Chapter 4

Bayesian model selection for high-dimensional data

Naveen Naidu Narisetty*
Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, United States
*Corresponding author: e-mail: naveen@illinois.edu

Abstract
High-dimensional data, where the number of features or covariates can even be larger than the number of independent samples, are ubiquitous and are encountered on a regular basis by statistical scientists both in academia and in industry. A majority of the classical research in statistics dealt with the settings where there is a small number of covariates. Due to the modern advancements in data storage and computational power, the high-dimensional data revolution has significantly occupied mainstream statistical research. In gene expression datasets, for instance, it is not uncommon to encounter datasets with observations on at most a few hundred independent samples (subjects) and with information on tens or hundreds of thousands of genes per each sample. An important and common question that arises quickly is—"which of the available covariates are relevant to the outcome of interest?" This concerns the problem of variable selection (and more generally model selection) in statistics and data science.

This chapter will provide an overview of some of the most well-known model selection methods along with some of the more recent methods. While frequentist methods will be discussed, Bayesian approaches will be given a more elaborate treatment. The frequentist framework for model selection is primarily based on penalization, whereas the Bayesian framework relies on prior distributions for inducing shrinkage and sparsity. The chapter treats the Bayesian framework in the light of objective and empirical Bayesian viewpoints as the priors in the high-dimensional setting are typically not completely based subjective prior beliefs. An important practical aspect of high-dimensional model selection methods is computational scalability which will also be discussed.

Keywords: Bayesian variable selection, High-dimensional data, Model comparison, Bayesian computation

Handbook of Statistics, Vol. 43. https://doi.org/10.1016/bs.host.2019.08.001

1 Introduction

The rapid developments in collecting, storing, transmitting, and managing massive amounts of data have led to unique opportunities and challenges in Statistics and the emerging field of Data Science. Variable selection is a fundamentally important problem for many modern datasets that have a large as the number of variables, which is a common feature of modern data sets from many applications including biology, climate sciences, behavioral and environmental sciences.

The linear regression model is the one of the most commonly used models in statistics and is also a building block for many general models. In this chapter, we primarily consider the linear regression model, but most of the ideas and methods discussed can be applied more generally to other models such as the generalized linear models and nonlinear regression models. Consider the linear regression model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \tag{1}$$

with standard assumptions on the error vector ϵ. The classical least squares approach for estimating β minimizes the loss function

$$L(\beta) = \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2, \tag{2}$$

where y_i denotes the ith response and x_i denotes the covariate vector for the ith observation. The well-known least squares estimator for β is

$$\hat{\beta}^{\text{ols}} := (X^\top X)^{-1} X^\top Y. \tag{3}$$

In the high-dimensional setting, the dimension p of the covariate vector can be quite large and potentially even larger than the sample size. For example, in gene expression datasets, there are at least thousands of genes as covariates but typically only a few dozen independent samples. In such cases, p is much larger than n and will be denoted by $p \gg n$. When $p \gg n$, even estimation of the regression parameter β is a challenging problem since the least squares minimization (2) does not have a unique solution and the least squares estimator is not well defined. This necessitates some simplifying assumptions on the data generating model and the most common assumption made is that most of the components of β are zero, often referred to as the "sparsity assumption" (van de Geer, 2016). Sparsity assumption is made in a lot of applications and is often reasonable. However, one cannot hope that the sparsity assumption remains valid for every application and hence relaxations of the sparsity assumption are also considered in the literature (Belloni and Chernozhukov, 2013; Belloni et al., 2011).

Even under the sparsity assumption, it is a very challenging problem to uncover the precise sparsity structure. The problem of detecting the nonzero

components of the parameter vector is often referred to as variable selection or model selection. In many scientific applications, a sparser model is generally desired for the reasons of parsimony, reduced variance, and ease of interpretability. This article attempts to provide a review of variable selection methods for high-dimensional datasets with a focus on the Bayesian approaches for variable selection. Fan and Lv (2010) and Bühlmann and van de Geer (2011) provide reviews of frequentist approaches for variable selection. While George and McCulloch (1997) and O'hara and Sillanpaa (2009) provide reviews of some Bayesian variable selection methods, the current article covers more recently developed strategies. There has been a tremendous research activity in this direction and it is not possible to cover all the work, so the article only presents a selection of ideas.

The remaining part of the chapter is organized as follows. In Section 2, a brief outline of some of the classical approaches for model selection, which mainly target the low-dimensional setting, are described. In Section 3, the penalization framework for variable selection is discussed. In Section 4, the Bayesian framework is introduced followed by a detailed description on spike and slab priors in Section 5 and on continuous shrinkage priors in Section 6. Section 7 discusses some computational aspects of Bayesian model selection and Section 8 gives an outline of the different types of theoretical results studied in the literature. Section 9 provides R packages available for implementation and Section 10 provides an example data analysis.

2 Classical variable selection methods

2.1 Best subset selection

For model selection, the best subset selection approach is to find the best combination of variables among all possible combinations. For example, if there are three predictors X_1, X_2, X_3, then we would consider all the possible models

$$\{X_1\}, \{X_2\}, \{X_3\}, \{X_1, X_2\}, \{X_1, X_3\}, \{X_2, X_3\}, \{X_1, X_2, X_3\}$$

and determine which of the above models is the best based on some criterion function. The criterion function used needs to penalize large models since large models tend to have better in-sample fit. A major difficulty is when the number of predictors p is large, in which case there are too many models to consider and it soon becomes infeasible to evaluate all the possible models.

More recently, modern optimization algorithms have been proposed to perform best subset selection in high-dimensional contexts. Bertsimas et al. (2016) developed modern mixed integer optimization methods to obtain optimal solutions to the best subset selection problem. These algorithms can be used in problems up to a thousand predictors. Recently, Hazimeh and Mazumder (2018) have developed coordinate descent algorithms for the best subset selection problem and demonstrated that their algorithms can be applied to simulated datasets with p nearly a million.

2.2 Stepwise selection methods

Motivated by the computational burden associated with traditional best subset selection algorithms, stepwise methods are developed for finding a small subset of "good models" to consider for further evaluation.

- Forward selection (FS): Starting from the null model which has no covariates, at each step of the FS algorithm, a new variable is added to the current model based on some criterion such as the decrease in residual sum of squares (RSS). This provides a sequence of p models and the model minimizing a criterion, such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) which are defined in Section 2.3, is used for selecting a final model.
- Backward elimination (BE): Very similar in spirit to the FS algorithm but the difference is that the BE algorithm starts from the full model (when it is possible to estimate the full model), and removes one variable at a time based on the increase in RSS.

For both FS and BE, it is not necessary to obtain the whole sequence of p models. One can instead terminate the algorithm early after a certain number of steps. In such a case, compared to BE, FS has the computational advantage since only relatively smaller models need to be fit especially when the number of steps is small. This is because working with the full model or models close to the full model in size is typically computationally more expensive. On the other hand, if a pair of important variables are not significant marginally but are jointly significant, then forward selection tends to miss both variables whereas backward elimination has higher chance of selecting them. It is not necessary that the models selected by BE and FS coincide. One simple strategy is to use a criterion function and select one final model from the two models selected by BE and FS.

A more general algorithm compared to FS and BE is to consider the possibilities of both addition and deletion of a variable at each step. A further generalization is to include swapping of a selected variable with a variable not yet selected in the model. While these strategies provide more general sequence of models and are likely to provide better model selection performance, they demand more computational power compared to forward selection.

The stepwise approaches described so far provide a sequence of models instead of specifying a stopping rule. Stopping criteria based on p-values or F-test statistics have been commonly considered in the literature (Bendel and Afifi, 1977; Finos et al., 2010; Grechanovsky and Pinsker, 1995). However, an important issue that needs to be carefully addressed in those cases is the control of false positives that could occur due to the large number of comparisons involved. A better alternative is to use a criterion function on the entire collection of models, which can incorporate the multiplicity of the comparisons. We now discuss some commonly used criterion functions.

2.3 Criterion functions

- Akaike information criterion (AIC): AIC (Akaike, 1973) for the model M_k with dimension k is defined as

$$\mathrm{AIC}(M_k) = -2 \log L(M_k) + 2k,$$

where $L(M_k)$ is the likelihood corresponding to the model M_k. The first term $-2 \log L(M_k)$ in AIC is twice the negative log likelihood, which turns out to be the residual sum of squares corresponding to the model M_k for the linear regression model with a Gaussian likelihood. That is, $-2 \log L(M_k) = \sum_{i=1}^{n} (y_i - x_i^\top \hat{\beta}(M_k))^2$, where $\hat{\beta}(M_k)$ is the least squares estimator for model M_k. Therefore, the first term acts as a measure of lack of fit to the data with smaller values to be preferred. The second term acts as a penalty term to penalize models having a large dimension. AIC aims to balance the lack of fit and the model complexity with models having smaller AIC values indicating a better balance between these two important aspects.

- Bayesian information criterion (BIC):
BIC (Schwarz, 1978) for the model M_k with dimension k is defined as

$$\mathrm{BIC}(M_k) = -2 \log L(M_k) + \log(n)\, k,$$

where n is the sample size. BIC is motivated by a Bayesian framework in the sense that the model minimizing BIC corresponds to the model with the highest posterior probability. Due to the larger penalty of $\log(n)$ on the model complexity as opposed to 2 for AIC, BIC often selects a sparser model compared to AIC.

- Extended Bayesian information criterion (EBIC):
Chen et al. (2008) proposed a generalization of BIC for the settings with $p > n$ in which case the regularization imposed by BIC on model complexity is not sufficient. A version of EBIC proposed by Chen et al. (2008) uses the following criterion function:

$$\mathrm{EBIC}(M_k) = -2 \log L(M_k) + \log(p \vee n)\, k,$$

where $(p \vee n)$ denotes the maximum of p and n. Note that it simply increases the penalty term of $\log(n)$ in BIC to $\log(p \vee n)$.

3 The penalization framework

When the covariates are highly correlated, the least squares estimator, although unbiased, suffers from inflated variance. This is because the matrix $G = X^\top X$ is nearly singular, which causes its inverse to be ill-conditioned even if it exists. Motivated by this problem, "ridge regression" introduces and additional term to the least squares objective function in an attempt to regularize the resultant estimator. The ridge regression objective function is given by

$$R(\beta) = \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where λ is a tuning parameter that controls the amount of penalization or regularization. In comparison to the information criterion functions that use the model size (or equivalently the L_0 norm of the regression vector β) to measure the model complexity, the regularization term of ridge regression use the L_2 norm of the regression vector. At one extreme with $\lambda = 0$, ridge regression estimator is the LSE and at the other extreme of $\lambda \to \infty$, it is the zero vector. For intermediate values of λ, it provides a shrinkage toward zero. The ridge regression estimator is given by

$$\hat{\beta}^{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top Y.$$

For $\lambda > 0$, the ridge estimator introduces some bias, but it helps reduce variance when $X^\top X$ is nearly singular. For the special case of the orthogonal design with $X^\top X = nI$, the ridge regression estimator has a simple relationship with the least squares estimator given by

$$\hat{\beta}^{\text{ridge}} = \frac{1}{n+\lambda} X^\top Y = \frac{1}{(1+\lambda/n)} \hat{\beta}^{ols}.$$

This provides an intuition for the type of shrinkage ridge estimator provides as it shrinks the least squares estimator toward zero. Therefore, the ridge estimator is biased but has less variance compared to the least squares estimator. In high dimensions with $p > n$, even though $X^\top X$ is necessarily singular, the ridge estimator is still well-defined unlike the LSE. However, ridge estimator is not sparse since none of its components are nonzero. While ridge regression may not perform variable selection by default, the properties of the ridge estimator for prediction and estimation have been studied for the large p setting (Dicker, 2016; Dobriban and Wager, 2018; Hsu et al., 2014).

3.1 LASSO and generalizations

Tibshirani (1996) proposed using an L_1 regularized estimator popularly known as the LASSO estimator. The LASSO estimator minimizes the objective function

$$L(\beta) = \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Although very similar in form to the ridge regression, the LASSO estimator is quite special since it is a sparse estimator. It is not only a valid estimator when $p \gg n$, but the number of nonzero components of the LASSO estimator can be much smaller than the sample size for appropriately chosen values of λ. In other words, LASSO estimator does both estimation and variable selection.

To see this, consider the orthogonal design case with $X^\top X = n \times I$, where the LASSO estimator can be written as:

$$\hat{\beta}_j^{\text{lasso}} = \begin{cases} \hat{\beta}_j^{\text{ols}} - \dfrac{\lambda}{2n} & \text{if } \hat{\beta}_j^{\text{ols}} > \dfrac{\lambda}{2n} \\ 0 & \text{if } |\hat{\beta}_j^{\text{ols}}| \leq \dfrac{\lambda}{2n} \\ \hat{\beta}_j^{\text{ols}} + \dfrac{\lambda}{2n} & \text{if } \hat{\beta}_j^{\text{ols}} < -\dfrac{\lambda}{2n}, \end{cases}$$

where $\hat{\beta}_j^{\text{ols}}$ denotes the jth component of the least squares estimator. The relationship between the LASSO estimator and the least squares estimator for the orthogonal design provides important insights about the LASSO estimator. When the magnitude of the least squares estimator is smaller than $\lambda/2n$, the LASSO estimator sets it to zero and otherwise, it shrinks the magnitude of the coefficient by $\lambda/2n$. Fig. 1 illustrates the relationship between the least squares estimator, ridge regression estimator, and the lasso estimator for the orthogonal design case. We can see from the figure that the ridge estimator's shrinkage is multiplicative and hence implies larger bias for coefficients with a large magnitude. On the other hand, the shrinkage of LASSO sets coefficients with small magnitudes exactly to zero while the bias remains constant for all the coefficients with a large magnitude.

Under quite weak regularity conditions, the LASSO estimator is shown to have optimal theoretical properties for the purposes of estimation and prediction (Bickel et al., 2009; Bühlmann and van de Geer, 2011; Meinshausen and Yu, 2009; Mousavi et al., 2017; van de Geer, 2008; Zhang and Huang, 2008). However, for LASSO to have desirable theoretical properties for selection,

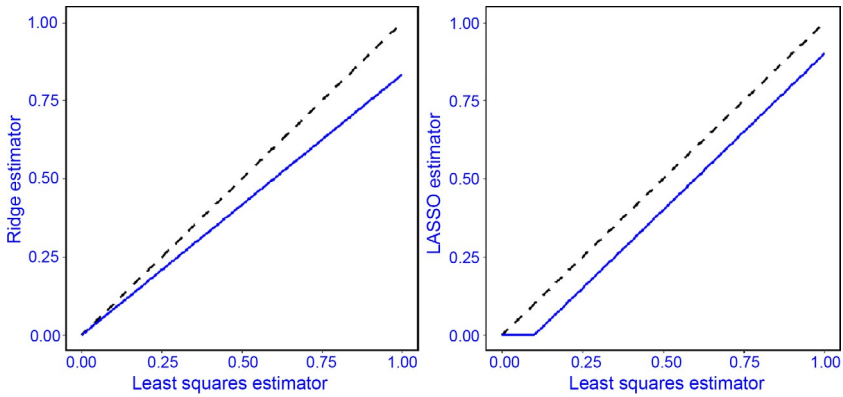

FIG. 1 The shrinkage of the ridge regression estimator and the LASSO estimator as a function of the least squares estimator for the orthogonal design.

it requires quite stringent conditions on the design matrix called as irrepresentable conditions (Zhao and Yu, 2006), a version of which will be described in the following.

3.1.1 Strong irrepresentable condition

Zhao and Yu (2006) discussed the strong irrepresentable condition under which the LASSO estimator can consistently perform perfect variable selection. However, this condition does not hold even if the correlations between the covariates are moderately high. Write the matrix $G = X^\top X$ as:

$$G = \begin{pmatrix} X_A^\top X_A & X_A^\top X_I \\ X_I^\top X_A & X_I^\top X_I \end{pmatrix} := \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}$$

The strong irrepresentable condition states that $|G_{21}G_{11}^{-1}\,\mathrm{sign}(\beta_A)| < 1$, where β_A is the regression vector restricted to the active variables. To appreciate the restrictive nature of this condition, Zhao and Yu (2006) considered the case $G = rJ + (1-r)I$, where J is the matrix of all 1's and $0 < r < 1$ is the common correlation between the predictors. In this case, the irrepresentable condition requires that $r < \frac{1}{(1+cs)}$, where s is the number of nonzero components of β, also referred to as the sparsity level. This is a strong condition especially when s is large and demonstrates that while LASSO is suitable for estimation and prediction, it may not be best suited for variable selection since an approximate version of the strong irrepresentable condition is also necessary for LASSO to be selection consistent (Zhao and Yu, 2006).

3.1.2 Adaptive LASSO

Given an initial consistent estimator $\hat{\beta}$, the adaptive LASSO estimator (Zou, 2006) minimizes

$$AL(\beta) = \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j|}.$$

This helps in providing adaptive shrinkage—larger penalty for smaller coefficients and smaller penalty for larger coefficients. When a good initial estimator is available, this turns out to be a good strategy as it also provides a natural scaling of the coefficients. While the adaptive LASSO helps to improve the performance of LASSO when there is a good initial estimator available, it still lacks model selection consistency under weak conditions.

Ideally, under the sparsity assumption, one would like to utilize the L_0 norm penalty for regularization to obtain model selection consistency. However, L_0 regularized objective function may not be feasible in practice with large p as the number of models to be evaluated is 2^p, which is huge. The L_1 norm is the smallest L_q norm which is convex and is therefore a natural

relaxation to the L_0 in view of its computational appeal. However, a drawback of the L_1 regularization is that the penalty term is linearly proportional to the magnitude of the coefficient β, which causes high bias for estimating the coefficients with large magnitude.

3.1.3 Elastic net

The elastic net (Zou and Hastie, 2005) penalty attempts to combine to advantages of both ridge regression and LASSO, namely shrinkage and sparsity together. The elastic net estimator minimizes

$$EN(\beta) = \sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}|\beta_j|^2.$$

Due to the ridge regularization, the elastic net estimator can handle correlations between the predictors better than LASSO and due to the L_1 regularization, sparsity is obtained. However, the bias issue present for LASSO is still present for elastic net.

3.2 Nonconvex penalization

Motivated by the bias induced by convex penalties such as the LASSO, Fan and Li (2001) and Fan and Peng (2004) proposed a nonconvex penalty called the smoothly clipped absolute deviation (SCAD) penalty. Although nonconvex penalized objective functions may not have a unique minimizer, several computational algorithms which attempt to provide good solutions have been proposed (Breheny and Huang, 2011; Fan and Li, 2001; Mazumder et al., 2011; Zhang, 2010). The SCAD penalty function is given by (Fig. 2):

$$\rho_\lambda^{\text{SCAD}}(\beta) = \begin{cases} \lambda|\beta|, & \text{if } |\beta| \le \lambda. \\ \dfrac{\left(2a\lambda|\beta| - \beta^2 - \lambda^2\right)}{2(a-1)}, & \text{if } \lambda < |\beta| \le a\lambda. \\ \dfrac{\lambda^2(a+1)}{2}, & \text{if } a\lambda \le |\beta|. \end{cases} \tag{4}$$

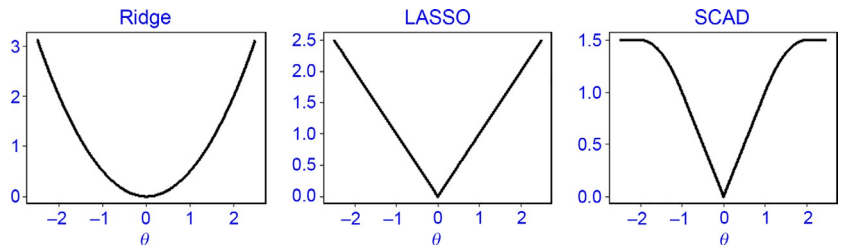

FIG. 2 Penalty functions corresponding to different penalization methods.

With a similar motivation Zhang (2010) proposed another nonconvex penalty called the minimax concave penalty (MCP). The idea behind these penalties is that they start penalizing the coefficients near zero in an L_1 manner similar to the LASSO penalty. However, as the magnitude of the coefficient becomes larger, the amount of their penalty smoothly decreases to zero. The rate at which this regularization decreases is more drastic for the MCP penalty compared to the SCAD penalty. In either case, these penalties avoid the bias induced by LASSO for large magnitude coefficients as the penalty becomes small.

An advantage of the nonconvex penalties is that they exhibit attractive theoretical properties for variable selection without stringent conditions as required for LASSO. For instance, Zhang (2010) and Loh and Wainwright (2017) showed that the nonconvex procedures achieve selection consistency under weaker conditions on the design matrix compared to LASSO.

3.3 Variable screening

When the dimension is extremely large, variable selection methods may not be feasible both in terms of computational implementation and theoretical performance. In such cases, screening methods play an important role as they reduce the dimension to a manageable size using computationally feasible approaches. Fan and Lv (2008) proposed using the marginal correlations between the covariates and the response for screening out covariates that have low correlation. Define $\rho_j = \text{cor}(X_j, Y)$, and define $\rho_{(j)}$ by ordering the ρ_j's based on their magnitudes so that $|\rho_{(1)}| < \cdots < |\rho_{(j)}| < \cdots < |\rho_{(p)}|$. Then the covariates having their magnitude of correlation smaller than $|\rho_{(p-K)}|$, that is, $\{j : |\rho_j| < |\rho_{(p-K)}|\}$, are screened out (excluded) from further analysis.

Under some conditions on the design matrix and the data generating model, Fan and Lv (2008) showed sure screening property that assures that all the relevant variables are selected by this marginal correlation-based screening. However, for such results to hold, the marginal correlation between each relevant covariate and the response should be large. This avoids the situations where some covariates may not have marginal correlation with the response but have joint effects in presence of other covariates.

Fan and Song (2010) generalized this to the GLM setting where marginal correlation is replaced by the marginal effect measured by fitting a one-covariate GLM model for each covariate. He et al. (2013) proposed screening based on the conditional quantiles of Y given each covariate marginally, motivated by the flexibility of quantile regression (Koenker, 2005; Koenker and Bassett, 1978). A nice feature of this approach is that it is much more robust to outliers. Moreover, quantile-based screening does not require specification of a model and can also handle heterogeneity of the observations in the data. For these reasons, it is a strong alternative to correlation-based screening.

4 The Bayesian framework for model selection

Model choice has been a very important topic within the Bayesian framework. Bayesian hypothesis testing can be viewed as a special case of the Bayesian approach to model selection. Consider the following hypothesis testing problem:

$$H_0 : M_0 \text{ vs } H_1 : M_1,$$

where M_0 and M_1 are two competing models (or two probability distributions). The Bayesian approach specifies prior probabilities for the hypotheses, say p_0 for H_0 and $p_1 = (1 - p_0)$ for H_1. If $\pi(Data \mid M_0)$ and $\pi(Data \mid M_1)$ denote the data generating distributions under the two different models, respectively, then the Bayesian approach involves computing the posterior probabilities of the models M_0 and M_1 given the data, namely

$$P[H_0 \,|\, Data] = \frac{P[Data \,|\, H_0]\, P[H_0]}{(P[Data \,|\, H_0]\, P[H_0] + P[Data \,|\, H_1]\, P[H_1])}. \tag{5}$$

Therefore, the ratio of the posterior probabilities is given by

$$\underbrace{\frac{P[H_1 \,|\, Data]}{P[H_0 \,|\, Data]}}_{\text{Posterior odds}} = \underbrace{\frac{P[Data \,|\, H_1]}{P[Data \,|\, H_0]}}_{\text{Bayes factor}} \underbrace{\frac{P[H_1]}{P[H_0]}}_{\text{Prior odds}}. \tag{6}$$

The first term in the RHS is the Bayes factor and the second term is the prior odds. The LHS is the posterior odds. Kass and Raftery (1995) provide concrete but ad hoc guidelines on how to interpret Bayes factors in terms of the strength of evidence they provide against the null hypothesis. For example, they suggest that Bayes factor larger than 10 provides a strong evidence for the alternate hypothesis. Note that, unlike with frequentist hypothesis testing, the Bayesian approach is symmetric in the hypotheses tested since the posterior odds could just as well be defined as $P[H_0 \,|\, Data]/P[H_1 \,|\, Data]$.

Another advantage of the Bayesian approach is that it can be easily generalized to any number of hypothesis by placing a prior distribution on all the hypotheses. However, an issue that arises with multiple hypotheses is that of adjusting for multiplicity. Scott and Berger (2010) studied Bayesian testing of multiple hypotheses and proposed strategies for multiplicity adjustment.

Let us now consider the normal means model which is an important special case of the linear regression model. Suppose that we have n independent observations

$$X_j \sim N(\mu_j, \sigma^2), j = 1, \ldots, n,$$

where the variance σ^2 is known and can potentially depend on n. We are interested in testing the hypotheses

$$H_0^j : \mu_j = 0 \text{ vs } H_1^j : \mu_j \neq 0.$$

In the single hypothesis testing context, it is perhaps most natural to assign a prior probability of $P[H_0] = P[H_1] = 0.5$ as an objective choice. However, under the multiple hypotheses context, such a prior of $P[H_0^j] = P[H_1^j] = 0.5$, independently across $j = 1, \ldots, n$ does not provide multiplicity control. Although this prior may seem like an objective prior on each of the hypothesis individually, Scott and Berger (2010) calls it a pseudo-objective since it has a prior expectation of $n/2$ hypothesis to be nonnull. This prior also leads to the probability that all nulls are true is $(1/2)^n$, which would be minuscule even for moderate n. One way to handle the issue of multiplicity in a Bayesian way is to first consider a prior on the possibility that all nulls are true followed by conditionally specifying prior probabilities on each individual hypothesis. Westfall et al. (Biometrika, 1997) adopted one such strategy that leads to a Bayesian version of Bonferroni correction for multiplicity. In general, a challenge with multiple hypothesis setting and more specifically with high-dimensional model selection is that there is no unique way of defining an objective prior.

Let us now consider the more general linear regression model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

The Bayesian framework relies on a likelihood specification for the observed data and a prior distribution on the parameters of interest. For linear regression, the most natural likelihood comes from specifying a Gaussian distribution on the errors. That is,

$$Y \mid (X,\beta) \sim N(X\beta, \sigma^2 I).$$

Our objective here is to select the model corresponding to the nonzero components of $\boldsymbol{\beta}$ from all the possible models. This problem can be formulated as a multiple hypothesis testing problem for the collection of the hypotheses

$$H_0^j : \beta_j = 0 \text{ vs } H_1^j : \beta_j \neq 0, j = 1, \ldots, p.$$

Define binary indicator variables Z_j to indicate whether the hypothesis H_1^j is true, and the binary vector $Z = (Z_1, \ldots, Z_p)$ that defines a model as it uniquely identifies the nonzero components of β. The idea is to obtain the posterior distribution of the model vector Z, which can be used to perform model selection. There are different ways to the posterior distribution $\pi(Z \mid Y)$ can be used for model selection:

- Maximum a posteriori (MAP) model: the MAP model maximizes the posterior distribution, that is, it finds the model that maximizes the posterior probability, that is, the model

$$\arg\max_k P[Z = k|Y],$$

where k denotes an arbitrary model coded as a binary vector with ones corresponding to the active covariates and zeroes corresponding to inactive covariates, and $P[Z = k|Y]$ is the posterior probability that k denotes the data generating model. As an example, $k = (1, 1, 1, 0, \ldots, 0)$ indicates a model with the first three covariates active. To obtain the MAP model, one would need to evaluate all the 2^p possible models which can be quite expensive if p is large. A good approximation for the posterior distribution is also difficult to obtain in high dimensions. Iterative methods (Hans et al., 2007; Yang et al., 2016) that aim to obtain a good model having posterior probability close to the MAP model are often used in practice.

- Median probability model: Find the set of covariates whose marginal posterior probabilities exceed 0.5, that is, $\{j : P[Z_j = 1|Y] > 0.5\}$. Barbieri and Berger (2004) called the model corresponding to this set of covariates as the median probability model and studied its theoretical properties.
- The threshold of 0.5 used by the median probability model is arguably ad hoc and an alternative threshold may be used for better model selection performance. An adaptive way to choose the threshold is to obtain the models corresponding to different threshold values followed by the use of a criterion function such as BIC to select a final model. This strategy was used by Narisetty and He (2014).

Several papers including Scott and Berger (2006, 2010) discuss prior distributions on the hypotheses which induce some level of multiplicity adjustment. Two of the most commonly used priors are given by

- The Z_j's have independent Bernoulli priors

$$P[Z_j = 1] = 1 - P[Z_j = 0] = q, \quad q \sim \text{Beta}(a, b).$$

Note that here q is an unknown parameter which has a Beta prior distribution. Scott and Berger (2010) discuss this fully Bayesian approach along with an empirical Bayes approach which estimates the prior probability q based on the data. Several other authors including (Narisetty and He, 2014; Scott and Berger, 2010; Yang et al., 2016) treat q as a hyperparameter and provide conditions on q for achieving appropriate multiplicity adjustment.

- Castillo et al. (2015) and Martin et al. (2017) used the following form of priors on Z:

$$\pi(Z = k) \propto \binom{p}{|k|}^{-1} f(|k|),$$

where $|k|$ is the number of nonzero components of k, also referred to as the size of the model k, and $f(\cdot)$ is a distribution on the model size. The intuition behind this prior is that the prior probability of a model depends only on its

size and the form of f determines the prior distribution on the model size. Castillo et al. (2015) proposed the following specific form for $f(\cdot)$:

$$f(|k|) \propto c^{-|k|} p^{-a|k|}, \; c, a > 0. \tag{7}$$

This prior places exponentially decreasing prior mass on models as their size increases, which encourages sparser models to be selected.

Once the prior distribution on the model space is provided, to carry out the Bayesian framework, we would need a prior distribution on the parameter vector β. A wide variety of prior distributions are considered for this purpose. We now discuss some of these priors, which are broadly called as spike and slab priors in the literature.

5 Spike and slab priors

In the Bayesian literature on variable selection and shrinkage, there are two primary classes of prior distributions for the regression coefficient β_j under the null hypothesis that $Z_j = 0$: (i) the point mass spike prior which is a degenerate distribution placing all of its probability mass at $\beta_j = 0$, and (ii) continuous (spike) priors which take the view that the magnitude of β_j under the null hypothesis is "small" but need not be exactly zero.

Before we proceed, we define some notation to be used in the rest of the chapter. We use k to denote a generic model which is a binary vector of length p indicating which covariates are active. For instance, $k = (1, 1, 1, 0, \ldots, 0)$ indicates a model with the first three covariates active. We use β_k and X_k to denote the components and columns of β and X corresponding to the nonzero components of k, respectively. Similarly, β_{k^c} and X_{k^c} denote the components and columns of β and X corresponding to the zero components of k, respectively.

5.1 Point mass spike prior

Under the hypothesis $Z_j = 0$, the point mass prior at zero is a natural prior choice and hence has been considered by several authors in the literature. Mitchell and Beauchamp (1988) considered the point mass spike prior $Z_j = 0$, and for the hypothesis $Z_j = 1$, proposed a proper uniform prior on β_j (which has a density that looks like a slab) giving raise to the well-known spike and slab prior terminology. See Fig. 3 for an illustration of the spike and slab priors with different slab distributions including the uniform slab prior.

5.1.1 g-priors

Zellner (1986) proposed a multivariate normal prior for the regression coefficients β given a model. That is, given that $Z = k$, $\beta_{k^c} = 0$ and the active part $\beta_k \,|\, \phi \sim N(0, \frac{g}{\phi}(X_k^\top X_k)^{-1})$ and ϕ has an improper diffuse prior given by

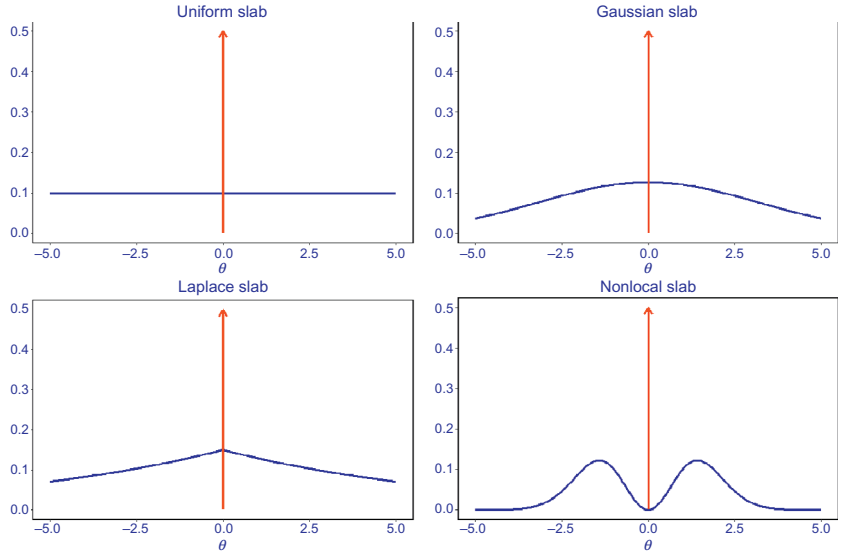

FIG. 3 Examples of point mass spike priors with different slab priors. The (*red*) *arrow* indicates a point mass at zero and the slab priors are (i) a Uniform prior on [−5,5], (ii) a Gaussian prior with variance 10, (iii) a Laplace prior with scale 0.15, and (iv) a nonlocal prior from the pMOM prior family with $r = 1$.

$\pi(\phi) \propto \frac{1}{\phi}$. In this framework, g is a hyperparameter to be chosen. Several authors had provided recommendations for the values of g. For example, Foster and George (1994) suggested $g = p^2$ and Kass and Raftery (1995) suggested $g = n$ and Fernández et al. (2001) suggested $g = \max(n, p^2)$.

To provide better variable selection properties, Liang et al. (2008) proposed using an additional prior distribution on g to obtain a mixture of g-priors. The prior distribution on g proposed by Liang et al. (2008) is given by

$$\pi(g) = \frac{(a-2)}{2}(1+g)^{-a/2}, \; g > 0, a > 2.$$

In the low-dimensional setting with p fixed, they showed model selection consistency of the mixture g-priors.

5.1.2 Nonlocal priors

Johnson and Rossell (2012) proposed using nonlocal priors for Bayesian variable selection. The main motivation behind the nonlocal priors is that the slab prior should ideally not have a fixed positive mass around zero since the slab prior should represent signals with nonzero magnitude. A nonlocal prior on the other hand has a density function that converges to zero as the magnitude of the parameter converges to zero. A more formal definition of local and nonlocal priors can be found in Johnson and Rossell (2012).

Examples of nonlocal priors are the product moment prior (pMOM) and the product inverse moment (piMOM) prior. A pMOM prior for a given model k is given by

$$\pi(\beta_k \mid \tau, \sigma^2) \propto \exp\left\{-\frac{1}{2\tau\sigma^2}\beta_k^\top \beta_k\right\} \prod_{j=1}^{k} \beta_{k_i}^{2r},$$

and a piMOM prior under a model k is given by

$$\pi(\beta_k \mid \tau, \sigma^2) \propto \prod_{j=1}^{k} \beta_{k_j}^{-(r+1)} \exp\left\{-\frac{\tau\sigma^2}{\beta_{k_j}^2}\right\},$$

where τ, σ^2 are hyperparameters. It can be seen that both the pMOM and piMOM priors have densities converging to zero as the magnitude of the regression coefficient converges to zero. With these priors, Johnson and Rossell (2012) showed that the posterior concentrates on the true model with probability going to one for $p \leq n$. In particular, if t denotes the true model, the authors showed that $P[Z = t \mid Y] \xrightarrow{\text{P}} 1$, where $\xrightarrow{\text{P}}$ indicates convergence in probability. This notion of consistency is called as global model selection consistency or strong model selection consistency, and is much stronger than the more commonly considered consistency in terms of pairwise Bayes factors, that is, $P[Z = k \mid Y]/P[Z = t \mid Y] \xrightarrow{\text{P}} 0$ for each model k marginally. However, even when this happens, it is still possible that the posterior probability of the true model tends to zero, that is $P[Z = t \mid Y] \xrightarrow{\text{P}} 0$. This is because there are $2^p - 1$ false models and the cumulative probability of all of them can be large even if each of them individually is small in comparison to the true model. In fact, the strong selection consistency is equivalent to having

$$\sum_{k \neq t} \frac{P[Z = k \mid Y]}{P[Z = t \mid Y]} \xrightarrow{\text{P}} 0,$$

and is a much stronger statement since the sum is over a large collection of models.

More interestingly, Johnson and Rossell (2012) also showed that using local priors, meaning that priors whose mass around zero does not tend to zero, it would not be possible to achieve such a posterior concentration. This is an important result since it provides a guideline on how to choose slab priors in the high-dimensional setting. It is important to note that while Gaussian priors with a fixed variance would be considered as local priors since the density at zero is positive, a Gaussian prior can still achieve the properties of a nonlocal prior if its variance parameter is allowed to increase to infinity as discussed by Narisetty and He (2014). This is because the Gaussian prior mass around zero would become small and tend to zero with an increasing variance.

The posterior distributions corresponding to nonlocal priors may not be computed using standard computational algorithms such as Gibbs sampler.

Johnson and Rossell (2012) proposed using Laplace approximations for approximate posterior computation. In high dimensions, their computational burden could be quite high. Therefore, more recently Shin et al. (2018) proposed a scalable computational algorithm for computing the posterior with nonlocal priors for large high dimensions. This algorithm called Simplified Shotgun Stochastic Search with Screening (S5) generalizes the shotgun stochastic search (SSS) algorithm of Hans et al. (2007) for more efficient sampling of the model space.

5.2 Continuous spike priors

For point mass spike priors, under each possible model k for Z, the dimension of the corresponding regression vector β_k is different. Methods for posterior computation for this setting tend to be computationally intensive due to the change in dimension which motivated several authors to consider continuous priors for the spike distribution $\beta_j \mid Z_j = 0$, for $j = 1, \ldots, p$. That is, priors of the form

$$\beta_j \mid Z_j = 0 \sim \pi_0(\beta_j), \quad \beta_j \mid (Z_j = 1) \sim \pi_1(\beta_j), \tag{8}$$

with both π_0 and π_1 being continuous distributions. The prior π_0 still focuses majority of its probability mass around zero. Priors on the model indicator Z are placed as discussed in Section 4.

A foremost example of the continuous spike and slab prior framework was proposed by George and McCulloch (1993):

$$\begin{gathered} \mathrm{Y} \mid (X, \beta, \sigma^2) \sim N(X\beta,\ \sigma^2 I), \\ \beta_j \mid \left(\sigma^2,\ Z_j = 0\right) \sim N(0,\ \sigma^2 \tau_0^2), \quad \beta_j \mid \left(\sigma^2,\ Z_j = 1\right) \sim N(0,\ \sigma^2 \tau_1^2), \\ P(Z_j = 1) = 1 - P(Z_j = 0) = q, \\ \sigma^2 \sim IG(\alpha_1, \alpha_2), \end{gathered} \tag{9}$$

where $0 < \tau_0^2 < \tau_1^2 < \infty$ are the variances of the spike and slab priors, respectively that may be tuned and IG denotes an inverse Gamma distribution. The intuition behind this setup is that the covariates with zero or very small coefficients will be identified with zero Z values, and the active covariates will be classified as $Z = 1$. We use the posterior probabilities of the latent variables Z to identify the active covariates (see Fig. 4).

Ishwaran and Rao (2005) studied several theoretical properties related to the framework defined by (9) when the dimension p is fixed. In particular, they studied the consistency properties of the posterior mean as an estimator. They also studied the variable selection properties of their selection method that is based on thresholding the posterior mean. While their results provide substantial insights about the shrinkage properties of the spike and slab prior framework, they are not applicable to the high-dimensional setting which is the current interest.

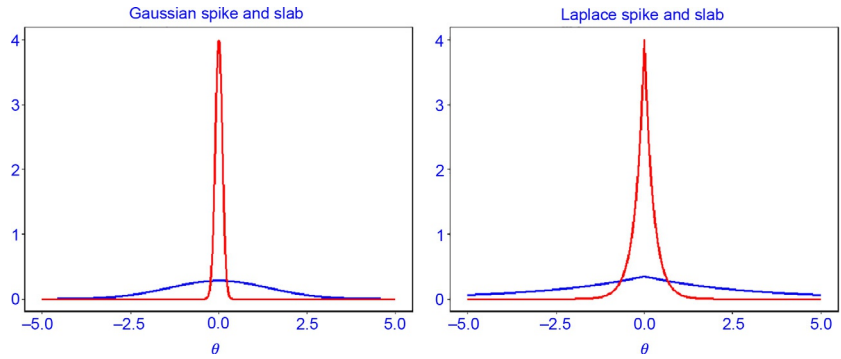

FIG. 4 Examples of continuous spike and slab priors: spike and slab Gaussian priors, and spike and slab LASSO priors.

Narisetty and He (2014) studied the model selection properties associated with these priors and provided insightful results on how the prior parameters τ_0^2, τ_1^2, and q should be selected to depend on n and p for achieving appropriate shrinkage and model selection performance. The spike and slab prior variances are set so that $\tau_0^2 \to 0$ and $\tau_1^2 \to \infty$ as n goes to ∞, where the specific rates of convergence depend on n and p. We refer the reader to Narisetty and He (2014) for more specific details about the requirements on these prior parameters. With these prior conditions, Narisetty and He (2014) provided two insightful results about the posterior distribution on the model space. The first is that as sample size n goes to ∞, even if the number of variables p is nearly exponentially large in n, the posterior probability of the true model goes to one under mild conditions, that is, $P[Z = t \mid Y] \xrightarrow{P} 1$. As will be discussed in Section 9, this is a much stronger result than the usual Bayes factor consistency commonly considered. Moreover, another insight is that the posterior on the model space induces a regularization similar to the L_0 penalty so that it acts as an information criterion asymptotically. For more detailed discussion on this, we refer to Narisetty and He (2014).

5.3 Spike and slab LASSO

Rockova (2018) and Rockova and George (2018) proposed and studied the spike and slab Laplace prior which places a two-component mixture of Laplace priors on the regression parameters (Fig. 5). More specifically, the spike and slab LASSO model is given by:

$$\begin{gathered} Y \mid (X, \beta, \sigma^2) \sim N(X\beta, \sigma^2 I), \\ \beta_j \mid (Z_j = 0) \sim \mathrm{LP}(\lambda_0), \quad \beta_j \mid (Z_j = 1) \sim \mathrm{LP}(\lambda_1), \\ P(Z_j = 1) = 1 - P(Z_j = 0) = q, \end{gathered} \tag{10}$$

where $\mathrm{LP}(\lambda)$ is the Laplace distribution with pdf given by $\psi(\beta|\lambda) = \frac{\lambda}{2}\exp\{-\lambda|\beta|\}$, and $\lambda_0 \gg \lambda_1$.

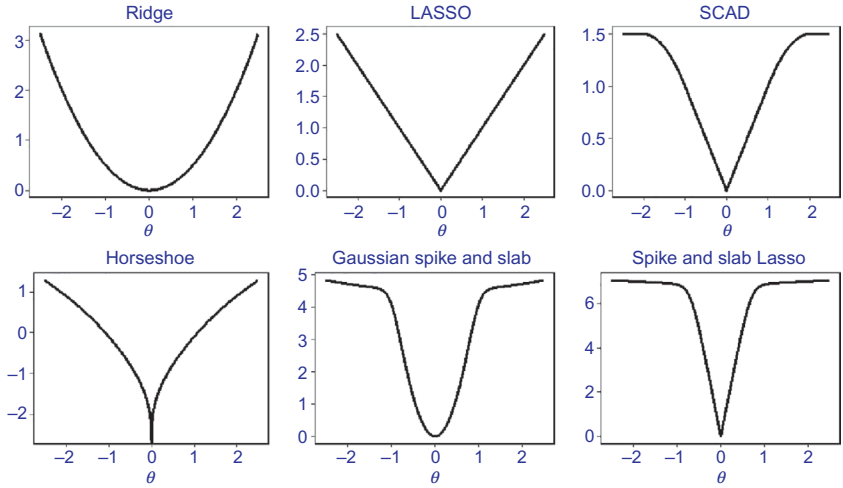

FIG. 5 Penalty functions corresponding to penalization methods on the top panel and penalty functions induced by different Bayesian methods on the bottom panel.

Unlike the previous literature which focused on using the posterior distribution of the Z's for variable selection, Rocková and George (2014) considered obtaining a point estimator for β which is the maximum a posterior (MAP) estimator corresponding to the posterior $\beta \mid (Y, X)$. The authors proposed a novel EM algorithm for obtaining this MAP estimator. More details about the EM algorithm for computation in Bayesian variable selection is deferred to Section 7.3.

Recently, Gan et al. (2018) used the spike and slab LASSO priors for estimation and sparsity recovery for graphical models and observed that optimal theoretical properties can be obtained due to this adaptive nature of the shrinkage. For a discussion about the shrinkage and regularization implicitly induced by different Bayesian methods including the spike and slab LASSO, see Section 6.4.

6 Continuous shrinkage priors

We now discuss the priors on the regression coefficients which provide shrinkage directly without introducing the binary latent variables as in the spike and slab approaches. Let us again consider the linear regression model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

With the Gaussian likelihood, a conjugate prior distribution on the regression vector β is again a Gaussian distribution. That is, $\boldsymbol{\beta}|\tau \sim N(0, \tau^2 I)$, where the parameter τ can either be treated as fixed or random. With a fixed τ, the posterior mean (and mode) corresponding to this Gaussian likelihood and Gaussian prior is the ridge regression estimator. More generally, there is a

correspondence between a specific prior distribution and a regularization imposed by the prior distribution at the MAP estimator. To see this, finding the MAP estimator that maximizes the posterior distribution is equivalent to minimizing $-\log\pi(\beta \mid Y)$, and is given by

$$\hat{\beta}^{MAP} = \arg\min_{\beta}\{-\log\pi(\beta \mid Y)\} = \arg\min_{\beta}\left\{-\log f(\beta;Y) + \underbrace{(-\log\pi(\beta))}_{\text{Bayesian-induced penalty}}\right\}. \tag{11}$$

Independent Laplace priors on the components of $\boldsymbol{\beta}$ leads to a posterior whose mode is LASSO, which forms motivation for Bayesian LASSO (Park and Casella, 2008).

6.1 Bayesian LASSO

Bayesian LASSO (Park and Casella, 2008) places independent Laplace prior distributions on the components of β:

$$\pi(\beta \mid \sigma,\lambda) \propto \prod_{j=1}^{p} \frac{\lambda}{2\sigma} \exp\left\{-\frac{\lambda|\beta_j|}{\sigma}\right\}.$$

The mode of the resultant posterior would be the same as the LASSO estimator with tuning parameter λ. The posterior distribution in addition provides uncertainty quantification for the regression parameters in the form of credible intervals. In classical low-dimensional settings, credible intervals obtained from a Bayesian posterior not only have a subjective Bayesian interpretation but also have valid frequentist properties in an asymptotic sense (van der Vaart, 1998). However, this is not the case in high dimensions. However, in the high-dimensional setting, credible intervals do not have an obvious interpretation in either sense since the prior is more aptly viewed as a shrinkage inducing tool rather than a belief inducing mechanism, and frequentist properties of the credible intervals from Bayesian LASSO are not established.

6.2 Horseshoe prior

The Laplace prior is a scale mixture of normal distributions. More specifically, if $\beta \mid \tau \sim N(0, \tau^2)$ and $\tau^2 \sim \text{Exp}(\lambda^2/2)$, then the marginal distribution of β is the double exponential distribution with the parameter λ. Therefore, when the variance of the normal distribution is exponentially distributed, it amounts to a Laplace distribution. The horseshoe prior takes an even more fat-tailed distribution, namely a Cauchy distribution, for the scale of the normal distribution to obtain the horseshoe prior (Carvalho et al., 2009a, b).

$$\beta \mid (\lambda,\tau) \sim N(0,\ \lambda^2\tau^2);\ \lambda \sim C^{+}(0,1),$$

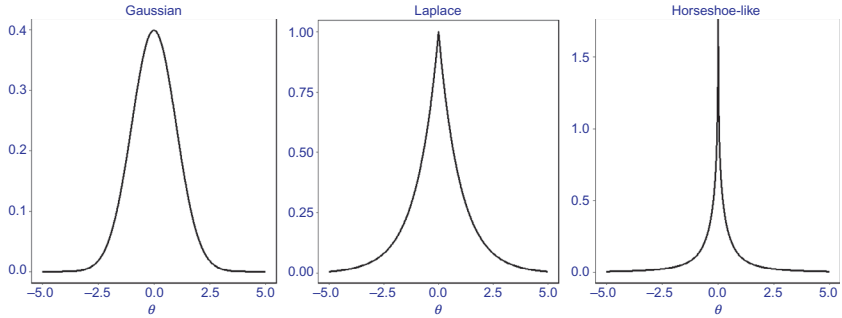

FIG. 6 Prior densities corresponding to some continuous shrinkage priors.

where $C^+(0, 1)$ denotes the half Cauchy distribution on the positive real line. Fig. 6 shows the density functions of Gaussian, Laplace, and horseshoe priors. The horseshoe prior is much more concentrated around zero compared to the Gaussian and Laplace priors inducing stronger shrinkage. Due to this, there has been a lot of interest in the use of horseshoe priors recently. We refer to Datta and Ghosh (2013) for an exposition on the Bayes risk properties of horseshoe estimator and Bhadra et al. (2017, 2019) for a comprehensive review of the horseshoe shrinkage approach and comparative studies with LASSO.

6.3 Global-local shrinkage priors

Polson and Scott (2010) noted that all continuous shrinkage priors can be written as global-local mixtures of normal priors in the following sense,

$$\beta_j \mid (\psi_j, \tau) \sim N(0, \psi_j \tau),\ \psi_j \sim f,\ \tau \sim g,$$

where τ represents the global shrinkage of β toward zero and ψ_j allow for differential shrinkage of each of the β_j's. Motivated by this, Bhattacharya et al. (2015) proposed the Dirichlet–Laplace prior framework (in the special case of the normal means model) which is a specific global-local shrinkage prior and is given by:

$$\beta_j \mid (\psi_j, \tau) \sim \mathrm{DE}(\psi_j \tau),\ \psi \sim \mathrm{Dir}(a, \ldots, a),\ \tau \sim \mathrm{gamma}(na, 1/2),$$

where DE is the double exponential distribution, Dir is the Dirichlet distribution, and a is a hyperparameter. They provide an augmented Gibbs sampling algorithm which can be used for posterior computation.

It is worth noting that there are many other continuous shrinkage priors considered in the literature and it is not possible to discuss them all here. The double Pareto shrinkage prior of Armagan et al. (2013) is one such example which is worth exploring by a reader interested in further reading. Finally, we note that continuous shrinkage priors such as the Horseshoe or the Dirchlet-Laplace priors do not directly provide a way to select the variables

since the posterior mode or mean corresponding to these methods need not be sparse. However, the estimator obtained from these methods can be thresholded to select variables. The choice of the threshold parameter would require some tuning. We discuss some strategies for selecting tuning parameters in the next section.

6.4 Regularization of Bayesian priors

As discussed previously, the prior distribution in the Bayesian framework implicitly induces shrinkage and regularization as indicated by the definition of the MAP estimator in Eq. (11). The Bayesian penalty function induced by the horseshoe prior can be approximately written as:

$$\log \pi_H(\beta) \approx \log \sum_{j=1}^{p} \log\left(1+\beta_j^{-2}\right).$$

The penalty function induced by the spike and slab priors can be written as:

$$\text{pen}(\beta_j) := -\log\left[(1-\theta)\psi(\beta_j|\lambda_0)+\theta\psi(\beta_j|\lambda_1)\right], \quad j=1,\ldots,p, \tag{12}$$

where $\psi(\cdot \mid \lambda_0)$, $\psi(\cdot \mid \lambda_1)$ are the densities of the spike and slab priors, respectively. As argued by Rockova and George (2018), these spike and slab priors have a desirable adaptive regularization property similar to nonconvex penalty functions such as the SCAD whose regularization decreases to zero as the magnitude of the coefficient increases. Fig. 5 provides a plot of the penalty functions corresponding to ridge regression, LASSO, and SCAD methods along with that corresponding to the Gaussian spike and slab prior, the spike and slab LASSO prior, and the horseshoe prior. It can be seen that regularization from all the Bayesian approaches are quite nonconvex and are closer to the L_0 penalty compared to that of the LASSO.

6.5 Prior elicitation—Hyperparameter selection

In the traditional subjective Bayesian context, one builds the prior distribution based on the belief or prior knowledge available. However, it is hard to be subjective in the high-dimensional context due to the vastness of the parameter space involved. Therefore, an objective Bayesian or an empirical Bayes stand is often taken under which the hyperparameters of the priors need to be selected based on some prespecified criterion or using the data themselves. We discuss a few such strategies for hyperparameter selection.

6.5.1 Empirical Bayes

Consider the following linear regression model with a generic prior indexed by a hyperparameter α:

$$Y \mid (X,\ \beta) \sim N(X\beta,\ \sigma^2 I),\ \beta \sim \pi_\alpha,$$

where α is the hyperparameter that needs to be selected. The basic idea of the empirical Bayes strategy is to treat the parameter of interest as latent and integrate it out to obtain a marginal likelihood for the hyperparameter α. Such a marginal likelihood can be used to obtain an estimator for α based on the observed data. The marginal likelihood is

$$ML(\alpha \mid Y) \propto \int L(Y \mid \theta)\pi_\alpha(\theta)d\theta. \tag{13}$$

Using the marginal likelihood for α, the empirical Bayes strategy is to find a point estimator. While any estimating method can be used in principle, a common approach is to maximize the marginal likelihood to obtain

$$\hat{\alpha} = \arg\max_{\alpha} \mathrm{ML}(\alpha \mid Y).$$

In some cases, especially with conjugate priors such as the Gaussian likelihood and the Gaussian prior case, closed form expressions are available to compute the above marginal likelihood as a function of α. However, it is not always possible to obtain the marginal likelihood in closed form. In such cases, Laplace approximation (discussed in Section 7) or MCMC-based strategies are also often used to approximate the integral in Eq. (13). For variable selection, George and Foster (2000) proposed a conditional empirical Bayes strategy in the context of g-priors and Yuan and Lin (2005) proposed an empirical Bayes method to select the prior hyperparameters in the context of point mass spike and Laplace slab priors.

6.5.2 Criterion-based tuning

The hyperparameters can also be selected based on some criterion function such as AIC or BIC discussed in Section 2. The idea here is that by using different values for the hyperparameter α, one would obtain different models among which the best one is selected based on a criterion function. More specifically, let $M(\alpha)$ denote the selected model corresponding to the hyperparameter value α. Then, BIC can be used to choose an optimal value for α as follows:

$$\hat{\alpha} := \arg\min_{\alpha} \mathrm{BIC}(M(\alpha)),$$

and $M(\hat{\alpha})$ will be the final model selected. Such a strategy is commonly used in the literature (Narisetty and He, 2014; Narisetty et al., 2019). Another alternative is to cross validated predicted error as the criterion to minimize in place of AIC or BIC.

7 Computation

Efficient computation is a crucial component of any statistical procedure in the Big Data era. There are a variety of computational approaches for Bayesian variable selection proposed in the literature. The objectives of some of

these computational approaches are quite different. For instance, the EM approaches attempt to obtain the maximum-a-posteriori estimator for β and do not necessarily attempt model selection directly. On the other hand, there are stochastic search approaches which operate only on the model space and do not explicitly consider parameter estimation. In the following, we will discuss some of the major computational approaches for Bayesian high-dimensional regression.

7.1 Direct exploration of the model space

7.1.1 Shotgun stochastic search

Hans et al. (2007) proposed this method which aims to search the space of models to obtain models having high posterior probabilities. The algorithm is similar to stepwise selection algorithms in the sense that at each step it considers a neighborhood of models and selects the model maximizing the posterior probability. Along this path, since many models having high posterior probabilities are visited, a set of models with high posterior probabilities are collected. In particular, a simplified version of the SSS algorithm starts with an initial model γ_0 and follows the following steps at iteration t to update the model from γ_{t-1} to γ_t for $t = 1, \ldots, T$.

- Compute $S(\gamma)$, the criterion function such as the posterior probability of the model γ, for all models in the neighborhood of γ_{t-1}. One way to define a neighborhood of a model is to consider all the models which either have one additional covariate, one covariate less, or one covariate different from the model γ_{t-1}. More specifically, these neighborhoods can be defined as:
  - Addition neighborhood: all the models that add one covariate to the current model γ_{t-1} are considered as neighbors. For instance, if the current model has covariates $\{1, 2\}$, the models $\{1, 2, 3\}, \{1, 2, 4\}, \ldots, \{1, 2, p\}$ will be in the neighborhood.
  - Deletion neighborhood: all the models that remove one covariate from the current model γ_{t-1} are considered as neighbors. If the current model has covariates $\{1, 2\}$, the models $\{1\}, \{2\}$ will be considered as neighbors.
  - Swap neighborhood: all the models that have one covariate different from the current model γ_{t-1} are considered as neighbors. If the current model has covariates $\{1, 2\}$, the models $\{1, 3\}, \{1, 4\}, \ldots, \{1, p\}, \{2, 3\}, \{2, 4\}, \ldots, \{2, p\}$ are neighbors.
- From the neighbors of γ_{t-1}, sample a model with probability proportional to the criterion function $S(\gamma)$ by normalizing the total probability within the neighboring set. Set the resultant model as γ_t.
- After the prespecified number of iterations T, choose the model which maximizes the criterion function among all the models visited.

This algorithm is a special case of Metropolis–Hastings random walk algorithms. These algorithms can be viewed as stochastic versions of stepwise algorithms which have been commonly used for variable selection. The performance of these methods could potentially depend on the initial model γ_0 and the neighborhood considered. Shin et al. (2018) generalized the shotgun stochastic search approach which they use for computation in their nonlocal prior setting. Liang et al. (2007) and Liang (2009) proposed another class of model space exploration approaches called stochastic approximation Monte Carlo (SAMC). SAMC operates by first partitioning the model space into disjoint subsets and by enforcing sampling from each of these subsets to avoid local trap issues often encountered by stochastic search algorithms. SAMC algorithms rely on selection of appropriate subsets of the model space and on estimating the posterior probabilities of the selected subsets. The details of these methods are quite involved and are beyond the scope of the current article, we refer to Liang (2009) and Liang et al. (2013).

7.2 Gibbs sampling

Gibbs sampling algorithms are quite commonly used for computation of Bayesian posterior distributions. In the context of variable selection, Gibbs sampling algorithms involving standard distributions can be used for computation when continuous spike and slab priors are used. For example, George and McCulloch (1993, 1997), Ishwaran and Rao (2005), and Narisetty and He (2014) used easy-to-sample-from Gibbs samplers based on Gaussian spike and slab priors. With the model (9), a standard Gibbs sampling algorithm would take the following form:

- The conditional distribution of β is given by $\beta \mid (Z, \sigma^2, Y, X) \sim N(V X^\top Y, \sigma^2 V)$, where $V = (X^\top X + D_z)^{-1}$, and $D_z = \text{Diag}(Z\tau_1^{-2} + (1-Z)\tau_0^{-2})$.
- The conditional probability for Z_j is

$$P(Z_j = 1 \mid (\beta, \sigma^2, Y, X)) = \frac{q\phi(\beta_j, 0, \sigma^2\tau_1^2)}{q\phi(\beta_j, 0, \sigma^2\tau_1^2) + (1-q)\phi(\beta_j, 0, \sigma^2\tau_0^2)}.$$

- The conditional of σ^2 is the Inverse Gamma distribution $IG(a, b)$ with $a = \alpha_1 + n/2 + p/2$, and $b = \alpha_2 + \beta^\top D_z\beta/2 + (Y - X\beta)^\top(Y - X\beta)/2$.

These are standard distributions that can be easily sampled. However, when the dimension of the design matrix p is large, the real challenge is that sampling from a p-variate normal distribution for β is computationally intensive. A direct sampling would typically require p^3 order operations as it requires the computation of the $p \times p$ matrix $(X^\top X + D_z)^{-1/2}$, which if computed using the eigenvalue decomposition of $(X^\top X + D_z)$ leads to p^3 order computational complexity.

The Skinny Gibbs algorithm (Narisetty et al., 2019) provides a simple and very effective modification of the Gibbs sampler to avoid the high computational

complexity in the case of large p. The idea is to split β into two parts in each Gibbs iteration, corresponding to the "active" (with the current $Z_j = 1$) and "inactive" (with the current $Z_j = 0$) subvectors. The active part has a low dimension, and is sampled from the multivariate normal distribution. The inactive part has a high dimension, but we simply sample it from a normal distribution with independent marginals. More specifically, the Skinny Gibbs sampler proceeds as follows, after an initialization.

Step 1 (for sampling β). Define the index sets A and I as the active (corresponding to $Z_j = 1$) and the inactive (corresponding to $Z_j = 0$) sets and decompose $\beta = (\beta_A, \beta_I)$ so that β_A and β_I contain the components of β corresponding to $Z_j = 1$ and $Z_j = 0$, respectively. Similarly rearrange the design matrix $X = [X_A, X_I]$. Then, the vector β is sampled as:

$$\beta_A \mid (Y,Z) \sim N(m_A, \sigma^2 V_A^{-1}), \quad \beta_I \mid (Y,Z) \sim N(0, \sigma^2 V_I^{-1}), \tag{14}$$

where $V_A = (X_A'X_A + \tau_1^{-2} I)$, $m_A = V_A^{-1} X_A' Y$, and $V_I = \mathrm{Diag}\,(X_I'X_I) + \tau_0^{-2} I = (n + \tau_0^{-2}) I$.

Step 2 (for sampling Z). Generate Z_j ($j = 1, \ldots, p$) conditioned on the remaining components of Z using the following conditional odds:

$$\begin{aligned} &\frac{P[Z_j = 1 \mid Z_{-j}, \beta, Y]}{P[Z_j = 0 \mid Z_{-j}, \beta, Y]} \\ &= \frac{q\phi(\beta_j, 0, \tau_{1,n}^2)}{(1-q)\phi(\beta_j, 0, \tau_{0,n}^2)} \times \exp\left\{\beta_j X_j' (Y - X_{C_j}\beta_{C_j})\right\}, \end{aligned} \tag{15}$$

where Z_{-j} is the Z vector without the jth component, and C_j is the index set corresponding to the active components of Z_{-j}, i.e., $C_j = \{k : k \neq j, Z_k = 1\}$.

Step 3 (for sampling σ^2). The conditional distribution of σ^2 given β and Z is the Inverse Gamma distribution $IG(a, b)$ with $a = \alpha_1 + n/2 + p/2$, and $b = \alpha_2 + \beta^\top D_z \beta/2 + (Y - X_A\beta_A)^\top (Y - X_A\beta_A)/2 + n\beta_I^\top \beta_I/2$, where the index sets A and I are the active and the inactive sets as defined in Step 1.

The main idea is that in Step 1, the update of β is modified such that the coefficients corresponding to $Z_j = 1$ and those corresponding to $Z_j = 0$ (denoted by β_I) are sampled independently, and the components of β_I are updated independently so that large matrix computations are avoided. That is, Skinny Gibbs modifies the precision matrix V_z as

$$V_z = \begin{pmatrix} X_A'X_A + \tau_1^{-2} I & X_A'X_I \\ X_I'X_A & X_I'X_I + \tau_0^{-2} I \end{pmatrix}$$

$$\Downarrow$$

$$\begin{pmatrix} X_A'X_A + \tau_1^{-2} I & 0 \\ 0 & (\mathrm{Diag}(X_I'X_I) + \tau_0^{-2}) I \end{pmatrix}.$$

As can be seen, the precision matrix is heavily modified in Step 1 which can alter the original Gibbs sampler. To compensate for the loss of correlation structure due to this modification, Step 2 of the Skinny Gibbs algorithm is designed to take into account this lost dependence structure. In spite of this modification, Narisetty et al. (2019) showed that the Skinny Gibbs algorithm retains the desired statistical properties such as the strong model selection consistency property which will be discussed in more detail in Section 8.

It is worth noting that the technique developed in the Skinny Gibbs algorithm is very general and can incorporate many modeling settings where the likelihood or priors involved can be written as mixtures of normal distributions. For instance, Narisetty et al. (2019) studied a Skinny Gibbs algorithm applied to logistic regression. This also applies to priors beyond normal priors which can be written as scale mixtures of normal priors such as Laplace priors.

Bhattacharya et al. (2016) considered an alternative approach to scale up the Gibbs sampling algorithms which involves large multivariate normal distributions. Their approach is to intelligently utilize properties of matrices to sample from the original high-dimensional normal distribution. While this algorithm has linear order complexity in terms of p, it has a quadratic complexity in terms of the sample size n while Skinny Gibbs has a linear order complexity in n. Moreover, Skinny Gibbs can be scaled up further by utilizing their matrix identities for sampling β_A when the size of $|A|$ is large.

7.3 EM algorithm

The EM algorithm is a popular technique to compute maximum likelihood estimators and maximum a posterior (MAP) estimators (Dempster et al., 1977). Even in high dimensions, Rocková and George (2014) found the EM algorithm to be effective for obtaining the MAP estimator corresponding to the spike and slab Gaussian prior specification given by (9) and Rockova and George (2018) generalized it to the spike and slab Lasso prior specification given by (10). The algorithm treats Z to be latent and implements an EM algorithm. Let us consider the model (9) having Gaussian spike and slab prior for illustration where the variances of the spike and slab priors are τ_0^2 and τ_1^2, respectively. The maximization problem involves the objective function

$$Q(Z,\beta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\tau_1^2 Z_j + \tau_0^2(1-Z_j)}.$$

E-Step: At the E-step, the conditional expectations of the Q function with respect to the Z variables given all the other parameters are obtained. That is, we need to find the conditional expectation

$$E\left(\frac{1}{\tau_1^2 Z_j + \tau_0^2(1-Z_j)}\right) = \frac{p_j^*}{\tau_1^2} + \frac{(1-p_j^*)}{\tau_0^2} := d_j,$$

where $p_j^* = E(Z_j \mid \beta) = \frac{a}{(a+b)}, a = \pi(\beta_j \mid Z_j = 1)\pi(Z_j = 1)$ and $b = \pi(\beta_j \mid Z_j = 0)$ $\pi(Z_j = 0)$.

M-Step: Then the conditional expectation of $Q(Z, \beta)$ given Z is essentially a weighted ridge regression penalized objective function with penalty weights given by the d_j's. Therefore, the maximization at the M-step has a closed form with the solution being

$$\hat{\beta}^{(k+1)} = (X^\top X + D)^{-1} X^\top Y,$$

where D is a $p \times p$ diagonal matrix with d_j as its diagonal elements. This yields a simple iterative algorithm for obtaining the MAP estimator corresponding to the spike and slab Gaussian prior specification.

However, one drawback of this approach is that the posterior probabilities of the Z variables are not obtained as part of the EM algorithm, which are important quantities for performing variable selection. As a proxy to these posterior probabilities, the conditional probabilities of $P[Z_j = 1 \mid \hat{\beta}]$, where $\hat{\beta}$ is the MAP estimator of β, are used for variable selection. Rockova (2018), Rockova and George (2018), and Gan et al. (2018) among others used the EM algorithm for MAP estimation and the conditional probability strategy for variable selection.

7.4 Approximate algorithms

7.4.1 Laplace approximation

In the context of point mass spike priors as discussed in Section 5.1, a direct computation of the posterior probability of a model requires evaluation of the marginal likelihood $P[Y \mid Z = k]$ as indicated by the posterior probability expression in Eq. (5). This can be written as

$$P[Y \mid Z = k] = \int_{\mathbb{R}^{|k|}} P[Y \mid Z = k, \beta_k]\, \pi[\beta_k \mid Z = k] d\beta_k.$$

The Laplace approximation for integrals uses quadratic Taylor's approximation for the integrand above (Raftery, 1996). This is similar to making a Gaussian approximation to the integrand above as a function of β_k which corresponds to the conditional posterior distribution of $\beta_k \mid (Y, Z = k)$. Therefore, performance of the Laplace approximation depends on how close this conditional posterior distribution is to a Gaussian distribution. The Laplace approximation is commonly used for posterior computation for Bayesian model selection. For instance, Yuan and Lin (2005) use the approximation for empirical Bayes variable selection using g-priors, Liang et al. (2007) use it in the mixture of g-priors context, and Johnson and Rossell (2012) use it for computation with their nonlocal priors.

7.4.2 Variational approximation

Variational approximation (Blei et al., 2017; Jordan et al., 1999) is a powerful and general strategy to approximate posterior distributions. The general idea

of variational approximation is to use a computationally feasible class of distributions from which the closest one to the posterior distribution is to be found. More specifically, suppose that $\mathbf{p}$ is the posterior distribution to be computed. For example, $\mathbf{p}$ can be the posterior distribution of $(\boldsymbol{\beta}, Z)$ given data corresponding to Model (9).

Consider the family of distributions Q which are easy to compute/sample from. From the set Q, the distribution $\hat{\mathbf{q}}$ which is closest to $\mathbf{p}$ in terms of minimizing the Kullback–Leibler distance is found. That is,

$$\hat{q} \leftarrow \arg\min_{\mathbf{q}\in Q} \mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}),$$

where $KL(\mathbf{p} \,||\, \mathbf{q})$ denotes the Kullback–Leibler distance between the distributions $\mathbf{p}$ and $\mathbf{q}$. In particular, if the family Q is indexed by a parameter vector θ so that $Q := \{\mathbf{q}_\theta : \theta \in \Omega\}$, then this corresponds to minimizing

$$\hat{\theta} \leftarrow \arg\min_{\theta\in\Omega} KL(\mathbf{p} \,\|\, \mathbf{q}_\theta)$$

The main challenge in this context is to find the family Q of distributions which is reasonably close to the posterior distribution π along with being computationally friendly. With point mass spike priors, Carbonetto and Stephens (2012) proposed the following Q family for variational approximation:

$$\mathbf{q}(\beta,Z,\theta) = \prod_{k=1}^{p} \mathbf{q}(\beta_k,Z_k,\theta_k),$$
$$\text{where } \mathbf{q}(\beta_k,Z_k,\theta_k) = \phi_k N(\beta_k \,|\, \mu_k,s_k^2)\mathbf{1}\{Z_k=1\} + (1-\phi_k)\delta_0(\beta_k)\mathbf{1}\{Z_k=0\} \tag{16}$$

This is a family of component-wise product distributions implying independence of different components. Although the posterior does not belong to this family, the hope is that there is one distribution in this family which is close to the posterior and that its important summary statistics are somewhat close to those of the posterior. Carbonetto and Stephens (2012) proposed a coordinate descent algorithm for the KL minimization problem which yields easily interpretable updates as follows:

$$s_k^2 = \frac{\sigma^2}{(X^\top X)_{kk} + (\sigma^2\tau_1^2)^{-1}}$$
$$\mu_k = \frac{s_k^2}{\sigma^2}\left((X^\top y)_k - \sum_{j\neq k}(X^\top X)_{jk}\,\phi_j\mu_j\right)$$
$$\frac{P[Z_k=1|(Y,X,\beta)]}{P[Z_k=0|(Y,X,\beta)]} \approx \frac{\phi_k}{(1-\phi_k)} = \frac{q}{(1-q)}\frac{s_k}{(\sigma\tau_1)}\exp\left\{\frac{\mu_k^2}{2s_k^2}\right\},$$

However, Huang et al. (2016) argued that the component-wise variational Bayes algorithm does not work well in the high dimensions. Motivated by

this, they proposed a new algorithm called the batch-wise variational Bayes algorithm where the parameters $(\mu_j)_{j=1}^p$, $(\alpha_j)_{j=1}^p$, and $(s_j)_{j=1}^p$ are updated simultaneously across different j instead of marginally updating for each j conditional on others. The simultaneous updates remain to be similar for the parameters α and s_j in the batch-wise algorithm whereas the update for μ changes as:

$$\mu = \left(\Phi X^\top X \Phi + n\Phi(1-\Phi) + \frac{1}{\tau_1}\right)^{-1} \Phi X^\top y,$$

where Φ is the $p \times p$ diagonal matrix with ϕ_k as diagonal elements, i.e., $\Phi = \text{Diag}(\phi_k)$.

While a direct computation of μ would be computationally expensive, μ can be sequentially updated much more efficiently based on the μ from the previous iteration. Huang et al. (2016) studied the model selection consistency associated with their variational Bayes in the high-dimensional setting when $p \geq n$. While majority of the variational approximations for Bayesian variable selection focused on the variational family given by (16), recently Ormerod et al. (2017) discussed some other choices and studied their properties when $p \leq n$.

8 Theoretical properties

In high dimensions, it is difficult to take an entirely subjective Bayesian view about the priors and think of the prior distribution as prior belief about the underlying parameters. This is because, due to high-dimensional probability involved, we do not quite have the right intuition about all the parts of the high-dimensional parameter space which can lead to some strong prior mass in some parts irrespective of which kinds of priors one uses. For example, if each of the coefficients has a standard normal prior which may be reasonable in a subjective Bayesian sense, the L_2 norm of the entire regression vector will be heavily concentrated around $\sqrt{p}$ which is extremely large in most applications. However, if the prior is chosen for convenience and does not quite reflect the prior uncertainty about the entire parameter space, the posterior uncertainty may not be scientifically interpretable. This provides the motivation to study objective properties associated with Bayesian procedures in the frequentist sense.

There are different types of theoretical properties considered in the Bayesian variable selection context. In this section, we provide a brief description of the different types of theoretical results considered and the related references. Let us denote the generic posterior distribution by $\pi(\beta, Z|Y)$. However, in some cases such as for continuous shrinkage priors (for example, Bayesian LASSO), the binary vector Z is not used in which case Z can be taken as the vector of all ones.

8.1 Consistency properties of the posterior mode

The posterior mode that maximizes the posterior $\pi(\beta \mid Y)$ is the maximum-a-posteriori (MAP) estimator of the parameter β. If obtaining a point estimator is sufficient in an application, studying the properties of the MAP estimator is quite natural. As discussed in Section 4, there is a natural correspondence between the MAP estimator and penalized estimators since the prior implicitly induces a penalty which is often termed Bayesian penalization or regularization. The penalty functions induced by many of the commonly used Bayesian priors are nonconvex and in some cases nonconcave. Rockova (2018) studied the estimation consistency properties of spike and slab LASSO and showed that the MAP estimator achieves the optimal rate of convergence in terms of L_2 norm error, L_1 norm error, and prediction loss. In the context of Bayesian graphical models, Gan et al. (2018) show that the MAP estimator using the spike and slab LASSO prior would also be consistent in terms of L_∞ norm, which is more relevant for model selection.

8.2 Posterior concentration

In the Bayesian context, studying the concentration properties of the posterior distribution is quite important (Ghosal et al., 2000). The posterior concentration for the regression vector β requires that

$$\sup_{\beta_0} E_{\beta_0} \pi(\beta : \| \beta - \beta_0 \| > \epsilon_n \mid Y) \xrightarrow{P} 0 \tag{17}$$

where $\|\cdot\|$ denotes the Euclidean norm and the rate of posterior concentration, ϵ_n, is defined implicitly to satisfy Eq. (17). The following are some existing works which studied the posterior concentration under various prior settings:

- Bhattacharya et al. (2015) studied the posterior concentration properties with their proposed Dirichlet–Laplace priors under the special case of the orthogonal design matrix $X^\top X = nI$. In particular, they show posterior concentration (17) with $\epsilon_n = O\left(\sqrt{\frac{s\log(n/s)}{n}}\right)$ with $s = \|\beta\|_0$, which is the optimal error rate for posterior concentration for the normal means model. Compared to the low-dimensional setting, the optimal error rate has an additional $\log(n/s)$ that represents the cost paid for the dimension, which is the same as the sample size n in this case.
- Castillo and van der Vaart (2012) and Castillo et al. (2015) studied the posterior concentration results extensively based on general choices of priors for the model size and for the slab prior. Under their prior choices, Castillo et al. (2015) showed posterior concentration with $\epsilon_n = O\left(\sqrt{\frac{s\log p}{n}}\right)$. In the high-dimensional setting with increasing p, the error rate ϵ_n pays a penalty factor of $\log p$ for not knowing the true model corresponding to the sparsity

of β_0, which is a standard phenomenon in high-dimensional analysis. Atchade (2017) further extended these results to modeling settings beyond the linear regression model.

- Martin and Walker (2014) and Martin et al. (2017) studied the posterior concentration results for an empirical Bayes approach that has prior distributions depending on the data. Their error rate for posterior concentration is $\epsilon_n = O\left(\sqrt{\frac{s\log(p/s)}{n}}\right)$, which matches the optimal error rate as discussed by Rigollet and Tsybakov (2012) and is slightly better compared to the error rate of Castillo et al. (2015).

8.3 Pairwise model comparison consistency

Casella et al. (2009) and Moreno et al. (2010) considered consistency in terms of the comparison of pairwise posterior probabilities. That is, if we consider any model $k \neq t$, where t is the true model, then the pairwise consistency requires that

$$\frac{P[Z=k \mid Y]}{P[Z=t \mid Y]} \xrightarrow{\mathrm{P}} 0.$$

Casella et al. (2009) showed that their intrinsic priors have the pairwise consistency property for fixed p case and Moreno et al. (2010) showed the same for potentially diverging dimension but with $p < n$. When the dimension p is fixed, there are finitely many models k so that this pairwise consistency implies that $P[Z=t \mid Y] \xrightarrow{\mathrm{P}} 1$. However, when the dimension p can grow to infinity, the pairwise consistency does not imply that the posterior probability of the true model converges to one. In fact, Johnson and Rossell (2012) argued that even under the pairwise consistency, the posterior probability $P[Z=t \mid Y]$ can go to zero, when p is not fixed. A stronger notion of consistency compared to the pairwise consistency in high dimensions can be defined as:

$$\max_{k \neq t} \frac{P[Z=k \mid Y]}{P[Z=t \mid Y]} \xrightarrow{\mathrm{P}} 0, \tag{18}$$

which assures that the posterior probability of the true model is uniformly larger in magnitude compared to any other model.

8.4 Strong model selection consistency

Johnson and Rossell (2012) considered a stronger version of Bayesian model selection consistency which requires that $P[Z=t \mid Y] \xrightarrow{\mathrm{P}} 1$. They showed that their proposed nonlocal priors satisfy this property for $p < n$ while local priors would not have this strong model selection consistency. Narisetty and He (2014) provided the strong selection consistency for spike and slab Gaussian priors with the spike prior variance decreasing to zero and the slab

prior variance increasing to infinity as sample size n increases, when the dimension p can be nearly exponentially large in sample size, that is, when $\log p = o(n)$.Narisetty et al. (2019) showed similar consistency for their proposed Skinny Gibbs algorithm under the logistic regression model.

It needs to be noted that the strong model selection consistency is not only stronger in theory than the pairwise consistency but is also practically very helpful. This is because in practice the posterior distribution can only be obtained approximately using MCMC algorithms or other approximation techniques such as the Laplace approximation. In these cases, it is desirable to have a substantial gap between the posterior probability of the true model when compared to the others. Even if the uniform consistency (18) holds true, it is still possible that $P[Z = t \mid Y] \xrightarrow{P} 0$, in which case it would be hard to distinguish the true model from the rest using approximate methods. Therefore, strong model selection consistency is indeed desired for Bayesian variable selection methods.

Along with studying strong selection consistency, Yang et al. (2016) simultaneously studied the algorithmic mixing properties of their proposed stochastic search algorithm in the context of point mass spike priors and slab g-priors. For a randomly generated dataset with Gaussian error distribution, they showed that after $O(nps^2 \log p)$ iterations, the posterior distribution estimated by their stochastic search algorithm is close to the actual posterior distribution corresponding to the prior specification and the true model will be selected with high probability. This is a distinct result in the Bayesian model selection literature as it accounts for the algorithmic approximation and ensures that it only takes nearly linear order computations for finding the the true model. Similar theoretical results for other computational algorithms such as the Skinny Gibbs algorithm are worthy of future study.

9 Implementation

In this section, we discuss some software implementations available in R for performing Bayesian variable selection. This is definitely not an exhaustive list but covers a range of methods developed under different perspectives. In Table 1, we provide a list of R packages and excerpts from the description provided in their package manual as a reference to some of the useful packages.

10 An example

In this section, we provide an illustration of some of the Bayesian methods for variable selection discussed in the chapter. We consider the data obtained from an experiment by Lan et al. (2006) to study the genetics of two inbred mouse populations. The data contain information about gene expression levels of 22,575 genes of 31 female and 29 male mice. The response variable we

TABLE 1 Some R packages for Bayesian computation.

Package	Description (excerpts from the package manual)
BAS	Package for Bayesian Variable Selection and Model Averaging in linear models and generalized linear models using stochastic or deterministic sampling without replacement from posterior distributions. Prior distributions on coefficients are from Zellner's *g*-prior or mixtures of *g*-priors corresponding to the Zellner–Siow Cauchy Priors or the mixture of *g*-priors.
BayesVarSel	Conceived to calculate Bayes factors in linear models and then to provide a formal Bayesian answer to testing and variable selection problems. From a theoretical side, the emphasis in this package is placed on the prior distributions and it allows a wide range of them.
BayesS5	A scalable stochastic search algorithm that is called the Simplified Shotgun Stochastic Search (S5) and aimed at rapidly explore interesting regions of model space and finding the maximum a posteriori (MAP) model.
BMA	Package for Bayesian model averaging and variable selection for linear models, generalized linear models and survival models (cox regression).
BMS	Bayesian model averaging for linear models with a wide choice of (customizable) priors. Built-in priors include coefficient priors (fixed, flexible, and hyper-*g*-priors), five kinds of model priors, moreover model sampling by enumeration or various MCMC approaches.
monomvn	Estimation of multivariate normal and Student-*t* data of arbitrary dimension where the pattern of missing data is monotone. The current version supports maximum likelihood inference and a full Bayesian approach employing scale mixtures for Gibbs sampling.
SAMCpack	We provide generic SAMC samplers for continuous distributions. User-specified densities in R and C++ are both supported. We also provide functions for specific problems that exploit SAMC computation.
Spikeslab	Spike and slab for prediction and variable selection in linear regression models. Uses a generalized elastic net for variable selection.
SSLASSO	Efficient coordinate ascent algorithm for fitting regularization paths for linear models penalized by Spike-and-Slab LASSO of Rockova and George (2018).
varSelectIP	Objective Bayes Variable Selection in Linear Regression and Probit models (Casella et al., 2009; Leon-Novelo et al., 2012).

consider is the phenotype called glycerol-3-phosphate acyltransferase (GPAT) which is measured by quantitative real-time PCR. The dataset is publicly available at GEO (http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330). Zhang et al. (2009), Bondell and Reich (2012), and Narisetty and He (2014) used this data for demonstrating their methods. Following Narisetty and

He (2014), we first perform a screening based on marginal correlation with the response and obtain $p = 200$ predictors (including the intercept and gender variable).

We consider the following Bayesian methods for illustration: (i) the g-prior and (ii) the hyper g-prior methods implemented using Clyde et al. (2010)'s BAS R packageWe, (iii) the mixture of g-priors method implemented using the GibbsBvs function from the R package BayesVarSel (Garcia-Donato and Martinez-Beneito, 2013), (iv) the Bayesian LASSO and (v) the horseshoe priors implemented using the monomvn R package, (vi) the spike and slab LASSO method using the SSLASSO package, and (vii) the BASAD method implemented using the publicly available code on the author's website. In addition, we also implement LASSO and SCAD using the glmnet and ncvreg R packages, respectively.

The models selected by various methods are tabulated in Table 2, which also provides the values of BIC for each of these selected models. Among all the methods considered, BASAD has the smallest BIC and the mixture g-prior takes a very close second smallest value. It is worth noting that the model selected by mixture g-prior is a submodel of the BASAD model. Due to this and given the large gap between the BIC values of models from BASAD and mixture g-prior in comparison with the remaining models, we would recommend selection of the BASAD model in this data example.

TABLE 2 Models selected for PCR data using for different methods.

Method	Model size	BIC	Selected variables
LASSO	10	144.22	24 46 74 101 129 175 176 180 181 191
SCAD	14	145.74	24 46 68 74 90 101 123 129 132 175 176 180 181 191
g-prior	3	139.17	86 90 147
Hyper g-prior	4	134.21	86 90 147 190
Mixture g-prior	4	114.02	90 101 104 123
SSLASSO	5	120.98	31 71 90 118 195
Bayesian LASSO	8	131.32	23 30 41 100 117 122 130 180
Horseshoe	7	126.54	23 30 41 100 122 140 180
BASAD	6	**113.66**	**24 86 90 101 104 123**

BASAD model has the smallest BIC value whereas the model chosen by mixture g-prior also has a very similar BIC with a smaller model. In fact, the model chosen by mixture g-prior is a submodel of the one chosen by BASAD. Bold indicates the best performing method in terms of having the smallest BIC value.

We would like to emphasize that the results from the analysis of this single data set should not be overly generalized. The purpose of this exercise is to give an idea about how these methods can be used in practice and to give a concrete example for the application of these methods.

The Bayesian methods based on *g*-prior, hyper *g*-prior, mixture *g*-prior, and BASAD provide the marginal posterior probabilities for all the covariates. As mixture *g*-prior and BASAD have the smallest BIC values, we provide the plots of their posterior probabilities in Fig. 7. It is interesting to note that the covariates having high posterior probabilities for both the BASAD and Mixture *g*-prior methods are quite similar, which makes sense as these two methods have the lowest BIC values.

The Bayesian methods of Bayesian LASSO and Horseshoe, which are based on continuous shrinkage priors, provide the posterior mean of the regression coefficient along with uncertainty measures in the form of boxplots for each of the regression coefficients. The posterior mean estimates along with their boxplots are provided in Fig. 8. As discussed earlier, the interpretation

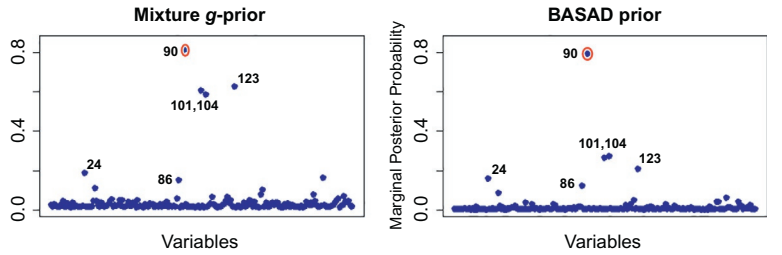

FIG. 7 Marginal Posterior probabilities for Mixture *g*-prior and BASAD. The covariate #90 (circled in *red*) has the highest posterior probability for both the methods.

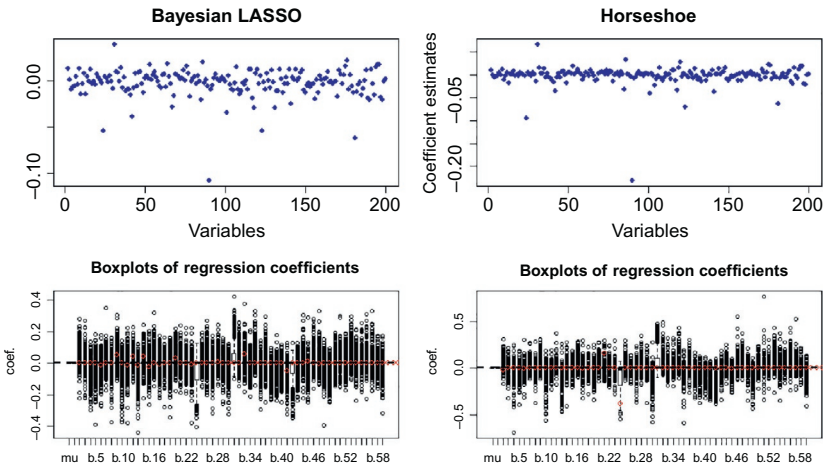

FIG. 8 Coefficient estimation and boxplots for continuous Shrinkage Priors.

of the uncertainty associated with the boxplots is ambiguous in general, but if the priors are viewed as subjective Bayesian priors, the boxplots provide a way to quantify uncertainty of the regression coefficients in the subjective Bayesian sense.

11 Discussion

In the article, we mainly focused on Bayesian variable selection for linear regression models. However, most of the ideas and methods discussed generalize naturally to many other models. There are existing works that extend these approaches to generalized linear models and group selection among others. For instance, Nott and Daniela (2004), Jiang (2007), Wang and George (2007), Chen et al. (2008), Liang et al. (2013), and Narisetty et al. (2019) studied various issues related to Bayesian variable selection in generalized linear models. Bayesian variable selection approaches for settings where the predictors are grouped together are also considered in the recent literature. Xu and Ghosh (2015) and Chen et al. (2016) proposed methods for Bayesian group selection.

It is well known that the least squares regression and the corresponding likelihood are sensitive to outliers. For this reason, least absolute deviation regression and more generally quantile regression (Koenker, 2005; Koenker and Bassett, 1978) is a more robust framework for regression when the errors are likely to be heavy tailed. Quantile regression (Koenker and Bassett, 1978) models the τ-th conditional quantile of the response y_i given the covariates. Unlike the least squares setting, quantile regression is a local model and does not explicitly assume a specific conditional distribution for Y given X. This means that there is no natural likelihood available for quantile regression and necessitates the use of working likelihoods for carrying out Bayesian inference. Yu and Moyeed (2001) proposed a working likelihood based on the Asymmetric Laplace Distribution (ALD). Computation of the posterior with the ALD likelihood is easy to implement using Gibbs sampling (Kozumi and Kobayashi, 2011) or Metropolis–Hastings (MH) algorithms. Yu et al. (2013) proposed spike and slab priors and a Gibbs sampling algorithm for performing Bayesian variable selection for quantile regression with the ALD likelihood. Variable selection with quantile regression has the potential to be much more robust than the more common linear regression while accommodating heterogeneity in the data.

Bayesian shrinkage and regularization has long been used in various contexts demonstrating the flexibility and power of this strategy in various statistical modeling contexts. Beyond statistical models, Bayesian regularization has also seen successful application for machine learning models. For instance, Snoek et al. (2015), Wang and Yeung (2016), and Gal et al. (2017) employed Bayesian regularization for neural networks and deep learning.

Acknowledgments

The author is grateful to two reviewers for their extensive and helpful feedback and graduate students Teng Wu and Ke Li for proofreading an initial version of the article. The author gratefully acknowledges support from NSF Award DMS-1811768.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, pp. 267–281.

Armagan, A., Dunson, D.B., Lee, J., 2013. Generalized double Pareto shrinkage. Stat. Sin. 23, 119–143.

Atchade, Y.A., 2017. On the contraction properties of some high-dimensional quasi-posterior distributions. Ann. Statist. 45 (5), 2248–2273. https://doi.org/10.1214/16-AOS1526.

Barbieri, M.M., Berger, J.O., 2004. Optimal predictive model selection. Ann. Stat. 32, 870–897.

Belloni, A., Chernozhukov, V., 2013. Least squares after model selection in high-dimensional sparse models. Bernoulli 19, 521–547.

Belloni, A., Chernozhukov, V., Hansen, C., 2011. Inference for high-dimensional sparse econometric models. In: Advances in Economics and Econometrics, 10th World Congress of Econometric Society.

Bendel, R.B., Afifi, A.A., 1977. Comparison of stopping rules in forward "stepwise" regression. J. Am. Stat. Assoc. 72, 46–53.

Bertsimas, D., King, A., Mazumder, R., 2016. The adaptive Lasso and its oracle properties. Ann. Stat. 44, 813–852.

Bhadra, A., Datta, J., Polson, N.G., Willard, B., 2017. The horseshoe+ estimator of ultra-sparse signals. Bayesian Anal. 12 (4), 1105–1131.

Bhadra, A., Datta, J., Polson, N.G., Willard, B., 2019. Lasso meets horseshoe: a survey. Stat. Sci. (in press).

Bhattacharya, A., Pati, D., Pillai, N.S., Dunson, D.B., 2015. Dirichlet–Laplace priors for optimal shrinkage. J. Am. Stat. Assoc. 110, 1479–1490.

Bhattacharya, A., Chakraborty, A., Mallick, B., 2016. Fast sampling with Gaussian scale mixture priors in high-dimensional regression. Biometrika 103, 985–991.

Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of Lasso and Dantzig selector. Ann. Stat. 37, 1705–1732.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational inference: a review for statisticians. J. Am. Stat. Assoc. 112 (518), 859–877. https://doi.org/10.1080/01621459.2017.1285773.

Bondell, H.D., Reich, B.J., 2012. Consistent high dimensional Bayesian variable selection via penalized credible regions. J. Am. Stat. Assoc. 107, 1610–1624.

Breheny, P., Huang, J., 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann. Appl. Stat. 5 (1), 232.

Bühlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data. Springer-Verlag, Berlin, Heidelberg.

Carbonetto, P., Stephens, M., 2012. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal. 7, 73–108.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2009a. Handling sparsity via the horseshoe. J. Mach. Learn. Res. 97 (5), 73–80.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2009b. The horseshoe estimator for sparse signals. Biometrika 97, 465–480.

Casella, G., Giron, F.J., Martinez, M.L., Moreno, E., 2009. Consistency of Bayesian procedures for variable selection. Ann. Stat. 37, 1207–1228.

Castillo, I., van der Vaart, A., 2012. Needles and straw in a Haystack: posterior concentration for possibly sparse sequences. Ann. Stat. 40, 2069–2101.

Castillo, I., Schmidt-Hieber, J., van der Vaart, A., 2015. Bayesian linear regression with sparse priors. Ann. Stat. 43, 1986–2018.

Chen, M.H., Huang, L., Ibrahim, J.G., Kim, S., 2008. Bayesian variable selection and computation for generalized linear models with conjugate priors. Bayesian Anal. 3, 585–614.

Chen, R.-B., Chu, C.-H., Yuan, S., Wu, Y.N., 2016. Bayesian sparse group selection. J. Comput. Graph. Stat. 25 (3), 665–683.

Clyde, M., Ghosh, J., Littman, M., 2010. Bayesian adaptive sampling for variable selection and model averaging. J. Comput. Graph. Stat. 20, 80–101.

Datta, J., Ghosh, J.K., 2013. Asymptotic properties of Bayes risk for the Horseshoe prior. Bayesian Anal. 8 (1), 111–132.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B (Methodological) 39, 1–38.

Dicker, L.H., 2016. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. Bernoulli 22 (1), 1–37. https://doi.org/10.3150/14-BEJ609.

Dobriban, E., Wager, S., 2018. High-dimensional asymptotics of prediction: ridge regression and classification. Ann. Statist. 46 (1), 247–279. https://doi.org/10.1214/17-AOS1549.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96, 1348–1360.

Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B 70, 849–911.

Fan, J., Lv, J., 2010. A selective overview of variable selection in high dimensional feature space. Stat. Sin. 20, 101–148.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Ann. Stat. 32, 928–961.

Fan, J., Song, R., 2010. Sure independence screening in generalized linear models with NP-dimensionality. Ann. Stat. 38, 3567–3604.

Fernández, C., Ley, E., Steel, M.F., 2001. Benchmark priors for Bayesian model averaging. J. Econ. 100, 381–427.

Finos, L., Brombin, C., Salmaso, L., 2010. Adjusting stepwise p-values in generalized linear models. Commun. Stat. Theory Methods 39, 1832–1846.

Foster, D.P., George, E.I., 1994. The risk inflation criterion for multiple regression. Ann. Stat. 22, 1947–1975.

Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep Bayesian active learning with image data. In: ICML'17. Proceedings of the 34th International Conference on Machine Learning, vol. 70JMLR.org, pp. 1183–1192. http://dl.acm.org/citation.cfm?id=3305381.3305504.

Gan, L., Narisetty, N.N., Liang, F., 2018. Bayesian regularization for graphical models with unequal shrinkage. J. Am. Stat. Assoc. (in press).

Garcia-Donato, G., Martinez-Beneito, M.A., 2013. On sampling strategies in Bayesian variable selection problems with large model spaces. J. Am. Stat. Assoc. 108, 340–352.

George, E.I., Foster, D.P., 2000. Calibration and empirical Bayes variable selection. Biometrika 87, 731–747.

George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. J. Am. Stat. Assoc. 88, 881–889.

George, E.I., McCulloch, R.E., 1997. Approaches for Bayesian variable selection. Stat. Sin. 7, 339–373.

Ghosal, S., Ghosh, J.K., van der Vaart, A.W., 2000. Convergence rates of posterior distributions. Ann. Stat. 28 (2), 500–531.

Grechanovsky, E., Pinsker, I., 1995. Conditional p-values for the F-statistic in a forward selection procedure. Comput. Stat. Data Anal. 20, 239–263.

Hans, C., Dobra, A., West, M., 2007. Shotgun stochastic search for "large p" regression. J. Am. Stat. Assoc. 102, 507–516.

Hazimeh H. and Mazumder R., Fast best subset selection: coordinate descent and local combinatorial optimization algorithms, arXiv 2018, arXiv:1706.10179.

He, X., Wang, L., Hong, H.G., 2013. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. Ann. Stat. 41, 342–369.

Hsu, D., Kakade, S.M., Zhang, T., 2014. Random design analysis of ridge regression. Found. Comput. Math. 14, 569–600.

Huang X., Wang J. and Liang F., A variational algorithm for Bayesian variable selection, arXiv 2016, arXiv:1602.07640.

Ishwaran, H., Rao, J.S., 2005. Spike and slab variable selection: frequentist and Bayesian strategies. Ann. Stat. 33, 730–773.

Jiang, W., 2007. Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. Ann. Stat. 35, 1487–1511.

Johnson, V.E., Rossell, D., 2012. Bayesian model selection in high-dimensional settings. J. Am. Stat. Assoc. 107, 649–660.

Jordan, M.I., Ghahramani, Z., Jaakkola, T., Saul, L., 1999. Introduction to variational methods for graphical models. Mach. Learn. 37, 183–233.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.

Koenker, R., 2005. Quantile regression. In: Econometric Society Monograph Series. Cambridge University Press. https://doi.org/10.1017/CBO9780511754098.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Kozumi, H., Kobayashi, G., 2011. Gibbs sampling methods for Bayesian quantile regression. J. Stat. Comput. Simul. 81 (11), 1565–1578.

Lan, H., Chen, M., Flowers, J.B., Yandell, B.S., Stapleton, D.S., Mata, C.M., Mui, E.T., Flowers, M.T., Schueler, K.L., Manly, K.F., Williams, R.W., Kendziorski, K., Attie, A.D., 2006. Combined expression trait correlations and expression quantitative trait locus mapping. PLoS Genet. 2, e6.

Leon-Novelo, L., Morenob, E., Casella, G., 2012. Objective Bayes model selection in probit models. J. Am. Stat. Assoc. 31, 353–365.

Liang, F., 2009. Improving SAMC using smoothing methods: theory and applications to Bayesian model selection problems. Ann. Stat. 37, 2626–2654.

Liang, F., Liu, C., Carroll, R.J., 2007. Stochastic approximation in Monte Carlo computation. J. Am. Stat. Assoc. 102, 305–320.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O., 2008. Mixtures of g priors for Bayesian variable selection. J. Am. Stat. Assoc. 103 (481), 410–423.

Liang, F., Song, Q., Yu, K., 2013. Bayesian subset modeling for high dimensional generalized linear models. J. Am. Stat. Assoc. 108, 589–606.

Loh, P.-L., Wainwright, M.J., 2017. Support recovery without incoherence: a case for nonconvex regularization. Ann. Statist. 45 (6), 2455–2482.

Martin, R., Walker, S.G., 2014. Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. Electron. J. Stat. 8, 2188–2206.

Martin, R., Mess, R., Walker, S.G., 2017. Empirical Bayes posterior concentration in sparse high-dimensional linear models. Bernoulli 23, 1822–1847.

Mazumder, R., Friedman, J.H., Hastie, T., 2011. Sparsenet: coordinate descent with nonconvex penalties. J. Am. Stat. Assoc. 106 (495), 1125–1138.

Meinshausen, N., Yu, B., 2009. Lasso-type recovery of sparse representations for high-dimensional data. Ann. Stat. 246–270.

Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression. J. Am. Stat. Assoc. 83, 1023–1032.

Moreno, E., Giron, F.J., Casella, G., 2010. Consistency of objective Bayes factors as the model dimension grows. Ann. Stat. 38, 1937–1952.

Mousavi, A., Maleki, A., Baraniuk, R.G., 2017. Consistent parameter estimation for LASSO and approximate message passing. Ann. Stat. 45, 2427–2454.

Narisetty, N.N., He, X., 2014. Bayesian variable selection with shrinking and diffusing priors. Ann. Stat. 42, 789–817.

Narisetty, N.N., Shen, J., He, X., 2019. Skinny Gibbs: a scalable and consistent Gibbs sampler for model selection. J. Am. Stat. Assoc.

Nott, D.J., Daniela, L., 2004. Sampling schemes for Bayesian variable selection in generalized linear models. J. Comput. Graph. Stat. 13, 362–382.

O'hara, R.B., Sillanpaa, M.J., 2009. A review of Bayesian variable selection methods: what, how and which. Bayesian Anal. 4, 85–117.

Ormerod, J.T., You, C., Muller, S., 2017. A variational Bayes approach to variable selection. Electron. J. Stat. 11, 3549–3594.

Park, T., Casella, G., 2008. The Bayesian LASSO. J. Am. Stat. Assoc. 103, 681–686.

Polson, N.G., Scott, J.G., 2010. Shrink globally, act locally: sparse Bayesian regularization and prediction. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M. et al., (Eds.), Bayesian Statistics 9. Oxford University Press, New York, pp. 501–538.

Raftery, A.E., 1996. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika 83 (2), 251–266.

Rigollet, P., Tsybakov, A.B., 2012. Sparse estimation by exponential weighting. Stat. Sci. 27, 558–575.

Rockova, V., 2018. Bayesian estimation of sparse signals with a continuous spike-and-slab prior. Ann. Stat. 46 (1), 401–437.

Rocková, V., George, E.I., 2014. EMVS: the EM approach to Bayesian variable selection. J. Am. Stat. Assoc. 109 (506), 828–846.

Rockova, V., George, E.I., 2018. The spike-and-slab LASSO. J. Am. Stat. Assoc. 113, 431–444.

Schwarz, G.E., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.

Scott, G.S., Berger, J.O., 2006. An exploration of aspects of Bayesian multiple testing. J. Stat. Plann. Inference 136 (7), 2144–2162.

Scott, G.S., Berger, J.O., 2010. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Ann. Stat. 38, 2587–2619.

Shin, M., Bhattacharya, A., Johnson, V.E., 2018. Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. Stat. Sin. 28, 1053–1078.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M. M. A., Prabhat, P., Adams, R.P., 2015. Scalable Bayesian optimization using deep neural networks. In: ICML'15. Proceedings of the 32nd International Conference on International Conference on Machine Learning, vol. 37JMLR.org, pp. 2171–2180. http://dl.acm.org/citation.cfm?id=3045118.3045349.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288.

van der Vaart, A. W., 1998. Bayes procedures. In: Gill, R., Ripley, B.D., Ross, O.S., Stein, M., Williams, D. (Eds.), Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, pp. 138–152.

van de Geer S. A., 2008. High-dimensional generalized linear models and the Lasso. Ann. Stat. 36, 614–645.

van de Geer, S., 2016. Estimation and Testing Under Sparsity. Lecture Notes in Mathematics Book Series, Springer ISBN 978–3-319–32774-7.

Wang, X., George, E.I., 2007. Adaptive Bayesian criteria in variable selection for generalized linear models. Stat. Sin. 667–690.

Wang, H., Yeung, D.-Y., 2016. Towards Bayesian deep learning: a framework and some existing methods. IEEE Trans. Knowl. Data Eng. 28 (12), 3395–3408. ISSN 1041–4347. https://doi.org/10.1109/TKDE.2016.2606428.

Xu, X., Ghosh, M., 2015. Bayesian variable selection and estimation for group Lasso. Bayesian Anal. 10 (4), 909–936. https://doi.org/10.1214/14-BA929.

Yang, Y., Wainwright, M.J., Jordan, M.I., 2016. On the computational complexity of high-dimensional Bayesian variable selection. Ann. Stat. 44, 2497–2532.

Yu, K., Moyeed, R.A., 2001. Bayesian quantile regression. Stat. Probab. Lett. 54 (4), 437–447.

Yu, K., Chen, C.W.S., Reed, C., Dunson, D.B., 2013. Partial correlation estimation by joint sparse regression models. Stat. Interface 6, 261–274.

Yuan, M., Lin, Y., 2005. Efficient empirical Bayes variable selection and estimation in linear models. J. Am. Stat. Assoc. 100, 1215–1225.

Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Bayesian Inference and Decision Techniques, New York, pp. 233–243.

Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. 38 (2), 894–942.

Zhang, C.-H., Huang, J., 2008. The sparsity and bias of the lasso selection in high-dimensional linear regression. Ann. Stat. 1567–1594.

Zhang, D., Lin, Y., Zhang, M., 2009. Penalized orthogonal-components regression for large P small N data. Electron. J. Stat. 3, 781–796.

Zhao, P., Yu, B., 2006. On model selection consistency of Lasso. J. Mach. Learn. Res. 7, 2541–2563.

Zou, H., 2006. The adaptive Lasso and its oracle properties. J. Am. Stat. Assoc. 101, 1418–1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320.