beta}$ where $\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta}|Y)$, and a $(1-\alpha)$ credible interval for y^{new} consists of the upper and the lower $\alpha/2$ quantile of the posterior predictive distribution, given by

$$\pi(y^{\text{new}}|Y) = \sum_{\boldsymbol{\gamma}} \int\int\int\int \phi(y^{\text{new}}; (x^{\text{new}})^\top \boldsymbol{\beta}, \sigma^2)\, \pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\lambda}, w|Y)\, d\boldsymbol{\beta}\, d\sigma^2\, d\boldsymbol{\lambda}\, dw\,. \quad (4)$$

Note that there are no analytical expressions for the above estimators and credible intervals, and we develop MCMC algorithms in sec. 4 to solve the computing problem.

4. COMPUTING FOR THE BAYSIAN MNET MODEL

In this section, we develop MCMC algorithms for the posterior distribution of the model in (2). First, we consider fixing $\boldsymbol{\lambda}$ and w, and denote the posterior distribution by $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2|Y, \boldsymbol{\lambda}, w)$. Note that the simple Gibbs sampler that updates each component of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2)$ in turn will not work. Because if $\beta_j = 0$ then γ_j will remain 0 given $(\boldsymbol{\beta}, \boldsymbol{\gamma}_{(-j)})$, and if $\beta_j \neq 0$ then γ_j will remain 1, making the Markov chain reducible. One solution is to include a latent variable β_j^* for $j = 1, \cdots, q$, such that $\beta_j = \beta_j^* \gamma_j$. A continuous prior can then be assigned to $\boldsymbol{\beta}^* = (\beta_1^*, \cdots, \beta_q^*)$, resulting in a posterior for $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}, \sigma^2)$, for which the simple Gibbs sampler is irreducible. Successful execution of this idea in certain BVS models include Kuo & Mallick (1998), Carlin & Chib (1995), and Dellaportas et al. (2002), to list a few. The latter two allow dependent priors for β_j^* and γ_j, hence requires specifying a "pseudo prior" on $\beta_j^*|\gamma_j = 0$. Pseudo priors are named so because they do not affect the posterior of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2)$, yet they can greatly influence the efficiency of the corresponding Gibbs samplers.

To avoid tuning pseudo priors in the above solution, we introduce a block Gibbs sampler in sec. 4.1 that updates the $(q+1)$ blocks $\{(\beta_1, \gamma_1), \cdots, (\beta_q, \gamma_q), \sigma^2\}$ in turn. This algorithm has a similar structure to one developed in Geweke (1996). It can be extended to compute Bayesian models that assign priors to all or some of the hyperparameters in $(\boldsymbol{\lambda}, w)$. For example, if a non-degenerate prior is assigned to $(\lambda_1, \lambda_2, \kappa, w)$, then one can run a block Gibbs sampler that updates the $(q+5)$ blocks $\{(\beta_1, \gamma_1), \cdots, (\beta_q, \gamma_q), \sigma^2, \lambda_1, \lambda_2, \kappa, w\}$ in turn. The challenge is that all conditional distributions of the hyperparameters given other components do not follow standard distributions. To solve this problem, we develop an adaptive random-walk Metropolis-Hastings (ARWMH) within Gibbs sampler in sec. 4.2. Finally, in sec. 4.3, we discuss how to estimate various posterior quantities using the Monte Carlo sample produced by our algorithm.

4.1. Updating $(\beta_1, \gamma_1), \cdots, (\beta_q, \gamma_q)$ with a Gibbs sampler

First, we consider updating $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ conditional on $\sigma^2, \boldsymbol{\lambda}$ and w. The posterior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}|\sigma^2, \boldsymbol{\lambda}, w, Y)$ is written in short as $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}|Y)$ in this section. Let $v_{(j)}$ denote the vector v deprived of its jth component. We form a block Gibbs sampler that updates (γ_j, β_j) as a block for $j = 1, \ldots, q$, according to $\pi(\gamma_j, \beta_j|\boldsymbol{\gamma}_{(j)}, \boldsymbol{\beta}_{(j)}, Y)$. Each update can be done by first drawing from $\pi(\gamma_j|\boldsymbol{\gamma}_{(j)}, \boldsymbol{\beta}_{(j)}, Y)$, a Bernoulli distribution with success rate $(1 + O_j)^{-1}$, and then drawing from $\pi(\beta_j|\boldsymbol{\gamma}, \boldsymbol{\beta}_{(j)}, Y)$, a piecewise normal density. Specifically, let $r_j = Y - X_{\boldsymbol{\gamma}_{(j)}}\boldsymbol{\beta}_{\boldsymbol{\gamma}_{(j)}}$, then

$$
\begin{aligned}
O_j &= \frac{f(\gamma_j = 0|\boldsymbol{\beta}_{(j)}, \boldsymbol{\gamma}_{(j)}, Y)}{f(\gamma_j = 1|\boldsymbol{\beta}_{(j)}, \boldsymbol{\gamma}_{(j)}, Y)} = \frac{f(\gamma_j = 0, \boldsymbol{\beta}_{(j)}, \boldsymbol{\gamma}_{(j)}, Y)}{f(\gamma_j = 1, \boldsymbol{\beta}_{(j)}, \boldsymbol{\gamma}_{(j)}, Y)} \\
&= \frac{f(Y|\boldsymbol{\beta}_{(j)}, \boldsymbol{\gamma}_{(j)}) f(\boldsymbol{\beta}_{(j)}|\boldsymbol{\gamma}_{(j)}) f(\boldsymbol{\gamma}_{(j)})}{\int_{\mathbb{R}} f(y|\beta_j, \gamma_j = 1, \boldsymbol{\beta}_{(j)}, \boldsymbol{\gamma}_{(j)}) f(\beta_j|\gamma_j = 1) f(\gamma_j = 1) f(\boldsymbol{\beta}_{(j)}|\boldsymbol{\gamma}_{(j)}) f(\boldsymbol{\gamma}_{(j)}) d\,\beta_j} \\
&= \frac{\exp\left\{-\frac{r_j^\top r_j}{2\sigma^2}\right\} \Pi_{s \neq j: \gamma_s = 1} \left[\frac{c_{\boldsymbol{\lambda}}}{\sigma} \exp\left\{-p_{\text{mn}}\left(\frac{\beta_s}{\sigma}; \boldsymbol{\lambda}\right)\right\}\right] w^{|\boldsymbol{\gamma}_{(j)}|} (1-w)^{q - |\boldsymbol{\gamma}_{(j)}|}}{\int_{\mathbb{R}} \exp\left\{-\frac{(r_j - X_j\beta_j)^\top (r_j - X_j\beta_j)}{2\sigma^2}\right\} \Pi_{s: \gamma_s = 1} \left[\frac{c_{\boldsymbol{\lambda}}}{\sigma} \exp\left\{-p_{\text{mn}}\left(\frac{\beta_s}{\sigma}; \boldsymbol{\lambda}\right)\right\}\right] w^{|\boldsymbol{\gamma}_{(j)}|+1} (1-w)^{q - |\boldsymbol{\gamma}_{(j)}| - 1} d\,\beta_j} \\
&= (1-w) \Bigg/ \left[w \frac{c_{\boldsymbol{\lambda}}}{\sigma} \int_{\mathbb{R}} \exp\left\{-\frac{1}{2\sigma^2}[X_j^\top X_j \beta_j^2 - 2 X_j^\top r_j \beta_j] - p_{\text{mn}}\left(\frac{\beta_j}{\sigma}; \boldsymbol{\lambda}\right)\right\} d\,\beta_j\right] \\
&=: (1-w) \Bigg/ \left[w \frac{c_{\boldsymbol{\lambda}}}{\sigma} (I_1 + I_2 + I_3)\right],
\end{aligned}
$$

where I_1, I_2 and I_3 are integrals over $(0, \lambda_1\sigma/\kappa)$, $(-\lambda_1\sigma/\kappa, 0)$ and $(-\infty, \lambda_1\sigma/\kappa) \cup (\lambda_1\sigma/\kappa, \infty)$, respectively. All three integrals have analytical expressions based on the standard Normal cdf and pdf, which we provide in the supplement. Further,

$$
\begin{aligned}
\pi(\beta_j|\boldsymbol{\gamma}, \beta_{(j)}, Y) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left[(Y - X_{\boldsymbol{\gamma}}\beta_{\boldsymbol{\gamma}})^\top (Y - X_{\boldsymbol{\gamma}}\beta_{\boldsymbol{\gamma}}) - \sum_{s: \gamma_s = 1} p_{\text{mn}}\left(\frac{\beta_s}{\sigma}; \boldsymbol{\lambda}\right)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[(r_j - X_j\beta_j)^\top (r_j - X_j\beta_j) - p_{\text{mn}}\left(\frac{\beta_j}{\sigma}; \boldsymbol{\lambda}\right)\right]\right\}.
\end{aligned}
$$

This is a mixture of three truncated Normal distributions weighted by (I_1, I_2, I_3), hence easy to draw samples from. Detailed expressions of the truncated Normals are again in the supplement.

4.2. Updating $(\sigma^2, \boldsymbol{\lambda}, w)$ with an adaptive random-walk Metropolis-Hastings within Gibbs sampler

To explore the posterior distribution $\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\lambda}, w|Y)$ for model (2) with prior (3), we next develop ways to update $(\sigma^2, \boldsymbol{\lambda}, w)$, where $\boldsymbol{\lambda}$ denotes $(\lambda_1, \lambda_2, \kappa)$ for the mnet prior. Let $|\boldsymbol{\gamma}| = \sum_j \gamma_j$. The distribution of w conditioning on all others is simply Beta$(|\boldsymbol{\gamma}| + a, p - |\boldsymbol{\gamma}| + b)$. For σ^2 and $\boldsymbol{\lambda}$, their respective conditional distributions are not standard, and require a Metropolis-Hastings (MH) step within the Gibbs sampler. Following the examples in Roberts & Rosenthal (2009), we choose to update the logarithm of each of these components one at a time, using a random-walk MH (RWMH) scheme with a $\text{N}(0, v^2)$ increment from the current value, subject to the usual MH acceptance rate. The key to an efficient RWMH algorithm is selecting a good "step size" v. If v is too large, the acceptance rate will be low, and the Markov chain rarely moves;

and if v is too small, despite a high acceptance rate, the Markov chain moves in tiny steps. In either situation, it takes a long time for the Markov chain to explore the state space of the posterior distribution properly. Note that optimal values of v differ for different components, and are highly dependent on the posterior distribution itself, which is influenced by model specification as well as the data. Therefore, it is very difficult to select good values of v in practice. Indeed, an earlier version of our RWMH algorithm based on predetermined values of v converges slowly, despite a lot of effort spent on tuning v for each data analysis problem. We resort to an adaptive strategy that adjusts v gradually, that drives the acceptance rate to near $.44$, the optimal acceptance rate for one-dimensional proposals under certain assumptions (Roberts et al., 1997). Also, Roberts & Rosenthal (2001) suggests there is no need to fine tune the acceptance rate, as any rate between 0.1 and 0.5 usually performs close to optimal for RWMH on smooth target densities.

Take the component κ for example, we start with an arbitrary $\log(\kappa^{(0)})$ and $v = v^{(0)}$. Then at the ith iteration, given the current value $\log(\kappa^{(i)})$, we propose $\log(\kappa') \sim \mathrm{N}(\log(\kappa^{(i)}), v^2)$, and accept it with probability $\alpha_i = \frac{\pi(\log(\kappa'))}{\pi(\log(\kappa^{(i)}))}$. Here, $\pi(\cdot)$ denotes the conditional posterior density of $\log(\kappa)$ given other variables. To update v every r iterations, let $\overline{\alpha}^{(q)} = r^{-1} \sum_{i=qr+1}^{(q+1)r} \alpha_i$, for $q = 0, 1, 2, \cdots$. Typically, acceptance rate decreases in v, which prompts us to lower v if $\overline{\alpha}^{(q)} > .44$, and vice versa. Hence, for the next r iterations, we use $v^{(q+1)} = v^{(q)} + \delta_q(\overline{\alpha}^{(q)} - .44)$, with $\delta_q = v^{(0)}/q$, following Atchadé & Rosenthal (2005). Trials of $r = 1, 5, 10$ did not yield much difference for our examples, and we set $r = 1$.

Further, since δ_q is small for a large q, changes in the adaptive algorithm are negligible after, say, 1000 iterations. For the examples in this project, the step size for each sampled variable appears stabilized by the end of the burn-in period, and running plots of the acceptance rate suggest they have reached close-to-optimal values. So, we run the adaptive version of the algorithm only during the burn-in period. This strategy avoids the theoretical complications that accompany adaptive algorithms, which we remark in sec. 4.3.

4.3. Computing posterior features based on the Monte Carlo sample

Denote the Monte Carlo sample by $\left\{ \left(\boldsymbol{\beta}^{(i)}, \boldsymbol{\gamma}^{(i)}, \left(\sigma^2\right)^{(i)}, \boldsymbol{\lambda}^{(i)}, w^{(i)} \right) ; \, i = 1 \ldots, N \right\}$, which is obtained after discarding a burn-in period of length $N/10$ to reduce the influence of the starting point of the Markov chain. We use the sample means and quantiles to estimate the posterior means and quantiles. For example, the posterior mean and the αth quantile of β_1 are estimated by the sample mean and the sample αth quantile of $\{\beta_1^{(i)}\}$, respectively, and a credible interval for $\mathrm{E}(y^{\text{new}}|Y)$ can be estimated using a pair of sample quantiles of $\{(x^{\text{new}})^\top \boldsymbol{\beta}^{(i)}, i = 1, \cdots, N\}$. Consider a more complicated example of approximating a $(1-\alpha)$ credible interval for the prediction of y^{new}. We need q_α, the αth quantile for $\pi(y^{\text{new}}|Y)$ in (4), which can be approximated as the following. Step one, we screen a coarse grid of values, say $Q_1 < \cdots < Q_K$, and estimate the lower tail probability of each using $\hat{F}(Q_k) = \frac{1}{N} \sum_{i=1}^{N} \Phi\big(Q_k; (x^{\text{new}})^\top \boldsymbol{\beta}^{(i)}, \sigma^{2(i)}\big)$. Here, Φ is the standard Normal cdf. Step two, we find the pair of grid points (Q_l, Q_u) such that $\hat{F}(Q_l) \le \alpha \le \hat{F}(Q_u)$. If $\hat{F}(Q_l)$ (or $\hat{F}(Q_u)$) is close enough to α, then Q_l (or Q_u) is our estimate for q_α. Otherwise, we zoom in on (Q_l, Q_u) by setting a new grid $Q_l = Q'_1 < \cdots < Q'_{K'} = Q_u$ for some $K' \ge 3$, and repeat steps one and two, until some grid point Q is found, for which $\hat{F}(Q)$ is as close as desirable to α. Then Q is our estimate for q_α.

All the aforementioned Monte Carlo estimators are consistent. Firstly, our bmnet models use proper priors, which result in proper posteriors. Secondly, for the non-adaptive version of our RWMH within Gibbs sampler, the transition density from any one point in the state space to any other is positive. Therefore, the corresponding Markov chain is Harris ergodic (Tan, 2009), and the ergodic theorem holds. That is, sample averages and quantiles are strongly consistent for the corresponding posterior means and quantiles, respectively.

Remark 1. *For the adaptive version of our algorithm, there is currently no easy-to-check condition to show that the ergodic theorem hold. The closest result that we are aware of concerns RWMH (Atchadé & Rosenthal, 2005), but the transition kernel of our algorithm is a composition of several such RWMH kernels. Hence, we avoid such theoretical issues by running the adaptive scheme only during the burn-in period. We enjoy benefits of both worlds: there is no need to manually tune the step sizes, and Monte Carlo estimators are easily consistent.*

5. SIMULATION STUDIES

For simulations conducted in this section, the design matrix X contains n observations of q-dimensional predictor vectors drawn from the multivariate normal distribution $\mathrm{N}_q(0, \Sigma)$, where Σ is a "uniform" covariance matrix such that $\Sigma_{ij} = \rho$ for $i \neq j$ and 1 otherwise. The response Y, is generated from $\mathrm{N}_n(X\boldsymbol{\beta}, \sigma^2 I_n)$, where $\boldsymbol{\beta}$ denotes the true vector of coefficients. We consider a sparse situation where the first 10 elements of $\boldsymbol{\beta}$ are 1, and the rest are 0. We consider two levels of correlation, $\rho = .3$ and $.9$.

Also, various signal to noise ratios, $s = \sqrt{\mathrm{E}[(X\boldsymbol{\beta})^\top (X\boldsymbol{\beta})]}/\sigma$ are investigated. We achieve different s by varying σ against fixed $\boldsymbol{\beta}$, to simulate various levels of contamination of fixed signals. For readers who understand relative effect sizes better, $s = (1, 2, 4, 8)$ convert to $|\beta_1|/\sigma = (.16, .33, .66, 1.32)$ for $\rho = .3$, and $(.10, .21, .41, .84)$ for $\rho = .9$, respectively, for the given setup. Note that Johnson & Rossell (2010) defined practical significance in linear regression to be $|\beta_1|/\sigma > 0.2$. Hence $s \in (1, 8)$ represents a fairly wide range of values.

5.1. Setups with various correlation and signal strength at $n = 100$ and $q = 150$

In this section, we demonstrate the potential advantage of using the bmnet model in high-dimensional regression problems with correlated predictors. We consider the simulation setup described above, with 200 datasets of size $(n, q) = (100, 150)$ generated at each combination of $\rho \in \{.3, .9\}$ and $s \in \{1, 2, 4, 8\}$. We compare eight different methods, the lasso, the enet, the mcp, the mnet, three bmnet models (benet, bmnet-fx, and bmnet-rd), and an example of the Bayesian model with a non-local prior (nlp). The penalty parameters in the first four methods are chosen by ten-fold cross-validations. Here, the λ's are screened over a fine grid, while the κ in the mcp and the mnet method is chosen from $\{1/6, 1/3\}$, based on the recommendation of Huang et al. (2016+). Also, we follow Zou & Hastie (2005) and refer to $\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\|Y - X\boldsymbol{\beta}\| + \lambda_1 \sum_j |\beta_j| + \frac{\lambda_2}{2} \sum_j \beta_j^2 \right)$ as the naive enet solution. The enet estimator is $(1 + \frac{\lambda_2}{2})\widetilde{\boldsymbol{\beta}}$, that corrects the bias in $\widetilde{\boldsymbol{\beta}}$ due to over-shrinkage. For the three bmnet models, the hyperpriors for $\boldsymbol{\lambda}$ are chosen as described in sec. 3.2. Specifically, the benet fixes κ at 0, the bmnet-fx fixes κ at $n/3$, while the bmnet-rd uses a hyperprior on κ. Recall setting $\kappa = n/3$ corresponds to specifying the maximum concavity parameter $\kappa^* = 1/3$ in the mnet PR method in some sense. Fixing κ this way is not necessarily optimal, rather, it is natural to want to compare this naive strategy to other methods. Indeed, the simulation results in sec. 5 suggest that the bmnet model with random κ either performs the best, or very close to being the best among the three bmnet models. For inference, the overall posterior mean $\mathrm{E}(\boldsymbol{\beta}|Y)$, namely the BMA, is used to estimate $\boldsymbol{\beta}$, and the MPM is the selected model. The last Bayesian model uses a non-local prior called the product inverse moment (piMOM) prior. Following the practice in Johnson & Rossell (2012) and Rossell & Telesca (2016+), the prior dispersion is set to $\tau = 0.133$, which assigns prior probability 0.01 to $|\beta_j|/\sigma < 0.2$. Then $\boldsymbol{\beta}$ is estimated by BMA, and numerical approximation to the highest posterior probability model (HPM) is used for variable selection.

Define the prediction mean squared error (pmse) as $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, where $\hat{\boldsymbol{\beta}}$ is the coefficient estimates. Figure 2 shows the median of the relative pmse of each method against a benchmark, the mnet. Figure 3 shows the false discovery proportion (FDP), the false negative

proportion (FNP), and the number of variables selected (NVS) by each method. Only 6 methods are graphed for clarity, with the lasso and the mcp left out due to their poor performances in several high correlation cases. Summary statistics for all methods are presented in the supplement, including the standard errors of the median pmse's.

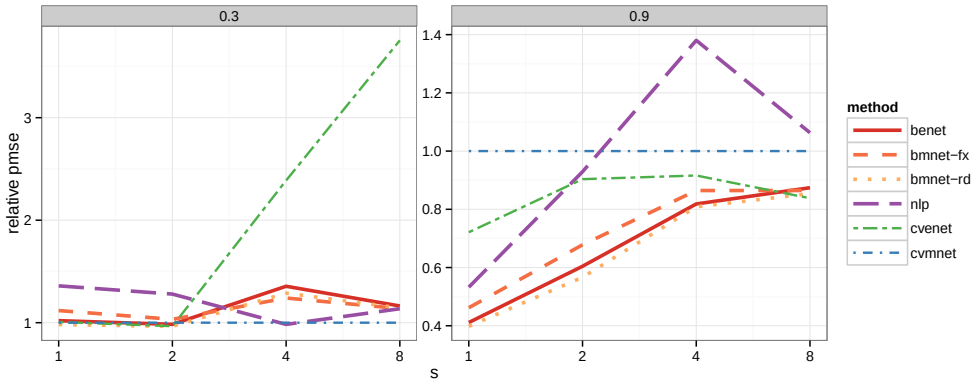


FIGURE 2: Medians of the relative pmse of different methods to that of the mnet, calculated over 200 replications at each (s, ρ). The left and the right panels correspond to the $\rho = 0.3$ and 0.9 cases, respectively. (Color online.)

Figure 2 and 3 show that, in the high correlation case, the bmnet methods have much lower pmse than the PR methods and the nlp. And in the low correlation case, the bmnet methods are comparable to the mnet in prediction and variable selection, when the latter is designed to be the most effective (Huang et al., 2016+).

For variable selection, FDP and FNP of the bmnet methods decrease toward 0 at comparable rate to the best performers for both ρ, as s increases. At low s, all Bayesian methods select very few predictors, which results in larger FNR and smaller FDR compared to the PR methods. This is a reasonable strategy of variable selection given the effect sizes are barely practically significant. (Nevertheless, the threshold for the posterior inclusion probabilities can be easily lowered from 0.5 for the MPM, if the user intend to find the most relevant predictors despite their small impact on the response.) At moderate to high s, models selected by the Bayesian methods are equal to or close to the true model. In comparison, the enet tends to over-fit, even when the signal is strong; while the mnet over-fits when the signal is weak, but becomes more accurate as the signal grows stronger. The comparison of variable selection performances among different methods is not clear at $\rho = .9$, which is understandable given all predictors, true or null, are highly correlated.

Regarding nlp, it appears the strongest in both prediction and variable selection in the less challenging cases of small ρ and large s, for which the bmnet methods perform almost as well. In the more challenging cases, especially when ρ is high, the bmnet has much smaller pmse than the nlp at any signal strength.

After all, the bmnet methods predict the best or close to the best in all cases, and have clear advantage over other methods in the high collinearity cases. Further, their FDP and FNP decrease toward 0 at comparable rates to the best performers. Among the three bmnet methods, bmnet-rd has the best overall prediction performance, though their differences are not great. We view the robustness of the benet and the bmnet methods to the hyperpriors as a valuable quality. In contrast, depending on whether the enet or the mnet penalty function is used in a PR framework,

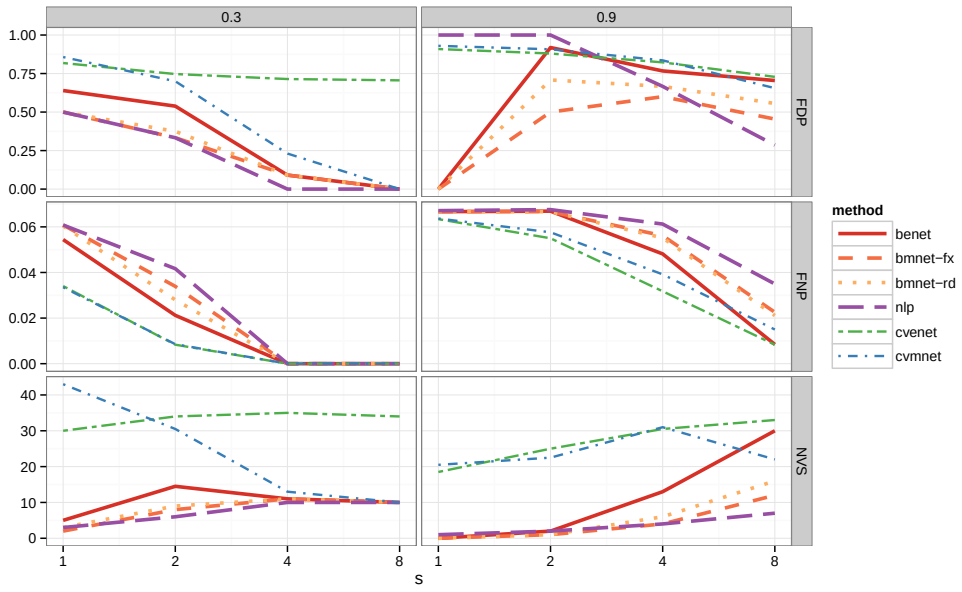


FIGURE 3: Variable selection performances of six methods at each (s, ρ). Variables are selected using the MPM for the three bmnet methods (including the benet), and the HPM for the nlp method, respectively. The top, the middle, and the bottom panels display the median over 200 replications of the FDP, the FNP and the NVS, respectively. The left and the right panels correspond to $\rho = 0.3$ and 0.9, respectively. (Color online.)

users often obtain fairly different models and inference results.

Remark 2. *The lasso and the enet models are fitted using the R package* glmnet *(Friedman et al., 2010); the naive enet, the mcp and the mnet methods are fitted using the R package* ncvreg *(Breheny & Huang, 2011); the nlp method is carried with the R package* mombf *(Rossell et al., 2014); and all Bayesian mnet models are fitted using R code developed for this paper. To carry out each bmnet method, we ran the MCMC algorithm for* 10^4 *iterations beyond a burn-in period of size* 10^3*. Each run took less than five minutes on a 2.93GHz Intel Xeon W3540 running Linux. We experimented with random starts for the Markov chains, and estimates for various posterior features such as the posterior inclusion probability and the posterior mean of coefficients were fairly stable. The Monte Carlo standard error of various estimates are calculated using the R package* mcmcse *(Flegal & Hughes, 2012), and are less than* 1% *the size of the estimates.*

5.2. Variable selection by the Bayesian mnet method and other methods as n increases

Recall we mentioned two ways to do variable selection based on a BVS model with the mnet prior, one using the MPM and the other using credible intervals of $\boldsymbol{\beta}$. Here, we demonstrate their selection consistency under the simulation setup described in the beginning of section 5. In particular, we fix the number of true predictors at 10, and let the number of null predictors to grow with n, such that $q = 5\sqrt{n}$. This setup is more challenging than those with fixed q. For clarity, only the bmnet method with random κ is shown. In addition, thanks to the suggestion of a referee, we also experiment with a version of the Bayesian model with the mnet that does not have the γ variable, for which a predictor is selected if the 80%, 90%, or 95% credible interval

for its coefficient excludes zero. A few other popular variable selection methods are also included for comparison.

Simulations are ran for 100 replications each for various s, at $(n, q) = (25, 25), (100, 50)$ and $(400, 100)$. Figure 4 displays result for $s = 8$ only, due to space limit. Recall $s = 8$ converts to $\beta_1/\sigma = 1.32$ and $.84$ for $\rho = .3$ and $.9$ respectively. The figure suggests that, the chance that the bmnet MPM (bmnet-rd, in solid line) pinpointing the true model increases towards 1 as n increases. As for variable selection based on credible intervals of the Bayesian mnet models, the bmnet model without γ (bmnet-rd-ng 0.9CI, in dotted line) performs poorly, which is not surprising given q is large and the true model is sparse. However, the credible interval methods are among the best when applied to the bmnet models with γ (bmnet-rd 0.9CI, in dashed line). Variable selection results based on 80% and 95% credible intervals are not much different from the ones shown, and are omitted for clarity of the graphs.

For other methods, the success rates of the cvmnet and the nlp also converge to 1 respectively, with nlp performing particularly well when $\rho = .3$. Whilst the enet always over-fit the true model.

When $\rho = .3$, the plot (not shown) at $s = 4$ looks similar to that at $s = 8$, but the probabilities are lower in general. For much smaller s, the signals are too weak for any of these methods to find the true model exactly. The results at $\rho = .9$ are similar, but the task is more challenging: at $s = 4$, none of the methods we studied perfectly identified the true model in any repetition.

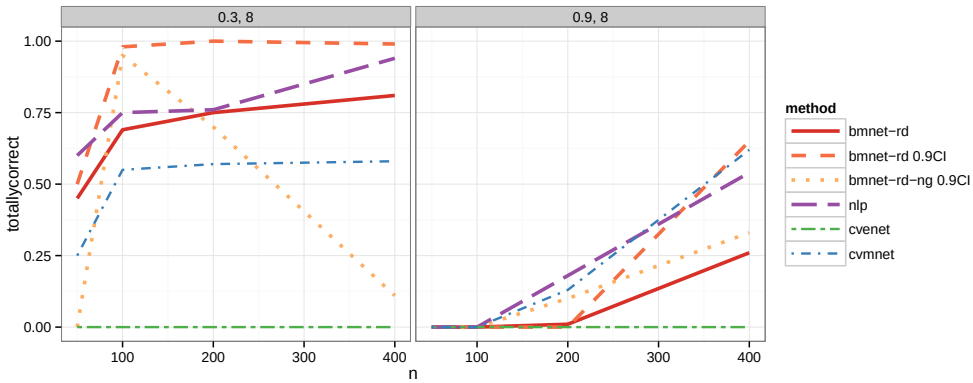


FIGURE 4: The left and the right plots show the proportion of times the selected model equals the true model for $\rho = 0.3$ and 0.9, at $s = 8$. (Color online.)

6. BREAST CANCER DATA EXAMPLE

Here, we study a breast cancer dataset obtained from *http://cancergenome.nih.gov*, that consists of expression measurements in the log scale of 17814 genes from 536 patients. One of the genes is BRCA1, which is the first gene identified that increases the risk of early onset breast cancer. It is of interest to find other genes that have expression levels related to BRCA1. An initial screening is used to find genes that have sufficient expression levels and variations across subjects. Besides BRCA1, 654 genes are identified that meet the following requirements: (a) the coefficient variation exceeds 1, (b) the range exceeds 1, (c) the standard deviation exceeds 0.5, and (d) the absolute value of the marginal correlation to the response variable exceeds 0.2.

We randomly split the data into a training set of 400 patients to build models, and a test set of 136 patients to help evaluate the models. For clarity of presentation, we focus on six methods. The three PR methods are the naive enet (enet.n), the mcp, and the mnet method, each based on a 10-

fold cross-validation. And the three BVS methods are the Bayesian enet model (benet), the bmnet model with $\kappa^* = \kappa/n$ fixed at $1/3$ (bmnet-fx), and the bmnet model that assigns a Gamma prior on κ^* (bmnet-rd) with mean $1/3$. For all three BVS methods, we assign the Beta$(2, 10)$ prior for w, so that the expected number of selected genes, $|\boldsymbol{\gamma}|$, follows a beta-binomial$(654, 2, 10)$ distribution. This reflects our prior belief that useful subsets of predictors include between 8 to 309 genes (these are the .005 and the .995 quantiles of $|\boldsymbol{\gamma}|$), and the prior mode equals $|\boldsymbol{\gamma}| = 65$, one-tenth the number of candidate genes.

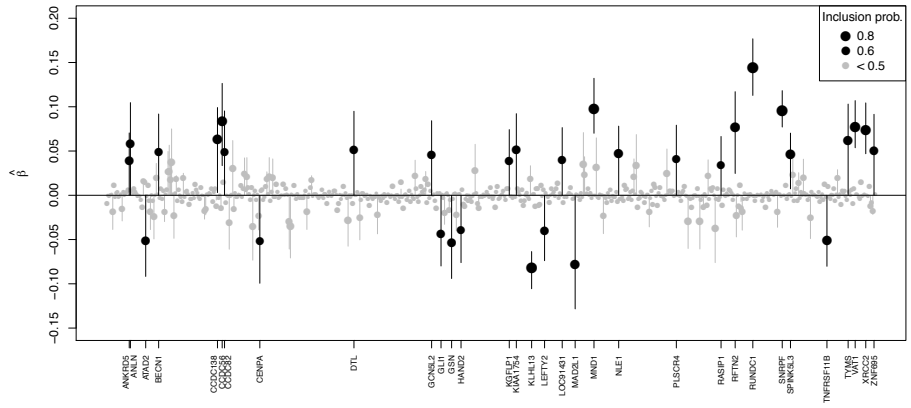
For each Bayesian method, a Markov chain is run for $N = 10^5$ iterations after a burn-in period of 10^4 iterations. We examined the trace plot of many parameters and found no flag for mixing problems. We also obtained Monte Carlo standard errors (mcse) of the estimates for various posterior means and quantiles using the batch means method with batch size $\sqrt{N}$, carried out using the R package mcmcse. For example, under the bmnet-fx model, the posterior mean of the coefficient for the gene "RUNDC1" is estimated to be 0.1442, with an mcse of 0.0008; and the first and the third posterior quartiles of this coefficient are estimated to be 0.1135 and 0.1767, respectively, each with a mcse of 0.0008, which is negligible relative to the size of the estimates.

For the bmnet-rd method, over 99% of the sample of κ^* falls in the interval $(0, 0.007)$. Since the mnet penalty function reduces to the enet penalty function when $\kappa^* = 0$, the above suggests that the posterior distribution of the bmnet-rd model and that of the benet model are very close to each other. Indeed, the MPMs selected by the benet and the bmnet-rd method contain 52 and 49 genes each, and they both include all 32 genes selected by the bmnet-fx method. Figure 5 demonstrates how we can visualize the selection results together with estimates for the coefficients (shown only for the most different pair of bmnet methods, the benet and the bmnet-fx, due to space limit). Selected genes are shown in black discs, where the size of a disc is proportional to the posterior inclusion probability of the corresponding gene. In addition, posterior means of the coefficients determine the position of the discs, and the 50% credible intervals correspond to the line segments that extend above and below the disc. (Note that credible intervals at any level can be easily provided. But had credible intervals at a higher level, say 90%, been drawn, the line segments would contain zero for any predictor with inclusion probability lower than 0.95, which holds for almost all predictors. Hence, we choose to plot at level 50% to convey more information in one graph.)

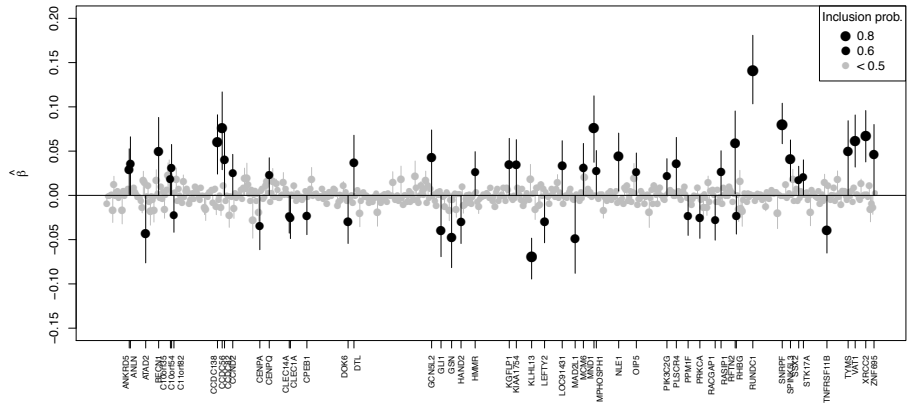
Despite the similarity in pattern of the two graphs in Figure 5, it is interesting to note how they differ in the details. Relative to the benet method, the bmnet-fx method selects fewer genes (fewer black discs), assigns most of the selected genes larger coefficients in absolute value (larger distance of black discs from 0), and assigns other genes smaller inclusion probabilities (smaller grey discs). For the genes selected by bmnet-fx, all but one has larger coefficient in absolute value than that by benet, and the estimates are on average 23% larger. This is a result of enforcing a relatively large κ^* value to discourage the shrinking of large coefficients.

Using (4), we make predictions and construct prediction intervals for the response variable of the subjects in the test set. As shown in Figure 6, the 90% (80%) prediction intervals contain the true responses for 84% (77%) of the subjects, respectively. That is, the coverage rates of the prediction intervals agree with their nominal levels.

We also analyzed the breast cancer data using several PR methods. The naive enet method, the mcp method, and the mnet method selected 112, 24 and 84 genes respectively. Note that the naive enet method and the mnet methods produced two sets of $\hat{\boldsymbol{\beta}}$ that agree well with each other, while the mcp method fitted a much more parsimonious model, with much higher coefficient estimates for the selected genes. With these methods, limited inference procedures were available, including obtaining point estimates of the coefficients, $\hat{\boldsymbol{\beta}}$, as shown in Figure 7, and making predictions for new observations using $(x^{\text{new}})^\top \boldsymbol{\beta}$. Although not the main focus of this section,



(a) The bmnet.fx method selected 32 genes in its median probability model.



(b) The benet method selected 52 genes in its median probability model.

FIGURE 5: Inference for coefficients based on (a) the bmnet model with fixed κ, and (b) the benet model. For either model, each disc represents a gene, where the posterior mean of its coefficient determines the center of the disc, and the 50% credible interval of the coefficient (the inter-quartile range) determines whiskers that extend above and below the disc. Further, the size of each disc is proportional to the inclusion probability of the corresponding predictor. Color is used to distinguish the selected predictors (in black) and the unselected ones (in grey) according to the MPM.

we evaluated the pmse of each method for the test data, which can be found in Table 2.

For the different Bayesian methods allowed in our bmnet framework, instead of claiming which one is preferred over the others for analyzing this dataset, we believe it is a desirable feature that different Bayesian models produce agreeable results and reasonable predictions, yet users with different goals in selecting variables have the flexibility to adjust the prior on hyper-parameters such as κ to achieve different models, all of which predict well.

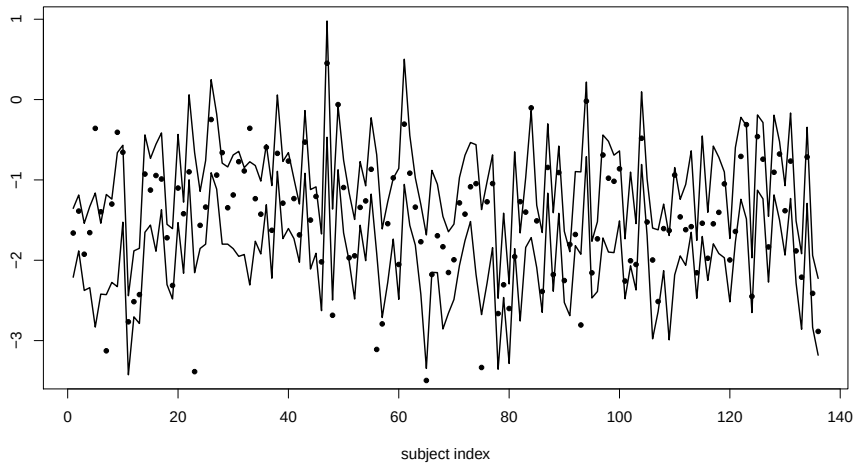


FIGURE 6: To predict the expression level of BRCA1 for the 146 subjects in the test set, the upper and the lower bounds of their 90% prediction intervals are connected and displayed using two lines. The observed expression levels are shown in dots, with 123, that is, a little over 84% of them captured by the corresponding prediction intervals.

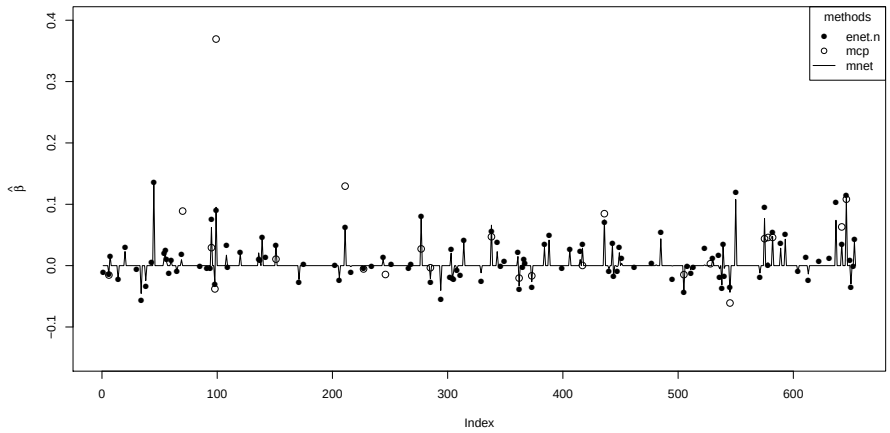


FIGURE 7: Estimated coefficients based on the naive enet, the mcp and the mnet method.

7. DISCUSSIONS

In this paper, we introduced a new class of prior, called the mnet prior, for BVS models. This results in a general bmnet framework that allows more flexibility in modeling and variable selection compared to existing methods. Compared to the mnet PR method that inspired the mnet prior, many more useful inferences can be done based on our bmnet model. We highlight the

TABLE 2: The breast cancer data was split in to a training set of size 400 and a test set of size 136. This table provides, the pmses and the number of genes selected (# VS) for each method.

	enet.n	mcp	mnet	benet	bmnet-fx	bmnet-rd
Pmse	32.54	38.20	33.25	30.21	30.81	30.36
# VS	112	24	84	52	32	49

type of graph that is shown in Figure 5, which displays the practical significance, the statistical significance and the importance of each potential predictor simultaneously.

On variable selection consistency We used simulation studies to heuristically check the variable selection consistency of our method, in the sense that the chance that the selected model equals the true model approaches one. Still, it is desirable to establish selection consistency in different situations in theory. Note that the mnet prior is a kind of "local prior" using the terminology of Johnson & Rossell (2012). They showed that a common problem for local priors is that, if all the hyperparameter values are fixed as n grows, strong consistency can not be achieved when q is greater than $O(\sqrt{n})$, in the sense that the posterior probability of the true model does not converge to one. However, Narisetty & He (2014), among others, has demonstrated that scaling the priors properly as functions of n and q will allow strong consistency under mild conditions, even when q grows with n at a nearly exponential rate. Still, the proper scaling is only known up to a constant, and hence simulation studies like those conducted in this paper are still helpful to obtain practical guidance. One major advantage of using local priors than the non-local priors in practice is that, local priors are much less sensitive to the choice of hyperparameters than that of non-local priors. Another challenge of using non-local prior is that they produce highly multi-modal posterior distributions for the coefficients unless when n is large and q is small (such as an example of $n = 1000$ and $q = 2$ in Rossell & Telesca (2016+)). The multi-modality makes it controversial to perform estimation and prediction through BMA and requires more investigation.

On computing An important part of the proposed bmnet method is an algorithm that handles its computation. Specifically, we developed an ARWMH within Gibbs sampler to draw samples from the posterior distribution, and listed several ways to estimate posterior quantities based on the samples. When applied to the simulation study and the real data examples, basic diagnostics for the convergence of the Markov chain suggest that our algorithm works reasonably well. However, there is much work to do to evaluate the convergence of the Markov chains. For example, despite reporting the Monte Carlo standard error for various estimates, we stopped short of proving that the Central limit theorems (CLT) hold for the estimators. Such result would hold (under additional moment conditions) if the Markov chain is geometrically ergodic, that is, if the chain converges to the posterior distribution at a geometric rate. But these are indeed hard analytical exercises, and remain open questions. Also, note that we did not try to identify the HPM for bmnet models in any of our examples. This is due to the large size of q we considered. Indeed, with any $q \geq 20$, the number of models, 2^q will exceed one million. And any affordable Monte Carlo sample size is unlikely to be large enough to well explore the entire model space, or to approximate the probability of individual models. This is indeed the curse of dimensionality that we do not expect to solve perfectly using any MCMC method, not even if iid samples can be drawn from the posterior distribution. Therefore, when q is large, we focus on inference procedures that only require quantities related to low-dimensional marginals of the posterior distribution. For example, the MPM is much less demanding to compute, and usually predicts better than the HPM.

While focusing on relatively easy inference targets in high-dimensional problems, it is desirable to lessen the computing burden by developing MCMC algorithms that are more efficient

than our ARWMH within Gibbs sampler. One idea is to explore within the current block Gibbs sampler framework. Our algorithm adopted a strategy that updates each block in turn, which is called a systematic scan. Other strategies are available, such as a random scan that updates the more important and the slower-mixing blocks more often. Indeed, the scanning scheme itself can be adaptive, as was carried out successfully in Richardson et al. (2010) for high-dimensional problems.

ACKNOWLEDGEMENTS

The authors thank Luke Tierney, Hani Doss, Patrick Breheny and three anonymous referees for helpful suggestions. Jian Huang was supported by the National Science Foundation Grant and the National Institutes of Health ???.

BIBLIOGRAPHY

Atchadé, Y. F. & Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* 11 815–828.

Barbieri, M. & Berger, J. (2004). Optimal predictive model selection. *The Annals of Statistics* 32 870–897.

Berger, J. & Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison. *Lecture Notes-Monograph Series* 135–207.

Breheny, P. & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics* 5 232–253.

Carlin, B. & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B* 57 473–484.

Casella, G. (2001). Empirical Bayes Gibbs sampling. *Biostatistics* 2 485–500.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics* 24 17–36.

Dellaportas, P., Forster, J. J. & Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing* 12 27–36.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96 1348–1360.

Flegal, J. M. & Hughes, J. (2012). *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA and Minneapolis, MN. R package version 1.0-1.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33 1–22. URL `http://www.jstatsoft.org/v33/i01/`.

George, E. & McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* 7 339–374.

George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88 881–889.

Geweke, J. (1996). Variable selection and model comparison in regression. *Bayesian Statistics* 5 609–620.

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association* 106 1383–1393.

Huang, J., Breheny, P., Lee, S., Ma, S. & Zhang, C. (2016+). The Mnet method for variable selection. *Statistica Sinica*. DOI:10.5705/ss.202014.0011

Ishwaran, H. & Rao, J. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics* 33 730–773.

Johnson, V. E. & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society Series B* 72 143–170.

Johnson, V. E. & Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* 107 649–660.

Johnstone, I. M. & Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics* 33 1700–1752.

Kuo, L. & Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B* 65–81.

Li, F. & Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association* 105.

Li, Q. & Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis* 5 151–170.

Liang, F., Paulo, R., Molina, G., Clyde, M. & Berger, J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* 103 410–423.

Lockhart, R., Taylor, J., Tibshirani, R. J. & Tibshirani, R. (2014). A significance test for the LASSO. *Annals of statistics* 42 413–468.

Mitchell, T. J. & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83 1023–1032.

Narisetty, N. N. & He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics* 42 789–817.

O'Hara, R. B. & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis* 4 85–117.

Park, T. & Casella, G. (2008). The Bayesian LASSO. *Journal of the American Statistical Association* 103 681–686.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92 179–191.

Richardson, S., Bottolo, L. & Rosenthal, J. S. (2010). Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Statistics* 9 539–569.

Roberts, G. O., Gelman, A. & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability* 7 110–120.

Roberts, G. O. & Rosenthal, J. S. (2001). Markov chains and de-initializing processes. *Scandinavian Journal of Statistics* 28 489–504.

Roberts, G. O. & Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* 18 349–367.

Rossell, D., Cook, J. D., Telesca, D. & Roebuck, P. (2014). *mombf: Moment and Inverse Moment Bayes factors*. R package version 1.5.9.

Rossell, D. & Telesca, D. (2016+). Non-local priors for high-dimensional estimation. *Journal of the American Statistical Association*. DOI:10.1080/01621459.2015.1130634

Smith, M. & Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* 75 317–343.

Tan, A. (2009). *Convergence rates and regeneration of the block Gibbs sampler for Bayesian random effects models*. Ph.D. thesis, Department of Statistics, University of Florida.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSP. *Journal of the Royal Statistical Society Series B* 58 267–288.

Yuan, M. & Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association* 100 1215–1225.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38 894–942.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B* 67 301–320.