Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Mallick and Yi, J Biomet Biostat 2013, S1
DOI: 10.4172/2155-6180.S1-005

Review Article Open Access

Bayesian Methods for High Dimensional Linear Models

Himel Mallick and Nengjun Yi*

Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, USA

Abstract

In this article, we present a selective overview of some recent developments in Bayesian model and variable selection methods for high dimensional linear models. While most of the reviews in literature are based on conventional methods, we focus on recently developed methods, which have proven to be successful in dealing with high dimensional variable selection. First, we give a brief overview of the traditional model selection methods (viz. Mallow's Cp, AIC, BIC, DIC), followed by a discussion on some recently developed methods (viz. EBIC, regularization), which have occupied the minds of many statisticians. Then, we review high dimensional Bayesian methods with a particular emphasis on Bayesian regularization methods, which have been used extensively in recent years. We conclude by briefly addressing the asymptotic behaviors of Bayesian variable selection methods for high dimensional linear models under different regularity conditions.

Keywords: Bayesian hierarchical models; Penalized regression; Regularization; Shrinkage methods; Bayesian variable selection; High dimensional linear models; Bayesian model selection; MCMC; Posterior consistency; Nonlocal priors; Bayesian subset regression

Introduction

Linear models are probably the most widely used statistical models to investigate the influence of a set of predictors on a response variable. In practice, only a small subset of potential covariates actually has an influence on the response variable, whereas the effect of most predictors is very small or even zero. Since, model misspecification can have a significant impact on a scientific result, correctly identifying relevant variables is an important issue in any scientific research. If more predictors are included in the model, a high proportion of the response variability can be explained. On the other hand, overfitting (inclusion of predictors with null effect) results in a less reliable model with poor predictive performance. The problem of variable selection becomes more challenging for high dimensional problems, particularly when the number of predictors greatly exceeds the number of observations. High dimensional problems arise in a variety of scientific fields, such as bioinformatics, medicine, genetics, etc. High dimensionality could lead us to models that are very complex, which poses serious challenges in estimation, prediction and interpretation. Therefore, many classical approaches to variable selection cease to be useful due to computational infeasibility, model non-identifiability, or both.

Various methods have been developed over the years for dealing with variable selection in high dimensional linear models. Very recently, much work has been done in the direction of Bayesian framework. Unlike non-Bayesian methods, Bayesian analysis enables one to deal with model uncertainty by averaging over all possible models. Moreover, Bayesian methods have the ability to significantly reduce the actual complexity involved in the estimation procedure by incorporating prior information within the data into the model estimation technique. With ever-increasing computing power, these methods are increasingly becoming popular and gaining more and more insight and considerations for high dimensional analysis.

In this review, we present a selective overview of some recent developments in Bayesian model and variable selection methods for high dimensional linear models. While most of the reviews in literature are based on conventional methods, we focus on recently developed methods, which have been used extensively over the years. The review is structured as follows. First, we give a brief overview of the traditional commonly used model selection methods followed by a discussion on some recently developed approaches for linear models, which have proven to be successful in dealing with high dimensional variable selection. Then, we mention some conventional Bayesian variable and model selection methods, along with some recently developed Bayesian approaches, with a particular emphasis on Bayesian regularization methods. We conclude by briefly addressing the asymptotic behaviors of Bayesian high dimensional variable selection methods for linear models under different regularity conditions.

Classical Model Selection Methods for Linear Models

In statistical tradition, commonly used methods for model selection are backward, forward and stepwise selection, where in every step, predictors are added to the model or eliminated from the model, according to a precisely defined testing rule. Besides accurate prediction, the primary goal in model selection is also to come up with meaningful, interpretable and parsimonious model. Traditional methods such as stepwise regression fall short in one or more of these criteria. To overcome these shortcomings, several information-type methods have been developed, which aim to provide a trade-off between model complexity and goodness-of-fit of a model. We describe the standard normal linear model and Bayesian hierarchical normal linear model in later subsections. Then, we briefly review four commonly used methods *viz*. Mallow's Cp [1], AIC [2] and BIC [3] for standard normal models, and DIC [4] for Bayesian hierarchical normal linear models, which remain the most popular model selection methods used for linear models and hierarchical linear models respectively.

Standard normal linear model

In the simplest case of a normal linear regression model, it is

***Corresponding author:** Nengjun Yi, Department of Biostatistics, Ryals Public Health Building 317F, University of Alabama at Birmingham, Birmingham, AL 35294, USA, Tel: (205) 934-4924; Fax: (205) 975-2540; E-mail: nyi@uab.edu

Received April 14, 2013; **Accepted** May 28, 2013; **Published** June 01, 2013

Citation: Mallick H, Yi N (2013) Bayesian Methods for High Dimensional Linear Models. J Biomet Biostat S1: 005. doi:10.4172/2155-6180.S1-005

assumed that the mean of the response variable can be described as a linear function of a set of predictors. Mathematically, in a normal linear regression setup, we have the following model

$$y = X\beta + \varepsilon, \tag{1}$$

where y is the $n\times 1$ vector of centered responses; X is the $n\times p$ matrix of standardized regressors; β is the $p\times 1$ vector of coefficients to be estimated and ε is the $n\times 1$ vector of independent and identically distributed normal errors with mean 0 and variance σ^2. The classical estimator in linear regression is the Ordinary Least Squares (OLS) estimator $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$, which is obtained by minimizing the residual sum of squares (RSS) given by

$$Q_{OLS}(\beta) = (y - X\beta)^T (y - X\beta). \tag{2}$$

The OLS estimator has some attractive statistical properties, in the sense that it is the best linear unbiased estimator (BLUE), provided the classical assumptions hold [5]. Also, it coincides with the maximum likelihood estimator (MLE) for normal linear models.

Bayesian hierarchical normal linear model

In Bayesian hierarchical normal linear model, we assume that the distribution of the dependent variable y is specified conditional on the parameters β and σ^2 as

$$y^{n\times 1} \Big| \beta, \sigma^2 \sim N(X\beta, \sigma^2 I_n). \tag{3}$$

Then, any prior information on (β,σ) is incorporated by specifying a suitable prior distribution $p(\beta,\sigma^2)$ on them. This second-level model has its own parameters known as hyperparameters, which are usually estimated from the data. After observing the data, the prior distribution, $p(\beta,\sigma^2)$ is updated by the corresponding posterior distribution, which is obtained as

$$p(\beta,\sigma^2 \big| y) = \frac{p(y|\beta,\sigma^2)p(\beta,\sigma^2)}{\int p(y|\beta,\sigma^2)p(\beta,\sigma^2)d(\beta,\sigma^2)}. \tag{4}$$

The posterior distribution contains all the current information about the parameters. Ideally one might fully explore the entire posterior distribution by sampling from the distribution, using Markov chain Monte Carlo (MCMC) algorithms [6]. Due to its ability to incorporate specific hierarchical structure of the data (correlation among the predictors), hierarchical modeling is often more efficient than traditional approaches [7].

Mallow's C_p

For a subset model with $k \le p$ explanatory variables, the C_p statistic proposed by Mallows [1], is defined as

$$C_p = \frac{RSS(k)}{s^2} - n + 2p, \tag{5}$$

where s^2 is the MSE for the full model containing p explanatory variables and $RSS(k)$ is the residual sum of squares for the subset model containing k explanatory variables. In practice, C_p is usually plotted against p for a collection of subset models under consideration and models with C_p approximately equal to p are taken as acceptable models, in the sense of minimizing the total bias of the predicted values [5]. Woodrofe [8] showed that C_p is a conservative model selector, which tends to overfit. Nishii [9] showed that C_p is not consistent in selecting the true model, and often tends to select a larger model as $n \to \infty$.

Akaike's Information Criteria (AIC)

The AIC of Akaike [2] is defined as

$$AIC = -2\log L + 2p, \tag{6}$$

where L is the likelihood function evaluated at the MLE. Given a set of candidate models, the 'best' model is the one with the minimum AIC value. Similar to Mallow's Cp, AIC is not model selection consistent [9]. Here, the consistency of a model selection criterion means that the probability of the selected model being equal to the true model converges to 1. More information about AIC can be found in Burnham and Anderson [10]. The asymptotic approximation on which the AIC is based is rather poor when n is small [11]. Therefore, Hurvich and Tsai [12] proposed a small-sample correction, leading to the AIC_c statistic defined by

$$AIC_c = AIC + \frac{2p(p+1)}{(n-p-1)}. \tag{7}$$

AIC_c converges to AIC as n gets larger, and therefore, it is preferred to AIC regardless of the sample size [11].

Bayesian Information Criteria (BIC)

While AIC is motivated by the Kullback-Leibler discrepancy of the fitted model from the true model, Schwarz [3] derived BIC from a Bayesian perspective by evaluating the leading terms of the asymptotic expansion of the Bayes factor. The BIC is defined as

$$BIC = -2\log L + p\log n, \tag{8}$$

where L is the likelihood function evaluated at the MLE. Similar to AIC, model with minimum BIC is chosen as the preferred model from a set of candidate models. It is well known that neither AIC nor BIC performs better all the time. However, unlike AIC, BIC is a consistent model selection technique, which means, as the sample size n gets large enough, the lowest BIC model will be the true model, with probability 1 [10,11]. For a comparison of AIC and BIC, refer to Kundu and Murali [13] or Yang [14].

Deviance Information Criteria (DIC)

For model selection in Bayesian hierarchical normal linear models, Spiegelhalter et al. [4] proposed the generalization of AIC and BIC defined as

$$DIC = D + 2pD, \tag{9}$$

where D= -2log L is the deviance evaluated at the posterior mean of the parameters, and pD is the effective number of parameters calculated as the difference between posterior mean deviance and deviance of posterior means. Like AIC and BIC, models with smaller DIC are better supported by the data. DIC is particularly useful when the MCMC samples are easily available, and is valid only when the joint distribution of the parameters is approximately multivariate normal [4]. DIC tends to select over-fitted models, which has been addressed by Ando [15], although very little is known about its performance in high dimensional models. As noted by Gelman et al. [16], various other difficulties (apart from overfitting) have been noted with DIC, but there has been no consensus on an alternative.

High Dimensional Methods

In the setting of a linear regression model, if the number of covariates p is of the polynomial order or exponential order of the

sample size , i.e., $p = O(n^{\kappa})$ or $p = O(\exp(n^{\kappa}))$ for $\kappa > 0$, then it is called a high dimensional problem [17]. Despite their popularity, classical model selection criteria such as Mallow's C_p, AIC, BIC tend to select more variables than necessary for high dimensional linear models, especially when the number of regressors increases with the sample size [18,19]. Also, Yang and Barron [20] argues that in some cases the overfitting problem can be substantial, resulting in severe selection bias, which damages predictive performance for high dimensional models. To overcome this, various model selection criteria for high dimensional models have been introduced recently. Wang et al. [21] proposed a modified BIC (mBIC), which is consistent when p is diverging slower than n. Chen and Chen [22,23] developed a family of extended Bayesian information criteria (EBIC), for variable selection for high dimensional problems. On the other hand, a large amount of effort has gone into the development of regularization methods for simultaneous variable selection and coefficient estimation. Regularization methods mitigate modeling biases and achieve higher prediction accuracy in high dimensional linear models by shrinking the coefficients and providing meaningful estimates, even if the model includes a large number of, and/or highly correlated predictors. We describe the extended Bayesian information criteria and regularization methods in the following subsections, before gradually moving to Bayesian methods for high dimensional linear models.

Extended Bayesian Information Criteria (EBIC)

The family of EBIC is indexed by a parameter γ in the range [0,1]. The extended BIC (EBIC) is defined as

$$EBIC = -2\log L + p\log n + 2\gamma \log p, \tag{10}$$

where L is the likelihood function evaluated at the MLE, and $\gamma > 0$ is a tuning parameter. The original BIC is a special case of EBIC with $\gamma = 0$. The mBIC is also a special case of EBIC in an asymptotic sense; i.e. it is asymptotically equivalent to the EBIC with $\gamma = 1$. Chen and Chen [23] established the model selection consistency of EBIC, when $p = O(n^{\kappa})$ and $\gamma > 1 - \frac{1}{2\kappa}$ for any $\kappa > 0$, where consistency implies that as $n \to \infty$, the minimum EBIC model will converge in probability to the 'true' model. Among other developments, General Information Criterion (GIC) proposed by Shao [24] is known to be consistent in high dimensions. Kim et al. [25] showed that EBIC is asymptotically equivalent to GIC.

Regularization methods

Similar to information-type methods, various regularization methods have been developed to overcome the problem of overfitting in high dimensional linear models. It is well known that OLS often does poorly in both prediction and interpretation in high dimensions. Despite its nice statistical properties, it is highly unstable in the presence of multicollinearity. Also, if $p \gg n$, it produces a non-unique estimator, since X is less than full rank (non-identifiability). Motivated by this, regularization methods (also known as penalized likelihood or shrinkage method) with various penalties have been developed, which have proven to be successful and model selection consistent for high dimensional linear models [26,27].

The problem of interest involves estimating a sparse vector of regression coefficients by minimizing an objective function Q that is composed of a loss function (without loss of generality, most commonly used least squares loss function (RSS) is considered, although least absolute deviation and negative log-likelihood is also common) plus a penalty function P, i.e.

$$Q(\beta) = (y - X\beta)^T (y - X\beta) + P_{\lambda}(\beta), \tag{11}$$

where P is a function of the coefficients indexed by a parameter $\lambda > 0$, which controls the degree of penalization. Typically, the penalty function P has the following properties [28]:

1. It is symmetric about the origin, i.e. $P(0) = 0$,

2. is non-decreasing in $(0, \infty)$.

This approach produces a spectrum of solutions depending on the value of λ. Such methods are often referred to, as regularization methods, and λ is called the regularization parameter (or tuning parameter). The penalty function serves to control the complexity of the model and provides criteria for variable selection and model comparison, by imposing some constraints on the parameters. The form of $P_{\lambda}(\beta)$ determines the general behavior of regularization methods. Small penalties lead to large models with limited bias, but potentially high variance; large penalties lead to the selection of models with fewer predictors, but with less variance. A variety of penalty terms have been proposed, among which the most popular ones are ridge regression, the lasso and the elastic net. We summarize these methods in Table 1. For a more comprehensive review on regularization methods, refer Bickel and Li [29].

The bridge family: Regularization methods date back to the proposal of ridge regression by Hoerl and Kennard [30], who suggested minimizing the following objective function:

$$Q_{Ridge}(\beta) = (y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^{p} \beta_j^2. \tag{12}$$

As a continuous shrinkage method, ridge regression achieves better prediction performance than OLS through a bias–variance trade-off (biased estimates with lower variance). However, ridge regression cannot produce a parsimonious model, as it always keeps all the predictors in the model.

Frank and Friedman [31] introduced bridge regression, a broad class of penalized regression, which is obtained by minimizing

$$Q_{Lasso}(\beta) = (y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^{p} \left|\beta_j\right|^{\alpha}, \tag{13}$$

Method	Tuning parameter	Penalty function
Lasso	λ	$\lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$
Ridge	λ	$\lambda \sum_{j=1}^{p} \beta_j^2$
Bridge	λ	$\lambda \sum_{j=1}^{p} \lvert\beta_j\rvert^{\alpha}$
Adaptive Lasso	$\lambda_1, \ldots, \lambda_p$	$\sum_{j=1}^{p} \lambda_j \lvert\beta_j\rvert$
Elastic Net	λ_1, λ_2	$\lambda_1 \sum_{j=1}^{p} \lvert\beta_j\rvert + \lambda_2 \sum_{j=1}^{p} \beta_j^2$
Group Lasso	λ	$\lambda \sum_{k=1}^{K} \sqrt{\sum_{j=1}^{m_k} \beta_{k_j}^2}$

Table 1: Different penalty functions.

where λ is the tuning parameter and α is the concavity parameter, as it controls the concavity of the objective function (13). It includes lasso and ridge as special cases (corresponding to $\alpha = 1$ and $\alpha = 2$ respectively). Although Frank and Friedman [31] did not solve for the estimator of bridge regression for any given $\alpha \geq 0$, they indicated that optimum choice of α would yield reasonable predictor. The bridge estimator does variable selection when $\alpha \leq 1$ and shrinks the coefficients when $\alpha > 1$ [32]. These three estimators viz. ridge, lasso and bridge are together referred to as the bridge family.

Among penalized regression techniques, the most popular and widely used method in statistical literature is the Least Absolute Shrinkage and Selection Operator (LASSO). The lasso of Tibshirani [33] is obtained by minimizing

$$Q_{Lasso}(\beta) = (y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|. \quad (14)$$

Compared to ridge regression, a remarkable property of lasso is that it can shrink some coefficients exactly to zero, and therefore, can automatically achieve variable selection. Intuitively, this can be explained by the fact that $|\beta_j|$ is much larger than $|\beta_j|^2$ for small β_j and thus the L_1-penalty $\lambda \sum_{j=1}^{p} |\beta_j|$ enforces some β_j's exactly to zero.

Figure 1 shows the behavior of these three penalty functions in a two-parameter case, β_1 and β_2. To obtain the regularized estimators, we essentially seek the points at which the objective function contour first "hits" the constraint. Lasso, ridge and bridge penalty functions have constraints shaped like a square, circle and star, respectively. As a consequence of the different shapes, lasso is likely to involve variable selection ($\beta_1 = 0$ or $\beta_2 = 0$), as well as parameter estimate shrinkage, and ridge yields mainly parameter estimate shrinkage; in contrast, bridge induces an even higher chance of variable selection than lasso, because the star shape of bridge makes the contour even more likely to hit one of the points ($\beta_1 = 0$ or $\beta_2 = 0$), than does the diamond shape of lasso.

Some generalizations: The lasso has demonstrated excellent performance in many situations. As a consequence, most of the developments in recent years are focused on the lasso and related problems. However, despite its promising nature, there are three inherent drawbacks of lasso [35]. Firstly, due to the nature of the convex optimization problem, the lasso method cannot select more predictors than the sample size. But, in practice, there are often studies that involve much more predictors than the sample size, e.g. microarray gene expression data analysis, cloud detection through analysis of satellite images, classification of spam emails, and many others. Secondly, when there is some group structure among the predictors, the lasso estimator usually selects only one predictor from a group, while ignoring others. Thirdly, when the predictors are highly correlated, lasso performs unsatisfactorily. We discuss some of the alternatives and generalizations that have been proposed to overcome the above limitations of lasso.

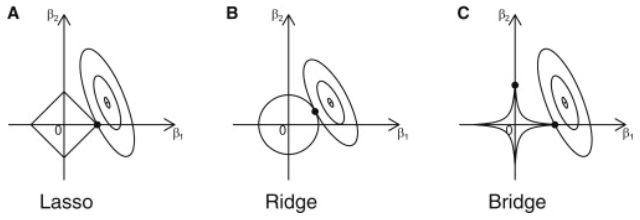

Figure 1: A graphical illustration of the properties of three different penalty functions. The eclipses represent the contours of the objective functions (A–C). The square, round, and star shapes represent the lasso, ridge, and bridge constraint, respectively. The dots are the points where contours are "tangent" to the constraints, i.e., the regularized estimates. Note that, in lasso (A) or bridge (C), the constraint is discontinuous at zero. If the contour first touches the constraint at point zero, the corresponding parameter estimate is zero and variable selection is achieved [34].

The lasso uses a unique tuning parameter λ to equally penalize all the coefficients. In practice, due to the single tuning parameter, lasso can either include irrelevant variables or over-shrink large coefficients [36], which is critical for high dimensional problems. To address the issue, Zou [37] introduced the adaptive lasso that uses a weighted L_1-penalty

$$P_\lambda(\beta) = \sum_{j=1}^{p} \lambda_j |\beta_j|, \quad (15)$$

where λ_j is the tuning parameter corresponding to the j^{th} coefficient β_j, $j = 1(1)p$. The intuition of adaptive lasso is to shrink coefficients differently by shrinking important variables slightly and unimportant variables heavily.

To address the issue of variable selection for grouped variables, Yuan and Lin [38] proposed the group lasso estimator, in which the penalty is given by

$$P_\lambda(\beta) = \lambda \sum_{k=1}^{K} \sqrt{\beta_{k_1}^2 + \beta_{k_2}^2 + \ldots + \beta_{k_{m_k}}^2}, \quad (16)$$

where K is the no. of groups with $\sum_{k=1}^{K} m_k = p$ and β_{k_j} is the coefficient corresponding to j^{th} predictor in the k^{th} group, $j = 1(1)m_k$, $k = 1(1)K$. With appropriately chosen tuning parameter, the group lasso can shrink all coefficients in a group to zero or keep all coefficients in a group in the model.

Zou and Hastie [35] proposed the elastic net estimator to achieve improved performance in situations when there is multicollinearity and grouping among predictors. The penalty term in elastic net is a convex combination of the lasso penalty and the ridge penalty, i.e.

$$P_\lambda(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2, \quad (17)$$

where $\lambda_1 > 0$, $\lambda_2 > 0$ are two tuning parameters.

The elastic net estimator can be interpreted as a stabilized version of the lasso, and it often outperforms the lasso, especially when there are groups of highly correlated predictors.

Other related regularization methods include the smoothly clipped absolute deviation (SCAD) method [28], the fused lasso [39], the adaptive elastic net [40], the minimax concave penalty (MCP) [41], the adaptive bridge estimator [32], the Dantzig selector [42], and the group bridge estimator [43].

Optimization algorithms: Various optimization algorithms have been proposed to obtain the lasso and related estimators [44]. Notably, the least angle regression (LARS) [45], and the coordinate descent algorithm [46,47], are the most computationally efficient ones. Given the tuning parameters, these algorithms are extremely fast (e.g. the computational load of LARS is same as that of a single OLS fit), thus making the penalized regression approaches extremely popular in high dimensional data analysis.

Limitations of regularization methods: In spite of being

theoretically attractive, here are at least three serious disadvantages of frequentist penalized regression approaches:

1. Penalized regression is essentially an optimization problem that only provides a point estimate of β. Nevertheless, people usually also need to know the level of confidence of the estimates, such as the confidence interval (or credible interval), and the p-value. This problem can be addressed by applying bootstrap sampling [33], but it is computationally intensive. Kyung et al. [48] showed that the bootstrap estimates of the standard errors of the lasso estimates might be unstable, and are not consistent if true $\beta_j = 0$.

2. The penalized regression approaches need to preset the tuning parameter(s). The commonly used method is to use cross-validation [49], to choose 'optimal' values of the tuning parameters. However, cross-validation can be computationally costly, e.g. for the adaptive lasso, it is very challenging to choose multiple tuning parameters. Moreover, cross-validation is a standard way to assess the predictive accuracy of the model. Choosing tuning parameters using cross validation of the prediction error can result in unsatisfactory performance, when the main goal is to identify relevant variables.

3. In frequentist framework, it may be challenging to deal with complicated multilevel/hierarchical structures of data, and to incorporate external information (for example, data with specific group level information) into the model.

High Dimensional Bayesian Variable and Model Selection Methods

In the Bayesian framework, the model selection problem is transformed to the form of parameter estimation: rather than searching for the single optimal model, a Bayesian will attempt to estimate the posterior probability of all models, within the considered class of models (or in practice, of all models with non-negligible probability) [50]. In many cases, this question is asked in variable-specific form (variable selection): i.e. the task is to estimate the marginal posterior probability that a variable should be in the model [50]. Thus, variable selection can be considered a special case of model selection.

In this section, we review some recently developed Bayesian methods for high dimensional variable selection. There is an enormous amount of literature on Bayesian variable selection methods [50-55]. Some conventional Bayesian variable selection methods are Gibbs Variable Selection (GVS) [50,53,56], Bayesian Model Averaging (BMA) [50,54,56], Stochastic Search Variable Selection (SSVS) [52], Unconditional Priors for Variable Selection [57], Product Space Search [58], and RJMCMC (Reversible Jump MCMC [59]). Other methods include approaches based on Zellner's g-prior [60-62], Bayes factor [63], fractional Bayes factor [64], objective Bayes [65,66], etc. Due to space constraint, it is impossible to discuss all the available methods in this review. Since, detailed and comprehensive reviews on the state-of-the-art Bayesian methods have previously appeared in literature [50-55], here we only mention some of those approaches. Instead, we focus on popular recently developed methods. We particularly focus on the Bayesian regularization methods, which have been proven successful for high dimensional variable selection.

Most of the conventional Bayesian variable selection methods rely on MCMC algorithms by specifying spike and slab priors on the coefficients subject to selection [50-52], requiring computation of marginal likelihood, which is computationally intensive for high dimensional models. Also, posterior sampling of these methods often requires stochastic search over an enormous space of complicated models facilitating slow convergence and mixing, when the marginal likelihood is not analytically tractable [67]. On the other hand, Bayesian regularization methods specify both the spike and slab as continuous distributions, which can be written as scale mixtures, leading to simpler MCMC algorithm with no marginal likelihood being computed. Also, unlike conventional Bayesian methods, Bayesian regularization methods specify shrinkage priors, which enable simultaneous variable selection and coefficient estimation. We also discuss two new model selection approaches *viz*. Bayesian model selection based on nonlocal prior densities proposed by Johnson and Rossell [67], and Bayesian subset regression (BSR) proposed by Liang et al. [68], which are shown to be model selection consistent for high dimensional linear models.

Bayesian regularization methods

Regularization methods are originally developed by the frequentists, and obtaining statistical inference on the regression coefficients is usually difficult, and often requires various kinds of asymptotic approximations. In contrast, a Bayesian approach enables exact inference, even when the sample size is small. Regularization methods naturally lend itself to a Bayesian interpretation, in which the penalization term in penalized regression is the negative log prior of the coefficient. Apart from their easy interpretability, Bayesian methods have some advantages over frequentist methods. Firstly, in MCMC-based Bayesian regularization methods, we have a valid measure of standard error obtained from the posterior distribution, and thus, we can easily obtain interval estimates of the parameters, along with other quantities. Secondly, it is more flexible in the sense that we can estimate the tuning parameter jointly with other parameters of interest. Thirdly, unlike in frequentist framework, it is fairly straightforward to extend a Bayesian model to incorporate multilevel information inherent in the data. Lastly, using MCMC and Gibbs sampler to search for the model space for the most probable variable models, without fitting all possible models is efficient, which avoids time-consuming computation [56].

Regularization methods as hierarchical models: Recent years have seen a resurgence of interest in Bayesian hierarchical modeling techniques. This increasing popularity can be contributed to the fact that hierarchical models are more easily interpreted and handled in the Bayesian framework. Hierarchical models can significantly reduce the 'effective' no. of parameters in a model by linking the coefficients together, or shrinking some of them. In Bayesian hierarchical models, the hyperparameters include shrinkage parameters, which control the complexity of the model, similar to the tuning parameter λ in penalized regression approaches. With the given prior distribution, the log posterior density for linear models become

$$\begin{aligned}&\log p(\beta,\sigma^2,\lambda|y,X) \propto \log p(y|X\beta,\sigma^2) + \\ &\log p(\beta|\sigma^2,\lambda) + \log p(\sigma^2) + \log p(\lambda),\end{aligned} \tag{18}$$

where $p(\sigma^2)$ and $p(\lambda)$ are the prior distributions of σ^2 and λ respectively, $p(y|X\beta,\sigma^2)$ is the likelihood function and $p(\beta|\sigma^2,\lambda)$ is the prior distribution of β. It is to be noted that the posterior mode (MAP) estimate can be obtained by maximizing the posterior density in equation 18. Therefore, the posterior mode estimate is equivalent to the penalized regression estimate, with $\log p(\beta|\sigma^2,\lambda)$ as the penalty. Thus, with particular priors, hierarchical models can lead to similar results as penalized regression approaches.

It is evident that the prior distribution $p(\beta|\sigma^2,\lambda)$ plays an important role in Bayesian regularization methods. For models with a

large number of potential variables, it is reasonable to assume that most of the variables have no or very weak effects, whereas only some have noticeable effects. Therefore, we should set up a prior distribution that gives each coefficient a high probability of being near zero. Such prior distributions are often referred to as shrinkage prior distributions.

A shrinkage prior distribution should have an infinite spike at zero and very heavy tails, thereby strongly shrinking small coefficients to zero, while minimally shrinking large coefficients, and also enable to incorporate hierarchical structure of the data. Therefore, the resulting hierarchical models can effectively remove unimportant variables and reliably estimate the coefficients of important variables simultaneously [6].

Different Hierarchical Formulations

Bayesian lasso: Tibshirani [33] suggested that the lasso estimates can be interpreted as posterior mode estimates, when the regression parameters are assigned independent and identical Laplace priors. A remarkable feature of the double exponential distribution is that it can be presented as a two-level hierarchical model [69], as scale mixtures of normal distributions with independent exponentially distributed variances. The two-level formulation of the double exponential distributions offers advantages of easily interpreting the model and developing computational algorithms. The latent variables $\tau_1^2,..,\tau_p^2$ directly control the amount of shrinkage in the coefficient estimates. If $\tau_j^2=\infty$, there is no shrinkage; if $\tau_j^2=0$, the j^{th} coefficient is shrunk to zero. Although these latent variables are not the parameters of interest, they are useful quantities that allow easy computation [70].

Park and Casella [71] introduced Gibbs sampling for Bayesian lasso, using a conditional Laplace prior specification of the form

$$p(\beta|\sigma^2)=\prod_{j=1}^{p}\frac{\lambda}{2\sqrt{\sigma^2}}\exp\{-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\} \tag{19}$$

and non-informative scale-invariant marginal prior on σ^2, i.e. $p(\sigma^2)\propto 1/\sigma^2$. They pointed out that conditioning on σ^2 is important, as it ensures unimodal full posterior. Lack of unimodality might slow down the convergence of the Gibbs sampler, and make the point estimates less meaningful [48]. The Bayesian formulation of the original lasso, as given in Park and Casella [71], is given by the following hierarchical model:

$$\begin{aligned} &y^{n\times 1}|X,\beta,\sigma^2\sim N(X\beta,\sigma^2 I_n),\\ &\beta^{p\times 1}|\sigma^2,\tau_1^2,..,\tau_p^2\sim\prod_{j=1}^{p}N(0,\sigma^2\tau_j^2),\\ &\tau_1^2,..,\tau_p^2|\lambda\sim\prod_{j=1}^{p}\frac{\lambda^2}{2}e^{-\lambda^2\tau_j^2/2},\\ &\sigma^2\sim p(\sigma^2). \end{aligned} \tag{20}$$

With this formulation, the posterior distribution of β is normal, while the reciprocals of the latent variables are distributed as inverse Gaussian distributions (Table 2). The posterior distribution of σ^2 is inverse gamma distribution. Based on this, Park and Casella [71] formulated the Gibbs sampler for the Bayesian lasso and achieved variable selection by interval estimation. The adaptive version of Bayesian lasso can be obtained similarly by specifying variable-specific tuning parameter (Table 2).

Method	$p(\beta\|\sigma^2,\tau_1,..,\tau_p)$	$p(\tau_1,..,\tau_p)$	$p(\beta\|y,X,\tau^2)$	Posterior Distributions of Latent Parameters
Bayesian lasso	$N_p(0,\sigma^2 D_\tau)$	$\prod_{j=1}^{p}\frac{\lambda^2}{2}e^{\frac{-\lambda^2\tau_j^2}{2}}$	$N_p(A^{-1}X^Ty,\sigma^2A^{-1})$	$\tau_j^{-2}\|\beta,\sigma^2,X,y \sim \text{IG}\left(\frac{\lambda^2\sigma}{\|\beta_j\|},\lambda^2\right)$
Bayesian ridge	$N_p(0,D'_\tau)$	$\prod_{j=1}^{p} lnv-\chi^2(v,s^2)$	$N_p(A'^{-1}X^Ty,\sigma^2A'^{-1})$	$\tau^{-2}\|\beta,X,y \sim \frac{\chi^2_{v+p}}{vs^2+\sum_{j=1}^{p}\beta_j^2}$
Bayesian adaptive lasso	$N_p(0,\sigma^2 D_\tau)$	$\prod_{j=1}^{p}\frac{\lambda_j^2}{2}e^{\frac{-\lambda_j^2\tau_j^2}{2}}$	$N_p(A^{-1}X^Ty,\sigma^2A^{-1})$	$\tau_j^{-2}\|\beta,\sigma^2,X,y \sim \text{IG}\left(\frac{\lambda_j^2\sigma}{\|\beta_j\|},\lambda_j^2\right)$
Bayesian elastic net	$N_p(0,\sigma^2 D^*_\tau)$	$\prod_{j=1}^{p}\frac{\lambda_1^2}{2}e^{\frac{-\lambda_1^2\tau_j^2}{2}}$	$N_p(A^{*-1}X^Ty,\sigma^2A^{*-1})$	$\tau_j^{-2}\|\beta,\sigma^2,X,y \sim IG\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}},\lambda_1^2\right)$
Bayesian group lasso* (K Groups)	$N_{m_k}(0,\sigma^2\tau_k^2 I_{m_k})$	$\prod_{k=1}^{K}Gamma\left(\frac{m_k+1}{2},\frac{\lambda^2}{2}\right)$	$N_{m_k}(A_k^{-1}X_k{}^Ty^*,\sigma^2A_k^{-1})$	$\tau_k^{-2}\|\beta,\sigma^2,X,y \sim \text{IG}\left(\sqrt{\frac{\lambda^2\sigma^2}{\|\beta_k\|^2}},\lambda^2\right)$

*For Bayesian Group Lasso the shrinkage prior distribution is $p(\beta|\sigma^2,\tau_1,..,\tau_K)$ with corresponding mixing density $p(\tau_1,..,\tau_K)$ and posterior distribution $p(\beta_k|\beta_{-k},\sigma^2,\tau_1,..,\tau_p)$, where $\beta_{-k}=(\beta_1,..,\beta_{k-1},\beta_{k+1},..,\beta_K)'$, $y^*=y-\frac{1}{2}\sum_{k\neq k'}X_{k'}\beta_{k'}$, $A_k=X_k{}^TX_k+1/\tau_k^2 I_{m_k}$, $k=1(1)K$. $A=(X^TX+D_\tau^{-1})$, $A^*=(X^TX+D^{*-1}_\tau)$, $A'=(X^TX+\frac{\sigma^2}{\tau^2}I_p)$, D'_τ is a diagonal matrix with diagonal elements $\tau^2=\sigma^2/\lambda$, D^*_τ is a diagonal matrix with diagonal elements $(\tau_i^{-2}+\lambda_2)^{-1}$, $i=1,..,p$ and D_τ is a diagonal matrix with diagonal elements τ_j^2. IG refers to inverse Gaussian distribution.

Table 2: Scale mixture representation of different shrinkage priors and corresponding posterior distributions.

Bayesian ridge: The hierarchical representation of Bayesian ridge estimator is obtained as follows:

$$\begin{aligned} &y^{n\times 1}\,|\,X,\beta,\sigma^2 \sim N_n(X\beta,\sigma^2 I_n),\\ &\beta^{p\times 1}\,|\,\tau^2 \sim \prod_{j=1}^{p} N(0,\tau^2),\\ &\tau^2 \sim Inv-\chi^2(v,s^2). \end{aligned} \tag{21}$$

Under this hierarchical prior, the posterior distribution of β is normal, while the reciprocal of the latent variable τ^2 is distributed as χ^2 distribution. The above Bayesian ridge regression can be extended to include variable-specific latent variables. In that case, the prior distributions become

$$\beta_j\,|\,\tau_j^{\,2} \sim N(0,\tau_j^{\,2}),\ \tau_j^{\,2} \sim Inv-\chi^2(v,s^2). \tag{22}$$

And the conditional posterior distribution of τ_j^2 becomes

$$\tau_j^{\,-2}\,|\,\beta,X,y \sim \frac{\chi^2_{v+p}}{vs^2+\beta_j^2}. \tag{23}$$

Bayesian group lasso: The hierarchical representation of Bayesian group lasso estimator is obtained as follows:

$$\begin{aligned} &y^{n\times 1}\,|\,X,\beta,\sigma^2 \sim N_n(X\beta,\sigma^2 I_n),\\ &\beta_k^{\,m_k\times 1}\,|\,\sigma^2,\tau_k^2 \sim N(0,\sigma^2\tau_k^2 I_{m_k}),\\ &k=1(1)K \text{ independently},\\ &\tau_1^2,..,\tau_K^2\,|\,\lambda \sim \prod_{k=1}^{K} Gamma\left(\frac{m_k+1}{2},\frac{\lambda^2}{2}\right),\\ &\sigma^2 \sim p(\sigma^2). \end{aligned} \tag{24}$$

Under this hierarchical prior, the posterior distribution of β_k is normal, while the reciprocals of the latent variables are distributed as inverse Gaussian distributions (Table 2). A non-informative scale-invariant marginal prior on σ^2 results in an inverse gamma posterior distribution of σ^2.

Bayesian elastic net: Similarly, the hierarchical representation of Bayesian elastic net estimator is obtained as follows:

$$\begin{aligned} &y^{n\times 1}\,|\,X,\beta,\sigma^2 \sim N_n(X\beta,\sigma^2 I_n),\\ &\beta^{p\times 1}\,|\,\sigma^2,\tau_1^2,..,\tau_p^2 \sim \prod_{j=1}^{p} N(0,\sigma^2(\tau_i^{-2}+\lambda_2)^{-1}),\\ &\tau_1^2,..,\tau_p^2\,\Big|\,\lambda_1,\lambda_2 \sim \prod_{j=1}^{p}\frac{\lambda_1^2}{2}e^{-\lambda_1^2\tau_j^2/2},\\ &\sigma^2 \sim p(\sigma^2). \end{aligned} \tag{25}$$

	Posterior distributions of the tuning parameters and their expectations	E-step	M-step
Bayesian lasso	$p(\lambda^2\,\vert.) \sim G(p+a,\frac{1}{2}\sum_{j=1}^{p}\tau_j^2+b)$ $E(\lambda^2\,\vert.) = \frac{p+a}{\frac{1}{2}\sum_{j=1}^{p}\tau_j^2+b}$	$E(\tau_j^{-2}\,\vert\beta,\sigma^2) = \frac{\lambda^2\sigma}{\vert\beta_j\vert}$	$\hat{\beta}=(X^{*T}A^{-1}X^*)^{-1}X^{*T}A^{-1}y^*$ $\hat{\sigma}^2=\frac{1}{n+p}(y^*-X^*\hat{\beta})^T A^{-1}(y^*-X^*\hat{\beta})$
Bayesian ridge	$p(s^2.) \sim G(v/2+a,\tau^{-2}v/2+b)$ $E(s^2.) = \frac{v/2+a}{\tau^{-2}v/2+b}$	$E(\tau^{-2}\,\vert\beta) = \frac{v+p}{vs^2+\sum_{j=1}^{p}\beta_j^2}$	$\hat{\beta}=(X^{*T}A'^{-1}X^*)^{-1}X^{*T}A'^{-1}y^*$ $\hat{\sigma}^2=\frac{1}{n+p}(y^*-X^*\hat{\beta})^T A'^{-1}(y^*-X^*\hat{\beta})$
Bayesian adaptive lasso	$p(\lambda_j^{\,2}\,\vert.) \sim G(1+a,\frac{1}{2}\tau_j^2+b)$ $E(\lambda_j^{\,2}\,\vert.) = \frac{1+a}{\frac{1}{2}\tau_j^2+b}$	$E(\tau_j^{\,-2}\,\vert\beta,\sigma^2) = \frac{\lambda_j^{\,2}\sigma}{\vert\beta_j\vert}$	$\hat{\beta}=(X^{*T}A^{-1}X^*)^{-1}X^{*T}A^{-1}y^*$ $\hat{\sigma}^2=\frac{1}{n+p}(y^*-X^*\hat{\beta})^T A^{-1}(y^*-X^*\hat{\beta})$
Bayesian elastic net	$p(\lambda_1^2.) \sim G(p+a_1,\frac{1}{2}\sum_{j-1}^{p}\tau_j^2+b_1)$ $p(\lambda_2^2.) \sim G(\frac{p}{2}+a_2,\frac{1}{2\sigma^2}\sum_{j-1}^{p}\beta_j^2+b_2)$ $E(\lambda_2^2\,\vert.) = \frac{\frac{p}{2}+a_2}{\frac{1}{2\sigma^2}\sum_{j=1}^{p}\beta_j^2+b_2}$ $E(\lambda_2^2\,\vert.) = \frac{\frac{p}{2}+a_2}{\frac{1}{2\sigma^2}\sum_{j=1}^{p}\beta_j^2+b_2}$	$E(\tau_j^{\,-2}\,\vert\beta,\sigma^2) = \sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}}$	$\hat{\beta}=(X^{*T}A^{*-1}X^*)^{-1}X^{*T}A^{*-1}y^*$ $\hat{\sigma}^2=\frac{1}{n+p}(y^*-X^*\hat{\beta})^T A^{*-1}(y^*-X^*\hat{\beta})$
Bayesian group lasso** (K Groups)	$p(\lambda^2\,\vert.) \sim G(p+\frac{K}{2}+a,\frac{1}{2}\sum_{k=1}^{K}\tau_k^2+b)$ $E(\lambda^2\,\vert.) = \frac{p+\frac{K}{2}+a}{\frac{1}{2}\sum_{k=1}^{K}\tau_k^2+b}$	$E(\tau_k^{\,-2}\,\vert\beta,\sigma^2) = \sqrt{\frac{\lambda^2\sigma^2}{\Vert\beta_k\Vert^2}}$	$\hat{\beta}=(X^{*T}\tilde{A}^{-1}X^*)^{-1}X^{*T}\tilde{A}^{-1}y^*$ $\hat{\sigma}^2=\frac{1}{n+p}(y^*-X^*\hat{\beta})^T \tilde{A}^{-1}(y^*-X^*\hat{\beta})$

* *where, $\tilde{A}=(X^TX+\tilde{D})$, $\tilde{D}=Diag(\tau_1^2 I_{m_1}..,\tau_K^2 I_{m_K})$.

Table 3: EM algorithms for various Bayesian regularization methods.

Under this hierarchical prior, the posterior distribution of β is normal, while the reciprocals of the latent variables are distributed as inverse Gaussian distributions (Table 2). Here also, a non-informative scale-invariant marginal prior on σ^2 results in an inverse gamma posterior distribution of σ^2. For a slightly different version of Bayesian elastic net estimator refer to the paper of Li and Lin [72]. Other related developments are Bayesian bridge [73], and a different version of Bayesian adaptive lasso [74].

Estimation of hyperparameters: In the Bayesian framework, typical approaches for estimation of the tuning parameters are based on incorporating them into the Gibbs sampler with an appropriate hyperprior [48]. Park and Casella [71] suggested using a gamma prior $G(a,b)$ for a proper posterior. The prior, which is put on λ^2 for convenience because of the way λ enters into the posterior [48], is given by

$$p(\lambda^2) = \frac{b^a}{\Gamma(a)}(\lambda^2)^{a-1}e^{-b\lambda^2},\ a,b>0. \tag{26}$$

For elastic net, we have two parameters, and we assign $G(a_1,b_1)$ and $G(a_2,b_2)$ for λ_1^2 and λ_2^2 respectively. When the prior (Equation 26) is used in the hierarchy, the full conditional distributions of all the tuning parameters are gamma distributions, and are listed in Table 3. For Bayesian ridge, a gamma prior $G(a,b)$ is put on s^2, which again results in a gamma posterior (Table 3). The tuning parameters can also be estimated through the marginal likelihood of λ, which can be implemented with an EM/Gibbs algorithm [48].

Algorithms for fitting Bayesian regularization methods: In this section, we briefly describe two popular model-fitting algorithms (*viz.* MCMC and EM), for estimating parameters in Bayesian regularization methods in linear regression. For the MCMC algorithms, we only describe the hierarchical formulations and the corresponding posterior distributions. Any feature of the posterior distribution is a legitimate candidate for Bayesian inference: moments, quantiles, highest posterior density regions, etc. [19]. Posterior median is one attractive measure as a robust estimator. Hans [75] emphasized using the posterior mean as point estimate, as it facilitates prediction of future observations *via* the posterior predictive distribution. However, implementation of MCMC algorithms for high dimensional problems may require excessive computing time. For practical and computational purposes, it is often desirable to have a fast algorithm that returns point estimates of the parameters and their standard errors. The EM algorithm [76] aims to address this issue by providing MAP estimates along with speedy inference and fast computation.

It must be emphasized that both non-Bayesian and Bayesian regularization methods are essentially optimization methods, with the common goal of determining the model parameters that maximize some objective function. A Bayesian approach can often lead to very different results than a traditional penalized regression approach [76]. Although, the Gibbs samplers discussed in this paper are extremely fast [48], one should be aware of the existence of problems relating to MCMC algorithms, e.g. slow convergence, poor mixing, etc. Therefore, once the simulation algorithm has been implemented and the simulations are drawn, it is absolutely necessary to check the convergence of the simulated sequences [6]. Inference based on the output of a slowly converging algorithm may require long runtime (many iterations e.g. thousands, millions, or more). If the convergence is not attained (painfully slow), one should resort to an alternative algorithm or other remedies, e.g. increasing burn-in period, thinning, etc. [6], for the inference to be valid.

MCMC algorithm

As shown above, the conditional posterior distribution for each parameter has standard form, and thus, can be directly sampled. Thus, the MCMC algorithm can be applied to fully explore the joint posterior distribution by sampling each parameter from its conditional posterior distribution. In summary, the MCMC algorithm proceeds as follows:

1. Initialize all the parameters with some plausible values.

2. Update β by sampling from its conditional posterior.

3. Update the latent variables and the variance parameter by sampling from their conditional posterior distributions.

4. If the hyperparameters are not prefixed, update them by sampling from their conditional posterior distributions.

EM algorithm

Given the latent variables τ_j^2's, the prior information of β can be incorporated in the linear model, as p 'additional data-points' with value 0 (prior means), and corresponding 'explanatory variables' equal to 0, except x_j which equals 1 and residual variance depending on τ_j^2 [6]. Therefore, given τ_j^2, the posterior mode of (β,σ^2) can be obtained by performing weighted linear regression on the augmented response variable y^*, augmented design matrix X* and augmented variance-covariance matrix Σ, where

$$y^{*n+p\times 1} = (y_1,\ldots,y_n,\underbrace{0,0..0}_{p})',\ X^{*n+p\times p} = \begin{pmatrix} X \\ I_p \end{pmatrix} \text{ and}$$

$$\sum^{n+p\times n+p} = \begin{pmatrix} \sigma^2 I_n & 0 \\ 0 & \sigma^2\Sigma_{\beta}{}^{p\times p} \end{pmatrix} \text{ and } \Sigma_\beta \text{ is a diagonal matrix}$$

containing prior variances. Therefore, by treating the latent variables as 'missing data' and averaging over them, we can estimate the posterior mode of (β,σ^2) by EM algorithm [76]. The algorithm proceeds as follows:

1. Initialize all the parameters with some plausible values.

2. E-Step: Update the latent variables by replacing their posterior conditional expectations.

3. M-Step: Update (β,σ^2) by weighted linear regression on the augmented data as follows:

$$\hat{\beta} = (X^{*T}\Sigma^{-1}X^*)^{-1}X^{*T}\Sigma^{-1}y^*,$$
$$\hat{\sigma}^2 = \frac{1}{n+p}(y^*-X^*\hat{\beta})^T\Sigma^{-1}(y^*-X^*\hat{\beta}). \tag{27}$$

4. Repeat 1,2, and 3, until convergence.

Table 3 gives the summary of E and M steps for different Bayesian regularization methods. In some cases, the hyperparameters need to be updated in the E-step, which can be easily done by plugging in their conditional posterior expectations. Other variants of EM algorithms are proposed by Figueiredo [77] and Xu [78].

Extension to GLM: A generalized linear model (GLM) consists of three components: the linear predictor, the link function and the distribution of the outcome variable [6,79]. The linear predictor can be expressed as: $\eta = X\beta$; the link function $g(.)$ relates the linear predictor η to the mean of the outcome variable y as: $E(y|X) = g^{-1}(X\beta)$ and the distribution of y depends on the linear predictor $X\beta$ and generally

also a dispersion (or variance) parameter ϕ and can be expressed as: $p(y|X\beta,\phi)=\prod_{i=1}^{n} p(y_i|X_i\beta,\phi)$, where $X_i\beta=\eta_i$ is the linear predictor for the i^{th} observation.

The standard IRLS algorithm proceeds by approximating the generalized linear model by a weighted normal linear regression, and estimating the maximum likelihood estimates of the parameters (β,ϕ) using weighted least squares, and then iterating the process [6]. At each iteration, the algorithm calculates pseudo-data z_i and pseudo-variances σ_i^2 for each observation based on the current estimates of the parameters $(\hat{\beta},\hat{\phi})$, approximating the generalized linear model likelihood $p(y_i|X_i\beta,\phi)$ by the normal likelihood $N(z_i|X_i\beta,\phi\sigma_i^2)$ and then updating the parameters by weighted linear regression. The iteration proceeds until convergence. The pseudo-data z_i and pseudo-variances σ_i^2 are calculated by

$$z_i = X_i\hat{\beta} - \frac{L'(y_i|X_i\hat{\beta},\hat{\phi})}{L''(y_i|X_i\hat{\beta},\hat{\phi})},$$
$$\hat{\sigma}_i^2 = 1/L''(y_i|X_i\hat{\beta},\hat{\phi}), \quad (28)$$

and the estimates are obtained as

$$\hat{\beta}_{WLS} = (X^T W X)^{-1} X^T W z,$$
$$\hat{\phi} = \frac{1}{n-p}(z-X\hat{\beta})^T W (z-X\hat{\beta}), \quad (29)$$

where $W^{-1}=Diag(\sigma_1^2,..,\sigma_n^2)$, L is the log-likelihood, L' and L'' are first and second order derivatives of L with respect to η_i and $(\hat{\beta},\hat{\phi})$ are the current estimates of (β,ϕ). Thus, all the methods described above can be easily extended to generalized linear models by approximating the GLM likelihood by normal likelihood [80,81].

Related Bayesian shrinkage methods: Bayesian regularization methods are not merely the Bayesian versions of different regularization methods. In fact, Bayesian regularization methods fall into the broad class of Bayesian shrinkage methods. Apart from the methods discussed above, there are tons of other Bayesian shrinkage methods available in literature, which include the normal/exponential-gamma model of Griffin and Brown [82]; the normal/gamma and the normal/inverse-Gamma model [70,83]; the horseshoe prior of Carvalho et al. [84]; the generalized double-Pareto model of Armagan et al. [85]; the orthant normal prior of Hans [86]; mixture of uniform prior of Knürr at al. [87]; the Laplace-Poisson model of Chen et al. [88], and the hypergeometric inverted-beta model of Polson and Scott [89]. For the literature of EM algorithms in Bayesian shrinkage methods, refer Gelman et al. [6], Green [90], Polson and Scott [89].

Bayesian model selection using nonlocal priors

Regularization methods identify only one model that maximizes a penalized likelihood function, or minimizes RSS subject to a penalty. On the other hand, most conventional Bayesian methods based on local prior densities [61-64] provide estimates of posterior model probabilities that are poor and unrealistic, and therefore, the posterior model probability estimates are often unreported for high dimensional linear models [67]. To overcome this deficiency, Johnson and Rossell [67] recently proposed Bayesian model selection methods by specifying nonlocal prior densities on the regression coefficients, which provide accurate estimates of the posterior probability that each identified model is correct. Unlike local prior densities, which are positive at null parameter value, nonlocal prior densities are identically zero whenever a model parameter is equal to its null value [91]. Johnson and Rossell [67] introduced two classes of nonlocal priors on the coefficients, along with their frequentist analogues for both densities, which we describe below.

Product moment (pMOM) densities: The first class of nonlocal densities is called product moment (pMOM) densities, which are defined as

$$p(\beta|\tau,\sigma^2,r) = d_p(2\pi)^{-p/2}(\tau\sigma^2)^{-rp-p/2}|A_p|^{1/2} \times \exp\left[-\frac{1}{2\tau\sigma^2}\beta' A_p \beta\right]\prod_{j=1}^{p}\beta_j^{2r}, \quad (30)$$

for $\tau>0$, A_p a $p\times p$ nonsingular matrix, and $r=1,2,...$ is called the order of the density and d_p is the normalizing constant independent of σ^2 and τ. Similar to BIC, an objective function that might be associated with this prior can be expressed as

$$Q_{pMOM}(\beta) = -2\log L + cp\log n - d\sum_j \log\left[\left(\frac{\beta_j^2}{\tau\sigma^2}\right)^r\right], \quad (31)$$

where L is the likelihood function and $d>0$. Similar to BIC, model with minimum value of the objective function (equation 31) should be selected as the best model.

Product inverse moment (piMOM) densities: The second class of nonlocal densities is called product inverse moment (piMOM) densities, which are defined as

$$p(\beta|\tau,\sigma^2,r) = \frac{(\tau\sigma^2)^{rp/2}}{\Gamma(r/2)^p}\prod_{j=1}^{p}|\beta_j|^{-(r+1)} \times \exp\left(-\frac{\tau\sigma^2}{\beta_j^2}\right), \quad (32)$$

for $\tau>0$ and r=1,2,.... The parameter τ in both pMOM and piMOM prior densities represents a scale parameter that determines the dispersion of the prior densities on β around the null vector, and it should be estimated carefully for efficient computation [67]. The objective function associated with this prior can be expressed as

$$Q_{piMOM}(\beta) = -2\log L + cp\log n + d\sum_j\left(\frac{\tau\sigma^2}{\beta_j^2}\right)^r, \quad (33)$$

where L is the likelihood function and $d>0$. Similar to above, model with minimum value of the objective function (Equation 33) should be selected as the best model

Based on these prior densities, Johnson and Rossell [67] formulated marginal densities in analytical form, leading to analytical posterior probabilities. Using these expressions, they proposed an MCMC scheme to obtain posterior samples. They demonstrated that the proposed method is model selection consistent (posterior probability of selecting the 'true' model approaches 1), as $n\to\infty$ and as long as

$n \le p$ under certain regularity conditions on the design matrix X. Also, the proposed method performed impressively, when compared with Bayesian methods based on local prior specifications [61-64]. Moreover, the proposed method performs either as well, or better than various regularization methods, as evident from their simulation studies. Although they provided frequentist versions of the nonlocal prior specifications, Johnson and Rossell [67] recommended using Bayesian methods, as it facilitates inference regarding the posterior probability that each model is true along with easy computation.

Bayesian Subset Regression (BSR)

Another recently developed method, which is particularly interesting to us and within the scope of current review, is the Bayesian Subset Regression (BSR) method proposed by Liang et al. [92], for high dimensional generalized linear models. They propose a new prior specification, which results in a Bayesian subset regression (BSR), with the negative log-posterior distribution approximately reduced to EBIC, when the sample size is large. In addition, they propose a variable screening procedure based on marginal inclusion probability, which is shown to have same theoretical properties of sure screening and is consistent, as the SIS procedure by Fan and Song [93], although both SIS [93,94], and its iterative extension ISIS [95] are outperformed by the proposed method, in terms of prediction accuracy. Also, the proposed method outperforms several popular regularization methods, including lasso and elastic net, suggesting that BSR is more suitable in high dimensional models than regularization methods. Here, we describe the method for high dimensional linear models.

Prior specification: To formulate the prior specification of their method, let us denote the number of explanatory variables as P_n. Following Liang et al. [92], let us denote by ξ_n a subset model (of size $\le P_n$) of the full model with P_n predictors. Let, $\beta_{\xi_n} = \left(\beta^*_1, \beta^*_2, .., \beta^*_{|\xi_n|}\right)$ denote the vector of true regression coefficients of the model ξ_n. With this formulation, Liang et al. [92] set up following priors on β_{ξ_n} and ξ_n

$$p(\beta_{\xi_n}) = \frac{1}{(2\pi\sigma_{\xi_n}{}^2)^{|\xi_n|/2}} \exp\left\{-\frac{1}{2\sigma_{\xi_n}{}^2}\sum_{j=1}^{n}\beta_j^{*2}\right\}, \tag{34}$$

$$p(\xi_n) = \nu_n^{|\xi_n|}(1-\nu_n)^{P_n-|\xi_n|} I\left[|\xi_n| < K_n\right], \tag{35}$$

where $\sigma_{\xi_n}{}^2$ is a pre-specified variance chosen, such that

$$\log p(\beta_{\xi_n}) = O(1), \tag{36}$$

which ensures that the prior information of β_{ξ_n} can be ignored for sufficiently large n; ν_n denotes the prior probability of each variable, independent of other variables to be selected for the subset model, and K_n is an upper bound on the model size $|\xi_n|$ facilitating calculation of the MLE $\hat{\beta}_{\xi_n}$.

Posterior distribution: With the above prior specifications and following prior probability specification

$$\nu_n = \frac{1}{1+P_n^{\gamma}\sqrt{2\pi}}, \tag{37}$$

the log-posterior distribution is approximated as the following whenever $|\xi_n| < K_n$

$$\log p(\xi_n|.) \approx C + \log f(y|\hat{\beta}_{\xi_n}, \xi_n, X) - \frac{|\xi_n|}{2}\log n - |\xi_n|\gamma \log P_n \tag{38}$$

and equals $-\infty$ otherwise. Thus, the negative of the log-likelihood approximately reduces to the EBIC. This leads to Bayesian subset regression, with the MAP model approximately equivalent to the minimum EBIC model. For the simulation of the posterior, they propose an adaptive MCMC algorithm. With extensive numerical studies, they establish that BSR outperforms penalized likelihood approaches significantly, especially when the dimension of P_n is increasing. Under mild conditions, they establish the consistency of the resulting posterior. In addition, they show that the posterior probability of the true model will converge to 1 as $n \to \infty$.

Asymptotic Behaviors of High Dimensional Bayesian Methods

In this section, we consider the asymptotic behaviors of Bayesian variable selection methods in high dimensional linear models. Here also we denote the number of explanatory variables as P_n which is possibly much larger than the sample size n and it is assumed that the regression coefficients satisfy the sparseness condition $\lim_{n\to\infty}\sum_{j=1}^{P_n}|\beta_j^*| < \infty$, where β_j^* is the j^{th} 'true' regression coefficient. The sparseness condition describes a general situation when all explanatory variables are relevant, but most of them have very small effects [96]. Several authors have studied the asymptotic properties of Bayesian variable selection methods under different regularity conditions [68,96-99]. Wang et al. [96], Jiang [97] and Jiang [98] used a framework developed by Wasserman [100], for showing consistency in the context of density estimation. According to Wasserman [100], 'density consistency' is defined as 'given a proper prior to propose joint densities $f(x, y)$ of response y and explanatory variables vector x, the posterior-proposed densities are often close to the true density for large n'. Given a proper prior to propose joint densities of response y and explanatory variables vector x, the above three works showed that the posterior-proposed joint densities are often consistent to the true joint density for large n. Jiang [98] defined a proper prior for different sets of explanatory variables to prove that the posterior-proposed densities with different parameterizations were consistent to the true density. Wang et al. [96] further proved the 'regression function consistency', which ensures good performance of high dimensional Bayesian variable selection methods, especially when some regression coefficients are bounded away from zero, while the rest are exactly zero. Jiang [97] also studied the convergence rates of the fitted densities for generalized linear models and established consistency under some realistic assumptions. In summary, the asymptotic results reveal that with appropriate prior specification, high dimensional Bayesian variable selection methods not only can identify the 'true' model with probability 1, but also can give consistent parameter estimates [96], facilitating tremendous applications in a variety of fields. For other asymptotic results refer Casella et al. [19] or Moreno et al. [101].

On the other hand, unlike frequentist methods, asymptotic behaviors of Bayesian regularization or shrinkage methods are less studied and poorly understood [68]. Armagan et al. [99] provided sufficient conditions on prior concentration for strong posterior consistency, when $P_n = o(n)$ as $n \to \infty$ for various Bayesian shrinkage methods, including Bayesian lasso [71], generalized double pareto [85], and horseshoe estimator [84]. Posterior consistency involves

examining a posterior probability of a set of parameter values as $n \to \infty$, where the set is any neighborhood of the true parameter value [102]. Posterior consistency is mainly verified after checking sufficient conditions for a general posterior consistency theorem, after defining a suitable topology of the parameter, and the neighborhood of the true value of the parameter [102]. Bhattacharya et al. [92] studied prior concentrations of various Bayesian shrinkage priors to investigate whether the entire posterior distributions of these methods concentrate at the optimal rate, i.e. the posterior probability assigned to a shrinking neighborhood (proportional to the optimal rate) of the true value of the parameter converges to 1. They argued that most of the Bayesian shrinkage methods are sub-optimal, in terms of posterior contraction, which is considered stronger optimality condition than posterior consistency alone. Due to lack of theoretical justification, Bayesian shrinkage methods are not easily adopted by subjective Bayesians and frequentists [68]. Therefore, much work in this direction needs to be done to better understand these approaches regarding their asymptotic behaviors in high dimensional linear models.

Discussion

Over the years, Bayesian methods have evolved immensely with the growth in modern computing power. These approaches based on modern statistical techniques are powerful tools to handle the associated challenges in high dimensional data analysis. This paper attempts to provide a selective overview of these methods for high dimensional linear models. Our discussion is only limited to linear models; we have not discussed nonparametric Bayesian methods [103-107], and various machine learning techniques [66,95,108], and several other approaches which have their own advantages, and are beyond the scope of current review. We particularly focus on the Bayesian regularization methods, which have enjoyed a great deal of applicability in recent years. We have summarized the model-fitting algorithms for fitting these methods for high dimensional linear models. For more comprehensive review of Bayesian formulations of various regularization methods as hierarchical models, we suggest the readers to read Kyung et al. [48]. Also, we have mentioned two new Bayesian model selection methods, which have shown excellent performance in high dimensional model selection. Moreover, we have addressed the asymptotic behaviors of Bayesian high dimensional methods for linear models under certain regularity conditions. The paper will be particularly helpful in understanding the basic methodology used in high dimensional Bayesian methods. It is to be noted that Bayesian variable or model selection is a much broader topic than what we have described here, which includes several data pre-processing steps in practice including variable transformations, coding of variables, removal of outliers, etc. Therefore, in real life applications, the general framework of Bayesian variable and model selection [109], should be applied to these methods to ensure accurate results and easy implementation.

Acknowledgements

This work was supported in part by the research grant NIH 5R01GM069430-08. Himel Mallick was supported in part by a cooperative agreement grant U01 NS041588 from the National Institute of Neurological Disorders and Stroke, National Institutes of Health and Department of Health and Human Services. We thank the editor for several thoughtful suggestions on our earlier version of the manuscript, which greatly improved the quality of the paper.

References

1. Mallows CL (1973) Some comments on Cp. Technometrics 15: 661-675.
2. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19: 716-723.
3. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6: 461-464.
4. Spiegelhalter DJ, Best NG, Carlin BP, Der Linde AV (2002) Bayesian measures of model complexity and fit. J R Stat Soc Series B Stat Methodol 64: 583-639.
5. Yan X, Su XG (2009) Linear regression analysis: theory and computing. World Scientific Publishing Company, Singapore.
6. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis. (2nd Edn), Chapman & Hall/CRC, Florida, USA.
7. Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York, USA.
8. Woodroofe M (1982) On model selection and the arc sine laws. Ann Stat 10: 1182-1194.
9. Nishii R (1984) Asymptotic properties of criteria for selection of variables in multiple regression. Ann Stat 12: 758-765.
10. Burnham KP, Anderson DR (2004) Multimodel inference understanding AIC and BIC in model selection. Sociol Methods Res 33: 261-304.
11. Dziak J, Li R, Collins L (2005) Critical review and comparison of variable selection procedures for linear regression (Technical report).
12. Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76: 297-307.
13. Kundu D, Murali G (1996) Model selection in linear regression. Comput Stat Data Anal 22: 461-469.
14. Yang Y (2003) Regression with multiple candidate models: selecting or mixing? Stat Sin 13: 783-810.
15. Ando T (2007) Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. Biometrika 94: 443-458.
16. Gelman A, Hwang J, Vehtari A (2013) Understanding predictive information criteria for Bayesian models.
17. Luo S, Chen Z (2012) Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. J Stat Plan Inference 143: 494-504.
18. Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J R Stat Soc Series B Stat Methodol 64: 641-656.
19. Casella G, Girón FJ, Martínez ML, Moreno E (2009) Consistency of Bayesian procedures for variable selection. Ann Stat 37: 1207-1228.
20. Yang Y, Barron AR (1998) An asymptotic property of model selection criteria. IEEE Trans Inf Theory 44: 95-116.
21. Wang H, Li B, Leng C (2009) Shrinkage tuning parameter selection with a diverging number of parameters. J R Stat Soc Series B Stat Methodol 71: 671-683.
22. Chen J, Chen Z (2012) Extended BIC for small-n-large-P sparse GLM. Stat Sin 22: 555-574.
23. Chen J, Chen Z (2008) Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95: 759-771.
24. Shao J (1997) An asymptotic theory for linear model selection. Stat Sin 7: 221-264.
25. Kim Y, Kwon S, Choi H (2012) Consistent model selection criteria on high dimensions. J Mach Learn Res 13: 1037-1057.
26. Zhao P, Yu B (2006) On model selection consistency of Lasso. J Mach Learn Res 7: 2541-2563.
27. Kim Y, Choi H, Oh HS (2008) Smoothly clipped absolute deviation on high dimensions. J Am Stat Assoc 103: 1665-1673.
28. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96: 1348-1360.
29. Bickel PJ, Li B (2006) Regularization in statistics. Test 15: 271-344.
30. Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12: 55-67.
31. Frank IE, Friedman JH (1993) A statistical view of some chemometrics regression tools. Technometrics 35: 109-135.

32. Park C, Yoon YJ (2011) Bridge regression: adaptivity and group selection. J Stat Plan Inference 141: 3506-3519.

33. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Statist Soc B 58: 267-288.

34. Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, et al. (2010) Insights into colon cancer etiology *via* a regularized approach to gene set analysis of GWAS data. Am J Hum Genet 86: 860-871.

35. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 67: 301-320.

36. Radchenko P, James GM (2008) Variable inclusion and shrinkage algorithms. J Am Stat Assoc 103: 1304-1315.

37. Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101: 1418-1429.

38. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Series B Stat Methodol 68: 49-67.

39. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. J R Stat Soc Series B Stat Methodol 67: 91-108.

40. Ghosh S (2011) On the grouped selection and model complexity of the adaptive elastic net. Stat Comput 21: 451-462.

41. Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38: 894-942.

42. Candes E, Tao T (2007) The Dantzig selector: Statistical estimation when p is much larger than n. Ann Stat 35: 2313-2351.

43. Huang J, Ma S, Xie H, Zhang CH (2009) A group bridge approach for variable selection. Biometrika 96: 339-355.

44. Hesterberg T, Choi NH, Meier L, Fraley C (2008) Least angle and ℓ1 penalized regression: A review. Statistics Surveys 2: 61-93.

45. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32: 407-499.

46. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models *via* coordinate descent. J Stat Softw 33: 1-22.

47. Wu TT, Lange K (2008) Coordinate descent algorithms for lasso penalized regression. Ann Appl Stat 2: 224-244.

48. Casella G, Ghosh M, Gill J, Kyung M (2010) Penalized regression, standard errors, and Bayesian lassos. Bayesian Anal 5: 369-411.

49. Craven P, Wahba G (1979) Smoothing noisy data with spline functions. Numer Mathe 31: 377-403.

50. O'Hara RB, Sillanpää MJ (2009) A review of Bayesian variable selection methods: what, how and which. Bayesian Anal 4: 85-118.

51. George EI, McCulloch RE (1997) Approaches for Bayesian variable selection. Stat Sin 7: 339-373.

52. George EI, McCulloch RE (1993) Variable selection *via* Gibbs sampling. J Am Stat Assoc 88: 881-889.

53. Dellaportas P, Forster JJ, Ntzoufras I (2002) On Bayesian model and variable selection using MCMC. Stat Comput 12: 27-36.

54. Wasserman L (2000) Bayesian model selection and model averaging. J Math Psychol 44: 92-107.

55. Raftery AE (1995) Bayesian model selection in social research. Sociological Methodology 25: 111-164.

56. Zhang Z, Zhao C (2009) On Bayesian Variable Selection: Methods and Implementations for Genetic Association Studies.

57. Kuo L, Mallick B (1998) Variable selection for regression models. Sankhya Ser B 60: 65-81.

58. Carlin BP, Chib S (1995) Bayesian model choice via Markov chain Monte Carlo methods. J R Statist Soc B 57: 473-484.

59. Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82: 711-732.

60. Zellner A (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti 6: 233-243.

61. Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008) Mixtures of g priors for Bayesian variable selection. J Am Stat Assoc 103: 410-423.

62. Bottolo L, Richardson S (2010) Evolutionary stochastic search for Bayesian model exploration. Bayesian Anal 5: 583-618.

63. Berger JO, Pericchi LR (1996) The intrinsic Bayes factor for model selection and prediction. J Am Stat Assoc 91: 109-122.

64. O'Hagan A (1995) Fractional Bayes factors for model comparison. J R Statist Soc B 57: 99-138.

65. Casella G, Moreno E (2006) Objective Bayesian variable selection. J Am Stat Assoc 101: 157-167.

66. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3: 1157-1182.

67. Johnson VE, Rossell D (2012) Bayesian Model selection in high-dimensional settings. J Am Stat Assoc 107: 649-660.

68. Liang F, Song Q, Yu K (2013) Bayesian subset modeling for high dimensional generalized linear models. J Am Stat Assoc.

69. Andrews DF, Mallows CL (1974) Scale mixtures of normal distributions. J R Statist Soc B 36: 99-102.

70. Yi N, Xu S (2008) Bayesian LASSO for quantitative trait loci mapping. Genetics 179: 1045-1055.

71. Park T, Casella G (2008) The bayesian lasso. J Am Stat Assoc 103: 681-686.

72. Li Q, Lin N (2010) The Bayesian elastic net. Bayesian Anal 5: 151-170.

73. Polson NG, Scott JG, Windle J (2011) The Bayesian Bridge. arXiv preprint arXiv:1109.2279.

74. Leng C, Tran MN, Nott D (2010) Bayesian Adaptive Lasso. arXiv preprint arXiv:1009.2300.

75. Hans C (2010) Model uncertainty and variable selection in Bayesian lasso regression. Stat Comput 20: 221-229.

76. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data *via* the EM algorithm. J R Statist Soc B 39: 1-38.

77. Figueiredo MA (2003) Adaptive sparseness for supervised learning. IEEE Trans Pattern Anal Mach Intell 25: 1150-1159.

78. Xu S (2010) An expectation–maximization algorithm for the Lasso estimation of quantitative trait locus effects. Heredity 105: 483-494.

79. McCullagh P, Nelder JA (1989) Generalized linear model. (2nd Edn), Chapman & Hall/CRC, New York, USA.

80. Yi N, Ma S (2012) Hierarchical shrinkage priors and model fitting for high-dimensional generalized linear models. Stat Appl Genet Mol Biol 11.

81. Yi N, Banerjee S (2009) Hierarchical generalized linear models for multiple quantitative trait locus mapping. Genetics 181: 1101-1113.

82. Griffin JE, Brown PJ (2005) Alternative prior distributions for variable selection with very many more variables than observations.

83. Griffin JE, Brown PJ (2010) Inference with normal-gamma prior distributions in regression problems. Bayesian Anal 5: 171-188.

84. Carvalho CM, Polson NG, Scott JG (2010) The horseshoe estimator for sparse signals. Biometrika 97: 465-480.

85. Armagan A, Dunson DB, Lee J (2013) Generalized double pareto shrinkage. Stat Sin 23: 119-143.

86. Hans C (2011) Elastic net regression modeling with the orthant normal prior. J Am Stat Assoc 106: 1383-1393.

87. Knürr T, Läärä E, Sillanpää MJ (2011) Genetic analysis of complex traits via Bayesian variable selection: the utility of a mixture of uniform priors. Genetics Research 93: 303-318.

88. Chen X, Wang ZJ, McKeown MJ (2011) A Bayesian Lasso *via* reversible-jump MCMC. Signal Processing 91: 1920-1932.

89. Polson NG, Scott JG (2012) Good, great, or lucky? Screening for firms with

sustained superior performance using heavy-tailed priors. Ann Appl Stat 6: 161-185.

90. Green PJ (1990) On use of the EM for penalized likelihood estimation. J R Statist Soc B 52: 443-452.
91. Johnson VE, Rossell D (2010) On the use of non local prior densities in Bayesian hypothesis tests. J R Stat Soc Series B Stat Methodol 72: 143-170.
92. Bhattacharya A, Pati D, Pillai NS, Dunson DB (2012) Bayesian shrinkage. arXiv preprint arXiv:1212.6088.
93. Fan J, Song R (2010) Sure independence screening in generalized linear models with NP-dimensionality. Ann Stat 38: 3567-3604.
94. Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Series B Stat Methodol 70: 849-911.
95. Fan J, Lv J (2010) A selective overview of variable selection in high dimensional feature space. Stat Sin 20: 101-148.
96. Wang S, Luan Y, Chang Q (2011) On model selection consistency of bayesian method for normal linear models. Commun Stat Theory Methods 40: 4021-4040.
97. Jiang W (2007) Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. Ann Stat 35: 1487-1511.
98. Jiang W (2006) On the consistency of Bayesian variable selection for high dimensional binary regression and classification. Neural Comput 18: 2762-2776.
99. Armagan A, Dunson DB, Lee J, Bajwa WU, Strawn N (2011) Posterior consistency in linear models under shrinkage priors. arXiv preprint arXiv:1104.4135.
100. Wasserman L (1998) Asymptotic properties of nonparametric Bayesian procedures. In: Practical nonparametric and semiparametric Bayesian statistics. Springer, New York, USA 293-304.
101. Moreno E, Girón FJ, Casella G (2010) Consistency of objective Bayes factors as the model dimension grows. Ann Stat 38: 1937-1952.
102. Shi JQ, Choi T (2011) Gaussian process regression analysis for functional data. Chapman & Hall, Boca Raton, USA.
103. Zou F, Huang H, Lee S, Hoeschele I (2010) Nonparametric Bayesian variable selection with applications to multiple quantitative trait loci mapping with epistasis and gene–environment interaction. Genetics 186: 385-394.
104. Dunson DB (2010) Nonparametric Bayes applications to biostatistics. Bayesian Nonparametrics 28: 223.
105. Dunson DB (2009) Bayesian nonparametric hierarchical modeling. Biom J 51: 273-284.
106. Lijoi A, Prünster I (2010) Models beyond the Dirichlet process. Bayesian Nonparametrics 28: 80.
107. Teh YW, Jordan MI (2010) Hierarchical Bayesian nonparametric models with applications. Bayesian Nonparametrics 158-207.
108. Tipping ME (2001) Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 1: 211-244.
109. Chipman H, George EI, Mcculloch RE (2001) The practical implementation of Bayesian model selection. Institute of Mathematical Statistics, USA.

This article was originally published in a special issue, **Advances in Markov Chain Monte Carlo Methods and Survival Analysis** handled by Editor(s). Dr. Faming Liang, Texas A&M University, USA; Dr. Nengjun Yi, University of Alabama at Birmingham, USA; Dr. Wenqing He, University of Western Ontario, Canada; Dr. Liuquan Sun, Institute of Applied Mathematics, Academy of Mathematics and Systems Science, China