Journal of Statistical Planning and Inference 207 (2020) 215–229

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Contraction properties of shrinkage priors in logistic regression

Ran Wei [*], Subhashis Ghosal

North Carolina State University, United States of America

ARTICLE INFO

Article history:
Received 14 April 2018
Received in revised form 17 November 2019
Accepted 16 December 2019
Available online 8 January 2020

Keywords:
Bayesian variable selection
Continuous shrinkage
Contraction rate
Logistic regression
Point mass prior

ABSTRACT

Bayesian shrinkage priors have received a lot of attention recently because of their efficiency in computation and accuracy in estimation and variable selection. In this paper, we study the contraction properties of shrinkage priors in a logistic regression model where the number of covariates is high. For a shrinkage prior distribution that is heavy-tailed and concentrated around zero with high probability such as the horseshoe prior, the Dirichlet–Laplace prior, and the normal-gamma prior with appropriate choices of hyper-parameters, estimates of the logistic regression coefficient are shown to asymptotically concentrate around the true sparse vector in the $\mathcal{L}_2$-sense. It is shown that the proposed contraction rate is comparable with the point mass prior that is studied in Atchadé (2017). The simulation study under the logistic regression model verifies the theoretical results by showing that the horseshoe prior and the Dirichlet–Laplace prior perform like the point mass prior for the estimation, variable selection and prediction, and yield much better results than Bayesian lasso and the non-informative normal prior.

1. Introduction

For statistical models such as linear regression, logistic regression or normal means model, high-dimensional data analysis is challenging due to the computational burden and inherent limitations of any procedure under the curse of dimensionality. In these models, it is essential to impose a lower-dimensional structure, such as sparsity, which leads to the problem of variable selection in a regression model. In the non-Bayesian approaches, penalization procedures such as the LASSO is popularly used. Bayesian variable selection procedures are generally more informative than penalization methods because they automatically address the model selection uncertainty. There is an extensive literature on Bayesian variable selection methods in recent years. Since the value zero of a regression coefficient is equivalent to having the corresponding predictor dropping out of the model, priors are designed to give special emphasis to the value zero. For example, a point-mass prior combines a probability mass at zero and a non-zero continuous distribution. Computation is generally carried out by Reversible Jump Markov Chain Monte Carlo methods (Green, 1995). A spike-and-slab prior (Ishwaran and Rao, 2005) is typically a mixture of two normal distributions with one highly concentrated at around zero. The Stochastic Search Variable Selection (SSVS) method (George and McCulloch, 1993) can be used to compute the posterior distribution corresponding to a spike-and-slab prior. Due to high computational costs, these methods are not scalable to very high dimensional situations commonly arising in recent applications.

[*] Corresponding author.
E-mail address: rwei@ncsu.edu (R. Wei).

https://doi.org/10.1016/j.jspi.2019.12.004
0378-3758/© 2020 Elsevier B.V. All rights reserved.

To address the limitations of these methods mentioned above under high dimensional problems, Bayesian LASSO (Park and Casella, 2008) uses a double exponential prior distribution on the coefficients. Clearly, the posterior mode of the Bayesian LASSO is the LASSO estimator in linear regression model, and the posterior can be computed by a simple Gibbs sampling procedure. However, the Bayesian LASSO does not make the posterior distribution concentrate near the true value in large samples (Castillo et al., 2015), although its mode the LASSO estimator has good estimation and variable selection properties under appropriate conditions. In recent years, a variety of continuous prior densities with good shrinkage properties have been introduced in the literature, such as the horseshoe prior (Carvalho et al., 2010), normal-gamma prior (Griffin and Brown, 2010), double-Pareto prior (Armagan et al., 2013a), Dirichlet–Laplace (DL) prior (Bhattacharya et al., 2015) and the horseshoe+ prior (Bhadra et al., 2017). Unlike a spike-and-slab prior, these priors have one single component like the double exponential prior, but has much higher concentration near zero, and have typically thicker tails, so that they mimic a point-mass prior. Defined as a global–local scale mixture of Gaussian distribution, shrinkage priors give computationally-efficient alternatives to point mass prior. These priors go by the names continuous shrinkage priors, or one-component priors or global–local priors. Recent contributions in the literature show their promising posterior concentration properties near the true value and ability to identify true non-zero coefficients. It may be noted that under such priors, the posterior probability of hitting the exact value zero is always zero, so for variable selection, some appropriate thresholding procedure needs to accompany the Bayesian procedure.

Earliest posterior concentration results in models of dimension increasing to infinity with the sample size are provided by Ghosal (1997, 1999, 2000), respectively for generalized linear models, regression models and exponential families. In his results, no sparsity conditions are assumed on the truth and posterior concentration at the truth and asymptotic normality of the posterior are established, provided that the growth of the dimension is sufficiently slow compared with the sample size. In more high dimensional settings, assuming sparsity conditions, Jiang (2007) first studied posterior contraction under the Hellinger distance. Castillo and van der Vaart (2012) and Belitser and Nurushev (2020) established posterior concentration and variable selection properties for certain point-mass priors in the many normal means model. The latter paper also established asymptotic coverage from frequentist perspective. Posterior concentration and variable selection in high dimensional linear models are obtained by Castillo et al. (2015), Martin et al. (2017) and Belitser and Ghosal (2019) for certain point-mass priors. The last one also showed that some suitable empirical Bayes Bayesian credible regions with optimal size for any sparsity level have adequate frequentist coverage under an "excessive bias restriction" condition, generalizing the result of Belitser and Nurushev (2020) from sparse normal mean setting to linear regression. Recent theoretical breakthroughs establish concentration properties of posterior distributions of continuous shrinkage priors. Armagan et al. (2013b) showed posterior consistency in a linear regression model with shrinkage priors for a low-dimensional setting where the number of covariates does not exceed the number of observations. Furthermore, Van Der Pas et al. (2014) showed that the posterior based on the horseshoe prior concentrates at the optimal rate for the many normal-means problem. Bhattacharya et al. (2015) obtained an analogous result using the DL prior. Under appropriate choice of the hyper-parameter in DL prior, the posterior contraction property is applied to the coefficients of high-dimension linear regression model. Song and Liang (2017) considered a general class of continuous shrinkage priors and obtained posterior contraction rate in linear regression models depending on concentration and tail properties of the density of the continuous shrinkage prior. Essentially their conclusion may be summarized as the following statement: under appropriate conditions, the posterior contraction rates and variable selection ability of continuous shrinkage priors are close to those of the point-mass priors in high dimensional linear regression models.

Compared to the papers in the literature which address convergence results on Bayesian variable selection in a linear regression model, similar literature focusing on contraction properties in generalized linear models such as logistic regression are limited. Shen and Ghosal (2016) considered estimating conditional density in a high dimensional setting using tensor products of B-splines where the true conditional density is assumed to be a function of only a few predictors. They showed that the oracle contraction rate can be matched adaptively up to a logarithmic factor. A similar result by using Dirichlet process mixtures was obtained by Norets and Pati (2017). Yang and Tokdar (2015) and Belitser and Ghosal (2019) obtained optimal posterior contraction rate in the setting of high dimensional additive regression models using a Gaussian process prior. The logistic regression model is another important model for practical applications and often comes with a large of predictors, among which the important ones need to be selected for precise estimation and sensible interpretation. Atchadé (2017), as a special case of his more general results on quasi-posterior distributions in possibly nonlinear models, derived posterior contraction properties in a logistic regression model using a point-mass prior. They showed that the posterior contracts at the rate $\sqrt{s_*(\log p)/n}$, where s_* is the true number of active predictors. However, the properties of posterior distributions under shrinkage priors in the logistic regression model has not been studied in the literature. Borrowing a few techniques from Song and Liang (2017), we extend the results of Atchadé (2017) on logistic regression to continuous shrinkage priors. Our finding can be summarized to the statement that, provided that a prior has a sufficient concentration near zero and has sufficiently thick tails, posterior concentrates near the true vector of coefficients at a described rate and with high posterior probability, only selects (effectively) sparse vectors like a point-mass prior. The theoretical result is amply supported by simulations.

The remainder of the paper is organized as follows. Section 2 presents the logistic regression model and the assumptions on shrinkage prior densities. The main results on the posterior contraction of the parameters being estimated are shown in Theorem 2.1. Several examples of shrinkage priors and continuous spike-and-slab prior are introduced in Section 3 to demonstrate the conditions for posterior contraction. For different type Bayesian variable selection technique, Section 4 evaluates the performance in recovering the logistic coefficients, identifying non-zero subsets of covariates and predicting on test design matrix. The proof of the main posterior contraction result is given in Appendix.

2. Contraction results in logistic regression

For two positive sequences a and b, the relation $a \prec b$ means $\lim a/b = 0$, $a \gtrsim b$ means a/b is bounded below, and $a \asymp b$ stands for $m_1 < \liminf a/b \le \limsup a/b < m_2$ for some positive constants m_1 and m_2.

We study the concentration properties of the posterior distribution obtained from the continuous shrinkage prior in logistic regression model. Suppose that $Z \overset{\text{def}}{=} (Z_1, \dots, Z_n)^T$ are independent binary random variables and $X_{n\times p} = (X_1, \dots, X_n)^T$ is the matrix of the predictor variables for the sample size n and the number of predictors p. We consider the logistic regression model

$$\mathrm{P}(Z_i = 1|X_i) = \frac{\exp(X_i^T\beta)}{1+\exp(X_i^T\beta)}, \tag{1}$$

for $i = 1, \dots, n$, where $\beta \in \mathbb{R}^p$ is the vector of regression coefficients. The likelihood function in the logistic regression model is written as

$$q_{n,\beta}(Z) = \exp\left[\sum_{i=1}^{n}\big(Z_i \cdot X_i^T\beta - g(X_i^T\beta)\big)\right], \tag{2}$$

where $g(x) \overset{\text{def}}{=} \log(1+e^x)$ for $x \in \mathbb{R}$. Let β_* be the true values of the coefficients of logistic regression. We denote the number of non-zero elements in β_* by $s_* \overset{\text{def}}{=} \|\beta_*\|_0$ and the set of non-zero elements indexes as $\xi_* = \{j : \beta_{*j} \ne 0\}$. Let E stand for a bound on the maximum absolute value of β_{*j}: $\max_{j\in\xi_*}|\beta_{*j}| < E$, where E may depend on n but is assumed to be non-decreasing without loss of generality.

On the p-dimensional coefficients β, we put a sequence of prior distributions, denoted as $\Pi_\alpha(\beta)$, where α is the hyper-parameter in the prior density. The prior density on β is chosen to be of the product form: $\Pi_\alpha(\beta) \overset{\text{def}}{=} \prod_{j=1}^{p}\pi_\alpha(\beta_j)$. For a sequence $\bar\epsilon_n \downarrow 0$ satisfying $n\bar\epsilon_n^2 \to \infty$ to be chosen later, we assume the following conditions on the prior density $\pi_\alpha(\cdot)$ for each entry of the coefficient vector:

$$1 - \int_{-a_n}^{a_n}\pi_\alpha(\beta)d\beta \le p^{-(1+\mu)} \text{ for some constant } \mu > 0, \tag{3}$$

$$-\log\left(\int_{\sum_{j=1}^{p-s_*}|\beta_j|\le\eta\bar\epsilon_n}\prod_{j=1}^{p-s_*}\pi_\alpha(\beta_j)d\beta_1\cdots d\beta_{p-s_*}\right) \prec n\bar\epsilon_n^2, \tag{4}$$

$$\log s_* + 2\log(1/\bar\epsilon_n) - \log\left(\inf_{\beta\in[-E,E]}\pi_\alpha(\beta)\right) \prec n\bar\epsilon_n^2/s_*, \tag{5}$$

for a small constant $\eta > 0$ and threshold value a_n satisfying the restriction $a_n < \bar\epsilon_n/p$.

Above, (3) gives sufficient prior concentration near the value zero, allowing the posterior to pick up sparsity in the model. The second condition is similar, implying that the total contribution of near-zero coefficients to be well-controlled. This is needed only because the absence of point mass at zero, since otherwise the required condition will be automatically satisfied. The third condition says that the tail of the prior is not too sharp, so that there is some minimum prior concentration near large non-zero values. Since such a condition cannot hold uniformly over the entire parameter space which is unbounded, a restriction on the signal sizes is imposed for recovery at the stated rate. If uniformly is sought over a bigger space (i.e., E is larger), then the contraction rate can slow down.

As we will see, the sequence $\bar\epsilon_n$ stands for the pre-rate quantifying prior concentration around the true coefficient, in that the prior around an $\bar\epsilon_n$-neighborhood of the truth is at least as much as $e^{-cn\bar\epsilon_n^2}$ for some $c > 0$. Note that, from the last condition, $n\bar\epsilon_n^2 \ge s_*$.

Let $L_n > 0$ be a sequence such that

$$-\log\left(1 - \int_{-L_n}^{L_n}\pi_\alpha(\beta)d\beta\right) \prec n\bar\epsilon_n^2. \tag{6}$$

This condition allows ignoring large values of the coefficients too unlikely under the prior, helping to control entropy bounds.

For simplicity of notation, we normalize the predictor to be bounded by 1, $\|X\|_\infty \overset{\text{def}}{=} \max_{i,j}|x_{ij}| \le 1$. We also define the following quantity for $s \in \{1, \dots, p\}$:

$$\kappa_{1,a_n}(s) \overset{\text{def}}{=} \inf\left\{\frac{\beta^T(X^TWX)\beta}{n\|\beta\|_2^2} : 1 \le \#\{j : |\beta_j| > a_n\} \le s\right\} \tag{7}$$

for $W \in \mathbb{R}^{n\times n}$ as the diagonal matrix with the ith diagonal entry given by $W_i = g^{(2)}(X_i^T\beta)$. Since the prior on β assigns zero probability at the point zero, the exact number of nonzero elements of β is always p. A meaningful comparison with the value s_* is made by considering s the number of elements of β exceeding in absolute value the threshold a_n. In other words, only elements with absolute value larger than a_n will be treated as significant and counted towards non-zero entries. Positivity of $\kappa_{1,a_n}(s)$ large enough s, which will be implicit in the assumption in the theorem on recovery, implies identifiability. The latter is clearly necessary for recovery.

We introduce the function

$$\mathcal{L}_{n,\beta}(Z) = \log q_{n,\beta}(Z) - \log q_{n,\beta_*}(Z) - \triangledown \log q_{n,\beta_*}(Z)^T(\beta - \beta_*). \tag{8}$$

This function plays a key role in statistical inference as it quantifies the curvature of the objective function $\beta \mapsto \log q_{n,\beta}(Z)$ around β_*.

The following theorem shows a posterior concentration property on the coefficients of the logistic regression model with shrinkage priors $\Pi_\alpha(\beta) = \prod_{j=1}^p \pi_\alpha(\beta_j)$ satisfying conditions (3)–(5). The proof of the theorem is given in Appendix.

Theorem 2.1. *Consider the logistic regression model defined in (1), where the vector of coefficients β is given a prior distribution with density $\Pi_\alpha(\beta)$ satisfying conditions (3)–(6). Let $\bar{s} = \max\{s_*, Ln\bar{\epsilon}_n^2/\log p\}$ for some sufficiently large $L > 0$, $\log L_n = O(\log p)$ and assume that $\kappa_{1,a_n}(\bar{s})$ is bounded away from zero with $a_n < \bar{\epsilon}_n/p$ satisfying (3). Then with $\epsilon_n = \sqrt{(\bar{s}\log p)/n} \asymp \bar{\epsilon}_n$ and a sufficiently large constant $M > 0$ and some constant $c_0 > 0$, for the events*

$$B_n = \{\textit{At least } \bar{s} \textit{ entries of } |\beta| \textit{ is greater than } a_n\}$$

and $C_n = \{\|\beta - \beta_*\|_2 > M\epsilon_n\}$, we have that

$$P^*\left(\mathrm{P}_\alpha(C_n \cup B_n|Z, X) > 2e^{-c_0 n\bar{\epsilon}_n^2}\right) \le 7/p \to 0, \tag{9}$$

as $p = p_n \to \infty$.

It may be noted that under the assumed setting, the rate ϵ_n coincides with the pre-rate $\bar{\epsilon}_n$, although the latter has to be introduced before to define $\bar{s}$. In particular, the theorem implies that the posterior selects at most $\bar{s}$ significantly non-zero coefficients and concentrates near the truth at the rate ϵ_n. The $\mathcal{L}_2$-norm contraction rate ϵ_n is given by the maximum of the contraction rate $\sqrt{(s_*\log p)/n}$ in linear regression model (Song and Liang, 2017). Clearly, this is the best possible rate derivable from the theorem as $n\epsilon_n^2 \ge s_*\log p$ and $\epsilon_n \ge \bar{\epsilon}_n$. If the prior is such that the prior concentration near zero for the sum of absolute values is not too low as measured by (4) with $\bar{\epsilon}_n = \sqrt{(s_*\log p)/n}$ and the individual prior concentration near true non-zero values are not too low as measured by (5), then $\bar{s}$ if of the order s_*, and hence the optimal rate $\sqrt{(s_*\log p)/n}$ is obtained. Note that, for (5) to hold for $\bar{\epsilon}_n = \sqrt{(s_*\log p)/n}$, the right side becomes $\log p$, so a polynomially thick tail for the density π_α will suffice.

3. Examples of shrinkage priors

The conditions imposed on the prior through (3)–(5) are somewhat abstract. We now give examples of prior densities which satisfy the required conditions and hence lead to posterior concentration near the truth.

3.1. Spike-and-slab prior

For the Spike-and-slab prior that continuously shrinks the vector of coefficients, we assume that the prior density of $\beta_j, j = 1, \dots, p$ has the following form:

$$\beta_j|\rho_j, \sigma_0^2, \sigma_1^2 \sim (1-\rho_j)\mathrm{N}(0, \sigma_0^2) + \rho_j\mathrm{N}(0, \sigma_1^2), \quad \rho_j|b \sim \mathrm{Bin}(1, b), \tag{10}$$

where $\mathrm{N}(\mu, \sigma^2)$ stands for the normal distribution with mean μ and variance σ^2, and Bin for the binomial distribution. After integrating out ρ_j, the prior density for $\beta_j, j = 1, \dots, p$, can be written as $\pi_\alpha(\beta_j) = (1-b)\phi(\beta_j; 0, \sigma_0^2) + b\phi(\beta_j; 0, \sigma_1^2)$, where $\phi(\cdot; \mu, \sigma^2)$ denotes the density for normal distribution with mean μ and variance σ^2. Here $(b, \sigma_0^2, \sigma_1^2)$ plays the role of the hyper-parameter α. The prior approximates a point-mass prior if the value of σ_0^2 is very small. The following theorem presents the posterior concentration of β in logistic regression model with spike-and-slab prior.

Theorem 3.1. *Consider the continuous spike-and-slab prior in (10) on the coefficients β of logistic regression model (1). If $\log\sigma_1 + E^2/2\sigma_1^2 \prec n\bar{\epsilon}_n^2/s_*$, $\sigma_0 \le a_n/\sqrt{(1+\mu')\log p} < \bar{\epsilon}_n/\big(p\sqrt{(1+\mu')\log p}\big)$ and $b = p^{-(1+\mu')}$ for a constant $\mu' > 0$, then the posterior concentration property in (9) holds.*

In particular, the optimal rate $\epsilon_n = \sqrt{(s_*\log p)/n}$ holds if $\max_{j\in\xi_*}|\beta_{*j}| < E$ with E and σ_0, σ_1 satisfying $\log\sigma_1 + E^2/2\sigma_1^2 \prec \log p$, $\sigma_0 \le 1/p\sqrt{n}$ and $b = p^{-(1+\mu')}$ for some $\mu' > 0$.

Proof. In order to show the posterior concentration results with the defined contraction rate ϵ_n, we need to show that the conditions (3)–(5) on the prior density hold for the spike-and-slab prior. First for $a_n < \epsilon_n/p$,

$$\begin{aligned} 1-\int_{-a_n}^{a_n}\pi_\alpha(\beta)d\beta =&2\int_{a_n}^{\infty}\left[(1-b)\phi(\beta;0,\sigma_0^2)+b\phi(\beta;0,\sigma_1^2)\right]d\beta \\ =&2(1-b)\int_{a_n}^{\infty}\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\beta^2/2\sigma_0^2}d\beta+2b\int_{a_n}^{\infty}\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\beta^2/2\sigma_1^2}d\beta \\ \leq&\sqrt{2}\exp\{-a_n^2/2\sigma_0^2\}(a_n\sqrt{\pi}/\sigma_0)^{-1}+b. \end{aligned} \tag{11}$$

Since $b = p^{-(1+\mu')}$ and $a_n/\sigma_0 \geq \sqrt{(1+\mu')\log p}$ for constant $\mu' > 0$, we can show that for any constant $0 < \mu < \mu'$, $1-\int_{-a_n}^{a_n}\pi_\alpha(\beta)d\beta \leq p^{-(1+\mu)}$. Therefore, (3) holds. Next note that for a small constant $\eta > 0$,

$$\begin{aligned} &\int_{\sum_{j=1}^{p-s_*}|\beta_j|\leq\eta\epsilon_n}\prod_{j=1}^{p-s_*}\left[(1-b)\phi(\beta_j;0,\sigma_0^2)+b\phi(\beta_j;0,\sigma_1^2)\right]d\beta_1\cdots d\beta_{p-s_*} \\ &\qquad\geq\left(1-p^{-(1+\mu')}\right)^{p-s_*}\mathrm{P}_\alpha\left(\sum_{j=1}^{p-s_*}|T_j|\leq\eta\epsilon_n\right) \end{aligned} \tag{12}$$

where $T_j, j = 1,\ldots,p-s_*$, are independently and identically distributed as $\mathrm{N}(0,\sigma_0^2)$. Since $\mathrm{E}_\alpha\left(|T_j|\right) = \sigma_0\sqrt{2/\pi}$, we have $\mathrm{E}_\alpha\left(\sum_{j=1}^{p-s_*}|T_j|\right) = (p-s_*)\sigma_0\sqrt{2/\pi}$. Therefore by the central limit theorem, for $\sigma_0 < \eta\bar{\epsilon}_n/p$, we have $\mathrm{P}_\alpha\left(\sum_{j=1}^{p-s_*}|T_j|\leq\eta\bar{\epsilon}_n\right)\geq \mathrm{P}_\alpha\left(\sum_{j=1}^{p-s_*}|T_j|\leq(p-s_*)\sigma_0\sqrt{2/\pi}\right)\to 1/2$. Thus for a constant $c'' > 0$, we have

$$\int_{\sum_{j=1}^{p-s_*}|\beta_j|\leq\eta\bar{\epsilon}_n}\prod_{j=1}^{p-s_*}\pi_\alpha(\beta_j)d\beta_1\cdots d\beta_{p-s_*}\geq\frac{1}{4}\left(1-p^{-(1+\mu')}\right)^p\geq\exp(-c''n\bar{\epsilon}_n^2). \tag{13}$$

Lastly for $E > 0$ nondecreasing in n, we obtain

$$\begin{aligned} &-\log\left(\inf_{\beta\in[-E,E]}\left[(1-b)\phi(\beta;0,\sigma_0^2)+b\phi(\beta;0,\sigma_1^2)\right]\right) \\ &\quad\leq-\log\left(p^{-(1+\mu')}\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\{-E^2/2\sigma_1^2\}\right) \\ &\quad=\frac{1}{2}\log(2\pi)+\log\sigma_1+(1+\mu')\log p+E^2/2\sigma_1^2\prec n\epsilon_n^2/s_* \end{aligned} \tag{14}$$

since $\log\sigma_1+E^2/2\sigma_1^2\prec n\epsilon_n^2/s_*$. Then (5) holds. Now Theorem 2.1 can be applied.

To obtain L_n, note that the tail of the prior is at most like $\mathrm{N}(0,\sigma_1^2)$, so $L_n = \sqrt{Cn\bar{\epsilon}_n^2}$ would satisfy (6) and clearly $\log L_n = O(\log p)$. Thus the posterior contraction rate is $\epsilon_n = \sqrt{(s_*\log p)/n}$.

3.2. Dirichlet-Laplace prior

In Bhattacharya et al. (2015), a global–local shrinkage prior is proposed on β. Specifically, the prior distribution has the following structure:

$$\beta_j|\phi_j,\tau\sim\mathrm{DE}(\phi_j\tau);\ (\phi_1,\ldots,\phi_p)|\alpha\sim\mathrm{Dirichlet}(\alpha,\ldots,\alpha);\ \tau|\alpha\sim\mathrm{Gamma}(p\alpha,2), \tag{15}$$

where $\mathrm{DE}(\lambda)$ stands for double exponential distribution with scale λ and $\mathrm{Gamma}(a,b)$ for the gamma distribution with shape parameter a and scale parameter b. The hyper-parameter α controls the level of shrinkage such that smaller α means more shrinkage and sparser model. For convenience, we equivalently represent the prior distribution through the following hierarchical structure:

$$\beta_j|\psi_j\sim\mathrm{DE}(\psi_j),\ \ \psi_j|\alpha\sim\mathrm{Gamma}(\alpha,2),\ j=1,\ldots,p. \tag{16}$$

The following theorem shows the consistency of β in the logistic regression model with the Dirichlet–Laplace prior.

Theorem 3.2. *Consider the Dirichlet–Laplace prior in (16) for the logistic regression model. Assume that the true regression coefficients satisfy the condition that* $\max_{j\in\xi_*}|\beta_{*j}| < E$ *with* $E\prec n\bar{\epsilon}_n^2/s_*$. *If* $\alpha\prec\min\{\bar{\epsilon}_n/p, p^{-(1+\nu)}\}$ *for some constant* $\nu > 0$, *then the posterior concentration property in (9) holds.*

In particular, if $E\prec\log p$ *and* $\alpha\prec p^{-1}\min\{n^{-1/2},p^{-\nu}\}$ *for some* $\nu > 0$, *then the optimal rate* $\epsilon_n = \sqrt{(s_*\log p)/n}$ *is obtained.*

Proof. We need to verify the prior conditions (3)–(5) hold for the Dirichlet–Laplace prior density π_α. First, for $a_n < \bar{\epsilon}_n/p$, following the proof of Lemma 3.3 in Bhattacharya et al. (2015), we obtain

$$\begin{aligned}1-\int_{-a_n}^{a_n}\pi_\alpha(\beta)d\beta &= \int_0^\infty e^{-a_n/\psi}\frac{1}{\Gamma(\alpha)2^\alpha}\psi^{\alpha-1}e^{-\psi/2}d\psi\\ &= \frac{1}{\Gamma(\alpha)2^\alpha}\left[\int_0^{4a_n}\psi^{\alpha-1}e^{-a_n/\psi-\psi/2}d\psi+\int_{4a_n}^\infty\psi^{\alpha-1}e^{-a_n/\psi-\psi/2}d\psi\right]\\ &\le \frac{1}{\Gamma(\alpha)2^\alpha}\left[C+\int_{4a_n}^\infty\psi^{-1}e^{-\psi/2}d\psi\right],\end{aligned} \tag{17}$$

where $C>0$ is a constant independent of a_n. Since $\int_{4a_n}^\infty\psi^{-1}e^{-\psi/2}d\psi \le \int_{2a_n}^\infty t^{-1}e^{-t}dt \le -\log a_n$, we conclude that for $a_n<\epsilon_n/p$ and a constant $C'>0$ free of a_n,

$$1-\int_{-a_n}^{a_n}\pi_\alpha(\beta)d\beta \le C'\alpha\log(1/a_n) = C'p^{-(1+\nu)}\log(1/a_n)\le p^{-(1+\mu)}, \tag{18}$$

where $0<\mu<\nu$ is a constant. Next, $\beta_j, j=1,\dots,p-s_*$, are independently and identically distributed as in (16), so $\mathrm{E}_\alpha\left(|\beta_j|\right)=\mathrm{E}_\alpha\left(\mathrm{E}\left(\beta_j|\psi_j\right)\right)=2\alpha$. Then, by the central limit theorem, $\mathrm{P}_\alpha\left(\sum_{j=1}^{p-s_*}|\beta_j|\le\eta\epsilon_n\right)\ge\mathrm{P}_\alpha\left(\sum_{j=1}^{p-s_*}|\beta_j|\le 2(p-s_*)\alpha\right)\to 1/2$ since $p\alpha\prec\epsilon_n$. Thus the second condition (4) on the prior density holds. Finally

$$\inf_{\beta\in[-E,E]}\pi_\alpha(\beta)=\inf_{\beta\in[-E,E]}\int_0^\infty\frac{1}{\psi}e^{-|\beta|/\psi}\frac{1}{\Gamma(\alpha)2^\alpha}\psi^{\alpha-1}e^{-\psi/2}d\psi,$$

which can be bounded below by

$$\frac{1}{\Gamma(\alpha)2^\alpha}\int_0^{\eta'}\psi^{\alpha-2}e^{-E/\psi-\psi/2}d\psi\ge\frac{C''}{\Gamma(\alpha)2^\alpha}\int_0^{\eta'}\psi^{-2}e^{-E/\psi}d\psi\ge\frac{C''}{\Gamma(\alpha)}\int_{\eta'^{-1}}^\infty e^{-Et}dt=\frac{C''}{\Gamma(\alpha)}\frac{1}{E}e^{-E/\eta'},$$

where $\eta'>0$ is a small constant and C'' is a constant independent of α. Therefore, for $E\prec n\bar{\epsilon}_n^2/s_*$ and $\log p\prec n\bar{\epsilon}_n^2/s_*$, we have $-\log\bigl(\inf_{\beta\in[-E,E]}\pi_\alpha(\beta)\bigr)\le-\log C''+(1+\nu)\log p+E/\eta'+\log E\prec n\bar{\epsilon}_n^2/s_*$. This completes the proof of (5).

For (6), proceeding as in (18), for any sequence $b_n>1$, we can write

$$\begin{aligned}1-\int_{-L_n}^{L_n}\pi_\alpha(\beta)d\beta &= \frac{1}{2^\alpha\Gamma(\alpha)}\left\{\int_0^{b_n}\psi^{\alpha-1}e^{-L_n/\psi-\psi/2}d\psi+\int_{b_n}^\infty\psi^{\alpha-1}e^{-L_n/\psi-\psi/2}d\psi\right\}\\ &\le \frac{1}{2^\alpha\Gamma(\alpha)}\left\{e^{-L_n/b_n}\int_0^{b_n}\psi^{\alpha-1}+\int_{b_n}\psi^{\alpha-1}e^{-\psi/2}d\psi\right\}\\ &\le \left\{b_n^\alpha e^{-L_n/b_n}+e^{-b_n/3}\right\},\end{aligned}$$

Therefore if $n\bar{\epsilon}_n^2\prec b_n$ and $L_n=b_n^2$, then the condition (6) holds, as well as $\log L_n=O(\log p)$.

3.3. Horseshoe prior

Another shrinkage prior we consider is the horseshoe prior introduced in Carvalho et al. (2009). The horseshoe prior assumes that each β_j is conditionally independent with density $\pi_\alpha(\beta_j)$ having the following structure:

$$\beta_j|\lambda,\alpha\sim\mathrm{N}(0,\lambda_j^2\alpha^2),\quad\lambda_j\sim\mathrm{C}^+(0,1), \tag{19}$$

where $\mathrm{C}^+(0,1)$ is the half-Cauchy distribution, that is, the distribution of the absolute value of a standard Cauchy distribution, and the hyperparameter $\alpha>0$ controls the global shrinkage. Therefore, we have the density function of β_j: $\pi_\alpha(\beta_j)=\int_0^\infty\phi(\beta_j;0,\lambda^2\alpha^2)f_{\mathrm{C}^+}(\lambda)d\lambda$, where $f_{\mathrm{C}^+}(\lambda)=(2/\pi)(1+\lambda^2)^{-1}$ stands for the density of the half-Cauchy distribution.

Theorem 3.3. *Consider the horseshoe prior (19) on β in a logistic regression model (1). Assume that the true regression coefficients satisfy the condition that $\max_{j\in\xi_*}|\beta_{*j}|<E$. If $\alpha<p^{-(2+\mu'')}/\sqrt{n}$ for a constant $\mu''>0$, and $\max\{\log E,-\log\alpha\}=O(\log p)$, then the posterior concentration property in (9) holds for the optimal rate $\epsilon_n=\sqrt{(s_*\log p)/n}$.*

Proof. In order to prove the posterior concentration, we need to show that under the stated conditions of the theorem, relations (3)–(5) hold. First, for a constant $\mu''>0$, and some $a_n<\bar{\epsilon}_n/p$,

$$\begin{aligned}&1-\int_{-a_n}^{a_n}\pi_\alpha(\beta)d\beta\\ &=2\int_0^\infty\left(1-\Phi\left(\frac{a_n}{\lambda\alpha}\right)\right)f_{\mathrm{C}^+}(\lambda)d\lambda\end{aligned}$$

$$
\begin{aligned}
&= 2\int_0^{p^{1+\mu''}} \left(1-\Phi\left(\frac{a_n}{\lambda\alpha}\right)\right) f_{C^+}(\lambda)d\lambda + 2\int_{p^{1+\mu''}}^{\infty} \left(1-\Phi\left(\frac{a_n}{\lambda\alpha}\right)\right) f_{C^+}(\lambda)d\lambda \\
&\leq 2\left(1-\Phi\left(\frac{a_n}{\alpha}\cdot p^{-(1+\mu'')}\right)\right) + \int_{p^{1+\mu''}}^{\infty} \frac{2}{\pi}\frac{1}{1+\lambda^2}d\lambda \\
&\leq \frac{\sqrt{2/\pi}\, e^{-a_n^2 p^{-2(1+\mu'')}/\alpha^2}}{a_n p^{-(1+\mu'')}/\alpha} + p^{-(1+\mu'')} \leq p^{-(1+\mu)}
\end{aligned}
\tag{20}
$$

for any $0 < \mu < \mu''$, given that $\alpha \leq p^{-(2+\mu'')}/\sqrt{n}$ and

$$
\int_{p^{1+\mu''}}^{\infty} \frac{2}{\pi}\frac{1}{1+\lambda^2}d\lambda = \frac{2}{\pi}\left(\frac{\pi}{2} - \tan^{-1}(p^{1+\mu''})\right) \asymp p^{-(1+\mu'')}.
$$

Since $a_n < \epsilon_n/p$, there exists $0 < \bar{\mu}'' < \mu''$, such that $\dfrac{\eta\epsilon_n/(p-s_*)}{T_p^2\alpha^2} > (1+\bar{\mu}'')\log p$, for $T_p = p^{1+\mu''}$. By the proof of (20), we have $\int_{-\eta\epsilon_n/(p-s_*)}^{\eta\epsilon_n/(p-s_*)} \pi_\alpha(\beta)d\beta \geq 1 - p^{-(1+\bar{\mu})}$ for any constant $0 < \bar{\mu} < \bar{\mu}'' < \mu''$. For each β_j following the horseshoe prior (19),

$$
\int_{\sum_{j=1}^{p-s_*}|\beta_j|\leq\eta\epsilon_n} \prod_{j=1}^{p-s_*} \pi_\alpha(\beta_j)d\beta_1\cdots d\beta_{p-s_*} \geq \left[\mathrm{P}(|\beta_j| \leq \eta\epsilon_n/(p-s_*))\right]^{p-s_*},
$$

which is bounded below by $\left(1-p^{-(1+\mu)}\right)^{p-s_*} \to 1$. That verifies the condition (4).

For (5), observe that by the bound (2) of Carvalho et al. (2010), $\inf_{\beta\in[-E,E]}\pi_\alpha(\beta)$ is

$$
\int_0^\infty \frac{1}{\sqrt{2\pi}\lambda\alpha} e^{-E^2/2\lambda^2\alpha^2} f_{C^+}(\lambda)d\lambda \geq \frac{1}{\sqrt{2\pi^3}}\log(1+4\alpha^2/E) \asymp \frac{\alpha^2}{E},
$$

which is bounded by the order $p^{-(1+\nu)}$. This gives the desired assertion $-\log\left(\inf_{\beta\in[-E,E]}\pi_\alpha(\beta)\right) \lesssim \log p \prec n\bar{\epsilon}_n^2/s_*$ for any sequence $\bar{\epsilon}_n \succ \sqrt{(s_*\log p)/n}$.

Finally, as argued in (20), we have for sequences $L_n > 0$ and $b_n > 0$,

$$
\begin{aligned}
1 - &\int_{-L_n}^{L_n} \pi_\alpha(\beta)d\beta \\
&= 2\int_0^{b_n} \left(1-\Phi\left(\frac{L_n}{\lambda\alpha}\right)\right) f_{C^+}(\lambda)d\lambda + 2\int_{b_n}^{\infty} \left(1-\Phi\left(\frac{L_n}{\lambda\alpha}\right)\right) f_{C^+}(\lambda)d\lambda \\
&\leq 2\left(1-\Phi\left(\frac{L_n}{b_n\alpha}\right)\right) + \int_{b_n}^{\infty} \frac{2}{\pi}\frac{1}{1+\lambda^2}d\lambda \\
&\lesssim e^{-L_n^2/b_n^2\alpha^2} + b_n^{-1} \leq b_n^{-1}
\end{aligned}
$$

Choosing $b_n = e^{-cn\bar{\epsilon}_n^2}$ and $L_n = \sqrt{c}\alpha b_n\sqrt{n}\bar{\epsilon}_n$, with an arbitrary $\bar{\epsilon}_n \succ \sqrt{(s_*\log p)/n}$ and a sufficiently large constant $c > 0$, clearly the requirements that the $\log L_n = O(\log p)$ and that the above expression in the display is bounded by $e^{-cn\bar{\epsilon}_n^2}$.

Thus the required conditions hold, and the optimal posterior contraction rate $\epsilon_n = \sqrt{(s_*\log p)/n}$ is obtained.

3.4. Normal-gamma prior

The Bayesian lasso, considered by Park and Casella (2008), considered double-exponential prior for each component β_j, which is ostensibly motivated by the lasso, which becomes the posterior mode. To compute, the Bayesian lasso uses the following representation of a double-exponential density:

$$
ce^{-c|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/2s} \frac{c^2}{2} e^{-c^2 s/2} ds
$$

for any $a > 0$. This allows to write $\beta_j \sim \mathrm{DE}(a)$ through the hierarchical representation

$$
\beta_j|\lambda_j \sim \mathrm{N}(0,\lambda_j), \quad \lambda_j \sim \mathrm{Gamma}(1, 1/c^2). \tag{21}
$$

The main drawback of the Bayesian lasso is that the mixing exponential distribution Gamma$(1, 1/c^2)$ does not put enough mass near zero, so that in order to increase concentration, the scale $1/c^2$ needs to be small, which unfortunately makes the tail too thin, not allowing enough mass near non-zero values. A possible rectification of the problem is to consider a more general gamma distribution Gamma$(\alpha, 1/c^2)$ to lead to the hierarchical representation

$$
\beta_j|\lambda_j \sim \mathrm{N}(0,\lambda_j);\ \lambda_j \sim \mathrm{Gamma}(\alpha, 1/c), \tag{22}
$$

where we have replaced the notation for the scale from c^2 to c. As before, smaller values of the hyper-parameter α makes more shrinkage near zero giving sparser model. The following theorem gives posterior concentration and selection properties under the normal-gamma prior.

Theorem 3.4. *Consider the normal-gamma prior in (22) for the logistic regression model. Assume that the true regression coefficients satisfy the condition that* $\max_{j\in\xi_*}|\beta_{*j}| < E$ *with* $E^2 \prec n\bar{\epsilon}_n^2/s_*$. *If* $\alpha \prec \min\{\bar{\epsilon}_n/p, p^{-(1+\nu)}\}$ *for some constant* $\nu > 0$, *then the posterior concentration property in (9) holds.*

In particular, if $E \prec \log p$ *and* $\alpha \prec p^{-1}\min\{n^{-1/2}, p^{-\nu}\}$ *for some* $\nu > 0$, *then the optimal rate* $\epsilon_n = \sqrt{(s_*\log p)/n}$ *is obtained.*

Proof. The proof is based on the estimates very similar to the Dirichlet–Laplace case, and hence will only be briefly sketched. We shall also put $c = 1$ to simplify the expressions; the general case essentially the same.

We need to verify the prior conditions (3)–(5) hold for the normal-gamma prior density π_α. We note that, using the tail estimate of the normal probability,

$$1 - \int_{-a_n}^{a_n} \pi_\alpha(\beta)d\beta \lesssim \int_0^\infty e^{-a_n^2/2\lambda^2}\frac{1}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\lambda}d\lambda.$$

This is exactly of the form (17), except that a_n there has been replaced by a_n^2 (and the scale 2 is replaced by 1 in the mixing gamma density). Thus the estimate a multiple of $\alpha\log(1/a_n^2) \lesssim p^{-(1+\mu)}$, with any $0 < \mu < \nu$ is obtained.

For the second condition, we note that β_j, $j = 1, \ldots, p - s_*$, are independently and identically distributed with $\mathrm{E}_\alpha\left(|\beta_j|\right) = \mathrm{E}_\alpha\left(\mathrm{E}\left(\beta_j|\lambda_j\right)\right)$ a constant multiple of α. Thus the argument using the central limit theorem given there will apply. Also,

$$\inf_{\beta\in[-E,E]}\pi_\alpha(\beta) \propto \inf_{\beta\in[-E,E]}\int_0^\infty \frac{1}{\psi}e^{-\beta^2/2\lambda}\frac{1}{\Gamma(\alpha)2^\alpha}\lambda^{\alpha-1}e^{-\psi/2}d\lambda.$$

The resulting expression is exactly as in the Dirichlet–Laplace case, except that $|\beta|$ is replaced by β^2, so the same estimate will apply with E replaced by E^2. Finally, to check (6), we have that

$$1 - \int_{-L_n}^{L_n}\pi_\alpha(\beta)d\beta \le \frac{1}{\sqrt{2\pi}\Gamma(\alpha)}\int_0^\infty \psi^{\alpha-1}e^{-L_n^2/2\lambda-\lambda}d\lambda.$$

This is exactly of the form in the Dirichlet–Laplace case except that L_n is replaced by L_n^2. Hence the arguments used earlier will give a bound of the form $e^{-cn\bar{\epsilon}_n^2}$ for any predetermined $c > 0$ with the choice $L_n \succ n\bar{\epsilon}_n^2$. This completes the proof and establishes the theorem.

4. Simulation study

We perform a simulation study to compare the different Bayesian variable selection techniques in the logistic regression model. In order to evaluate the performance for both $n > p$ and $n < p$ scenarios, we generate $N = 100$ datasets with each dataset consisting (1): $n = 500$ observations and $p = 200$ predictors; (2): $n = 200$ observations and $p = 500$ predictors. The design matrices X are generated from the multivariate normal distribution $\mathrm{N}_p(0, \Sigma)$, where the following two data generating methods are considered:

(1) the covariance matrix is $\Sigma = \mathrm{I}_p$ obtained by mutually independent covariates;
(2) correlated covariates in autoregressive structure $\Sigma_{j,k} = 0.8^{|j-k|}$.

Each covariate is further standardized to have mean zero and standard deviation one. The true values of coefficients are $\beta = (1, 1.5, -2, 2.5, 0, 0, \ldots, 0)^T$ so that only the first 4 covariates are included in the model and the remaining $p - 4$ covariates have no effects on the response variable. With each simulated design matrix X, the response variable Y is generated from Bernoulli distribution with success probability $\mathrm{P}(Y = 1|X) = \exp(X\beta)\big/\big(1 + \exp(X\beta)\big)$.

Under each scenario, Bayesian logistic regression (1) is implemented on each simulated data set, with different priors. We examine the performance of the methods corresponding to the following priors on the coefficients β in the logistic regression model:

1. Point mass prior: With probability$1-b$, β_j is from the point mass distribution at 0; and with probability b, β_j follows normal distribution with mean 0 and variance $\sigma_0^2 = 100^2$. The prior on b is beta distribution $b_j \sim \mathrm{Beta}(a_0, b_0)$ with $a_0 = b_0 = 1/2$.
2. Non-informative normal prior: We assume independently and identically distributed normal prior on $\beta_j \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_0^2)$, where we choose $\sigma_0^2 = 100^2$ for a non-informative distribution.
3. Bayesian LASSO: We implement the Bayesian LASSO estimator in Park and Casella (2008). Each coefficient β_j is given a double exponential prior with parameter $\lambda_j\tau$, where λ_j are independent and identically distributed standard exponential representing local shrinkage and $\tau \sim \mathrm{C}^+(0, 1)$ induces global shrinkage.

Table 1
Summary of the simulation study for $n = 500$ and $p = 200$ scenario: comparing the Bayesian logistic regression models with point-mass prior, non-informative normal prior (Normal), Bayesian LASSO, Dirichlet–Laplace prior (DL) and horseshoe prior for independent and correlated covariates. The values in parentheses are standard errors.

Independent	Point mass	Normal	Bayes LASSO	DL	Horseshoe
$\mathcal{L}_1$-error	0.74(0.05)	1090.21(6.02)	20.69(2.99)	4.47(0.05)	3.84(0.20)
$\mathcal{L}_2$-error	0.43(0.03)	128.28(0.18)	2.34(0.48)	0.62(0.02)	0.70(0.05)
Coverage	99.89%	77.61%	99.03%	99.96%	99.83%
FP	0.00(0.00)	0.20(0.01)	0.01(0.01)	0.00(0.00)	0.01(0.01)
FN	0.01(0.01)	0.00(0.00)	0.01(0.01)	0.01(0.01)	0.00(0.00)
Brier score	0.10(0.01)	0.19(0.01)	0.12(0.01)	0.10(0.01)	0.10(0.01)
Correlated	Point mass	Normal	Bayes LASSO	DL	Horseshoe
$\mathcal{L}_1$-error	2.24(0.23)	1307.71(13.00)	12.88(0.58)	5.73(0.19)	3.39(0.29)
$\mathcal{L}_2$-error	1.31(0.13)	126.15(0.49)	2.09(0.08)	1.28(0.10)	0.98(0.11)
Coverage	99.27%	69.15%	98.95%	99.77%	99.70%
FP	0.00(0.000)	0.29(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)
FN	0.25(0.04)	0.00(0.00)	0.26(0.04)	0.35(0.04)	0.15(0.04)
Brier score	0.13(0.02)	0.22(0.03)	0.14(0.01)	0.13(0.01)	0.13(0.02)

4. Dirichlet–Laplace prior: As defined in (15), β_j follows a double exponential prior with scale parameter $\phi_j\tau$, where $(\phi_1, \ldots, \phi_p)^T$ are given Dirichlet$(\alpha, \ldots, \alpha)$ distribution and τ is given Gamma$(p\alpha, 2)$ prior. We fix the shrinkage level at $\alpha = 0.1$.
5. Horseshoe prior: As defined in (19), we assume that $\beta_j \sim \mathrm{N}(0, \lambda_j^2\alpha^2)$ and λ_j follows the standard half-Cauchy distribution. The hyper-parameter α is also given the standard half-Cauchy prior.

Among the above Bayesian methods, logistic regression model with point mass prior is implemented by R function `logit.spike(·)` in R package `BoomSpikeSlab`. For the logistic regression model using the other priors, we sample the coefficients using the algorithm by Polson et al. (2013) that represents the logit function as a Gaussian mixture distribution. The proposed sampling algorithm introduces latent variables following a Pólya-gamma prior distribution. We use the `rpg.devroye(·)` function in R package `BayesLogit` on the latent variable and combine with Gibbs sampling procedures on other model parameters for each prior specification. After the burn-in period consisting of 10,000 Markov chain Monte Carlo (MCMC) iterations, we generate 10,000 posterior samples and save the samples of model parameters at every 5th iteration.

For each prior, we evaluate the accuracy of the estimates in terms of the $\mathcal{L}_2$-error $\|\beta_* - \hat{\beta}\|_2 = \sqrt{\sum_{j=1}^{p}(\beta_{*j} - \hat{\beta}_j)^2}$, $\mathcal{L}_1$-error $\|\beta_* - \hat{\beta}\|_1 = \sum_{j=1}^{p} |\beta_{*j} - \hat{\beta}_j|$ and also monitor the coverage of 95% credible intervals, where the estimates $\hat{\beta}$ are calculated by the posterior means of model parameters. We also compare the variable selection performance using the average proportion of including zero effects (FP) and the average proportion of excluding non-zero effects (FN). For point mass prior, the subset of covariates is selected by posterior non-zero probability larger than 0.5. For the continuous shrinkage priors, the subset of covariates of covariates is identified by whether the 95% posterior credible intervals contain zero or not. For the newly-generated data sets $\{X_i', Y_i'\}_{i=1}^{n'}$ with the sample size $n' = 1000$, we monitor the prediction performance using the Brier score defined by $\mathrm{BS} = \sum_{i=1}^{n'}(Y_i' - \hat{\pi}_i')^2/n'$. The estimates of success probability are the posterior means of $\pi_i' = \exp(X_i'\beta)/(1 + \exp(X_i'\beta))$ at each iteration.

Tables 1 and 2 summarize the simulation results for different priors in the logistic regression model for $p = 200,\ n = 500$ and $p = 500,\ n = 200$ scenarios, respectively. As expected, the point-mass prior achieves the best overall performance in terms of the estimation, variable selection, and prediction accuracies for both $n > p$ and $n < p$ cases. However, the disadvantage of a point-mass prior is the heavy computational cost in generating the MCMC samples, especially when the number of predictors is large. The global–local shrinkage priors such as DL and horseshoe priors mimic the composition of a point mass prior through a dominated peak at zero while retaining heavy tails in a simple distributional format so that computational efficiency is largely improved.

For the simulated datasets with mutually independent predictors, all five priors performed very well in identifying the non-zero effects under both scenarios of dimensions. However, the advantage of the point-mass prior is demonstrated in terms of estimation accuracy. DL and horseshoe priors showed reasonably comparable average $\mathcal{L}_1$- and $\mathcal{L}_2$-errors that are only slightly worse than that of the point-mass prior, and the difference actually became less under $p > n$ scenario, which is obviously a more difficult problem that even the point-mass prior did not work very well. This observation has further supported the benefits of certain continuous shrinkage priors over the point-mass prior for high dimensional logistic regression model. For newly generated testing data sets, the average Brier score for the point-mass prior and the two shrinkage priors are quite close for all scenarios. When the covariates are correlated through an autoregressive structure, similar conclusions can be drawn. Under $n > p$ case, the DL prior and the horseshoe prior even outperform the point-mass prior in terms of the $\mathcal{L}_2$-error for the estimation accuracy, and the horseshoe prior efficiently distinguishes the significant covariates from the trivial ones so that the best variable selection performance is observed. The simulation study demonstrates that DL and horseshoe priors achieve the best model performance compared to the point-mass prior in variable selection, prediction accuracy and computational time.

Table 2
Summary of the simulation study for $n = 200$ and $p = 500$ scenario: comparing the Bayesian logistic regression models with point-mass prior, non-informative normal prior (Normal), Bayesian LASSO, Dirichlet–Laplace prior (DL) and horseshoe prior for independent and correlated covariates. The values in parentheses are standard errors.

Independent	Point mass	Normal	Bayes LASSO	DL	Horseshoe
$\mathcal{L}_1$-error	8.66(1.15)	1530.49(13.28)	64.38(4.90)	10.45(0.95)	21.33(1.01)
$\mathcal{L}_2$-error	3.02(0.67)	160.90(2.13)	19.03(0.98)	4.33(0.53)	8.71(0.64)
Coverage	95.22%	60.31%	90.81%	93.11%	92.89%
FP	0.02(0.01)	0.29(0.02)	0.09(0.01)	0.03(0.00)	0.01(0.01)
FN	0.01(0.01)	0.10(0.01)	0.01(0.01)	0.01(0.01)	0.01(0.01)
Brier score	0.13(0.05)	0.20(0.07)	0.19(0.02)	0.15(0.01)	0.13(0.01)
Correlated	**Point mass**	**Normal**	**Bayes LASSO**	**DL**	**Horseshoe**
$\mathcal{L}_1$-error	19.35(0.99)	2097.33(19.11)	52.90(1.25)	16.03(0.51)	15.94(0.57)
$\mathcal{L}_2$-error	7.39(0.78)	183.44(1.29)	15.45(2.03)	6.41(0.32)	6.02(0.32)
Coverage	90.27%	48.36%	84.37%	89.68%	91.45%
FP	0.05(0.01)	0.41(0.22)	0.12(0.05)	0.07(0.01)	0.05(0.01)
FN	0.35(0.08)	0.20(0.03)	0.29(0.04)	0.33(0.06)	0.39(0.04)
Brier score	0.23(0.08)	0.57(0.10)	0.33(0.08)	0.21(0.07)	0.28(0.04)

5. Conclusion

In this paper, we study the posterior concentration of different priors for variable selection in the logistic regression model. The theoretical results show the consistency properties of the coefficient estimates if the prior distribution satisfies certain properties such that it simultaneously has heavy-tail and significant concentration within a small neighborhood around zero. Specifically, we demonstrate the spike-and-slab prior and some global–local shrinkage priors such as the Dirichlet–Laplace, horseshoe and normal-gamma can give the optimal contraction rate in the logistic regression model, provided that the prior hyper-parameters are chosen following certain restrictions. The continuous shrinkage priors are more computationally efficient than the traditional point mass prior but can obtain optimal contraction rate that is comparable with that of the point-mass prior in terms of the $\mathcal{L}_2$-norm of the estimation error. The simulation study supports the theoretical results on estimation consistency, variable selection performance, and prediction accuracy. One interesting problem that has not been intensely discussed is the thresholding method for continuous shrinkage priors which cannot shrink the trivial effect to exactly zero. The numerical example we had in this manuscript used a simple posterior credible interval method, but a more carefully chosen thresholding technique will be necessary in practical implementation.

Credit authorship contribution statement

Both Dr. Wei and Dr. Ghosal contributed to this manuscript. Dr. Ghosal contributed mainly on the theoretical derivation of this manuscript, and Dr. Wei contributed on writing the details proof derivation and performing the numerical evaluations.

Acknowledgments

Research of the second author was partially supported by NSF grant DMS-1510238.

Appendix. Proof of Theorem 2.1

In order to prove Theorem 2.1, we first present the following general lemma, whose proof follows from the standard arguments used in Bayesian nonparametrics; see Bernardo et al. (1998) and Ghosal and van der Vaart (2017). Below, P^* (or P_{β_*}) and E^* (or E_{β_*}) stand respectively for the probability and expectation under the true distribution, P_α and E_α stand respectively for the probability and expectation for the model parameter uncertainties with hyper-parameter α.

Lemma A.1. *Let B_n and C_n be subset of the parameter space Θ, and ϕ_n be a test function $\phi_n(Z) \in [0, 1]$ based on the data. If $\mathrm{P}(B_n) \le b_n$, $\mathrm{E}_{\beta_*}(\phi_n(Z)) \le b_n'$, $\sup_{\beta \in C_n} \mathrm{E}_\beta\big(1 - \phi_n(Z)\big) \le c_n$, and*

$$\mathrm{P}^*\left\{ \int_{\mathbb{R}^p} \frac{q_{n,\beta}(Z)}{q_{n,\beta_*}(Z)} \pi(\beta) d\beta \ge d_n \right\} \ge 1 - d_n', \tag{23}$$

then for any $\delta_n \to 0$,

$$\mathrm{P}^*\left(\mathrm{P}(C_n \cup B_n | Z, X) \ge \frac{b_n + c_n}{d_n \delta_n} \right) \le \delta_n + b_n' + d_n'. \tag{24}$$

To derive posterior concentration and variable selection results, we use the above lemma with $B_n = \{\text{At least } \bar{s}$ entries of $|\beta|$ are greater than $a_n\}$ and $C_n = \{\|\beta - \beta_*\|_2 > M\epsilon_n\}$ with $\bar{s} = Ln\epsilon_n^2/\log p$ for a sufficiently large $L > 0$. We need to show that a test function ϕ_n, and constants $b_n, c_n, d_n, b_n', d_n'$ such that $b_n', d_n' \to 0$ and $(b_n + c_n)/d_n \to 0$.

Proof of Theorem 2.1. We first show that there exists a test function $\phi_n : \mathcal{Z}^{(n)} \to [0,1]$ such that for some positive constant $c_1 > 0$,

$$\mathrm{E}_{\beta_*}\big[\phi(Z)\big] \leq 4/p \tag{25}$$

$$\sup_{\beta \in C_n} \mathrm{E}_{\beta}\big[1 - \phi(Z)\big] \leq \exp(-c_1 n\epsilon_n^2). \tag{26}$$

The existence of test function can be verified by following the proof of Lemma 14 in Atchadé (2017) for the test of model parameter β under the defined likelihood function $q_{n,\beta}(Z)$. For a non-empty set $\bar{\Theta} \overset{\text{def}}{=} \big\{\beta \in \mathbb{R}^p : |\beta_j| \leq L_n, j = 1, \ldots, p, \#\{j : |\beta_j - \beta_{*j}| > a_n\} \leq \bar{s}\big\}$, define a subset of the sample space by

$$\bar{\mathcal{E}}_n \overset{\text{def}}{=} \Big\{Z \in \mathcal{Z}^{(n)} : \text{for all } \beta \in \beta_* + \bar{\Theta}, |\triangledown \log q_{n,\beta}(Z)^T(\beta - \beta_*)| \leq 4\sqrt{n\bar{s}\log p}\|\beta - \beta_*\|_2/2\Big\},$$

and the restriction of the density $q_{n,\beta}(Z)$ to $\bar{\mathcal{E}}_n$ is denoted by $\mathrm{Q}_{n,\beta}$.

For $M > 2$, $B_p(\bar{\Theta}, M\epsilon_n) \overset{\text{def}}{=} \{\beta \in \beta_* + \bar{\Theta} : \|\beta - \beta_*\|_2 \leq M\epsilon_n\}$ can be represented as $B_p(\bar{\Theta}, M\epsilon_n) = \cup_{l \geq 1} B(l)$, where $B(l) = \{\beta \in \beta_* + \bar{\Theta} : lM\epsilon_n < \|\beta - \beta_*\|_2 < (l+1)M\epsilon_n\}$. For each $l \geq 1$, let S_l be the set of maximal points separated by the distance $lM\epsilon_n/2$ in $B(l)$. Therefore, there is a test function $\phi_{\beta_k} : \mathcal{Z}^{(n)} \to [0,1]$ for each $\beta_k \in S_l$ such that

$$\mathrm{E}_{\beta_*}\big[\phi_{\beta_k}(Z)\big] \leq \sup_{\mathrm{Q} \in \mathrm{conv}(\mathcal{P}_{\beta_k})} \mathcal{H}_{1/2}\big(q_{n,\beta_*}, \mathrm{Q}\big) \tag{27}$$

$$\sup_{\mathrm{Q} \in \mathcal{P}_{\beta_k}} \int_{\mathcal{Z}^{(n)}} \big(1 - \phi_{\beta_k}(Z)\big)\mathrm{Q}(Z)dZ \leq \sup_{\mathrm{Q} \in \mathrm{conv}(\mathcal{P}_{\beta_k})} \mathcal{H}_{1/2}\big(q_{n,\beta_*}, \mathrm{Q}\big), \tag{28}$$

where $\mathcal{H}_{1/2}(q_1, q_2) \overset{\text{def}}{=} \int_{\mathcal{Z}^{(n)}} q_1^{1/2}(z)q_2^{1/2}(z)dz$ is the Hellinger transform for any two integrable non-negative functions on $\mathcal{Z}^{(n)}$, $\mathcal{P}_{\beta_k} \overset{\text{def}}{=} \{\mathrm{Q}_{n,\mu} : \mu \in \beta_* + \bar{\Theta} \text{ and } \|\mu - \beta_k\|_2 \leq lM\epsilon_n/2\}$ and any $\mathrm{Q} \in \mathrm{conv}(\mathcal{P}_\beta)$ can be written as a finite convex combination $\mathrm{Q} = \sum_m \alpha_m \mathrm{Q}_{n,\mu_m}$, where $\alpha_m \geq 0$, $\sum_m \alpha_m = 1$, $\mu_m \in \beta_* + \bar{\Theta}$ and $\|\mu_m - \beta_k\|_2 \leq lM\epsilon_n/2$. Since for all $\beta_k \in B(l)$, $\|\beta_k - \beta_*\| > lM\epsilon_n$, we have $\|\mu_m - \beta_*\|_2 > lM\epsilon_n/2 > \epsilon_n$. Thus, by the definition of the Hellinger distance,

$$\mathcal{H}_{1/2}\left(q_{n,\beta_*}, \mathrm{Q}\right) = \int_{\mathcal{Z}^{(n)}} \sqrt{\sum_m \alpha_m \mathbb{1}_{\bar{\mathcal{E}}_n(Z)} \frac{q_{n,\mu_m}(Z)}{q_{n,\beta_*}(Z)}} q_{n,\beta_*}(Z)dZ. \tag{29}$$

Now for $Z \in \bar{\mathcal{E}}_n$ and $\mu_m \in \beta_* + \bar{\Theta}$, we have that

$$|X_i^T(\mu_m - \beta_*)| \leq \|X\|_\infty \|\mu_m - \beta_*\|_1 \leq \sqrt{\bar{s}}\|\mu_m - \beta_*\|_2 + (p - \bar{s})a_n \leq 2\sqrt{\bar{s}}\|\mu_m - \beta_*\|_2$$

since $(p - \bar{s})a_n \leq (p - \bar{s})\epsilon_n/p < \epsilon_n < \|\mu_m - \beta_*\|_2$. Hence

$$\begin{aligned}
\mathcal{L}_{n,\mu_m} &\leq -\frac{n}{2 + \max_i |X_i^T(\mu_m - \beta_*)|}(\mu_m - \beta_*)^T \frac{X^TWX}{n}(\mu_m - \beta_*) \\
&\leq -\frac{n}{2 + 2\sqrt{\bar{s}}\|\mu_m - \beta_*\|_2}(\mu_m - \beta_*)^T \frac{X^TWX}{n}(\mu_m - \beta_*) \\
&\leq -\frac{n}{2 + 2\sqrt{\bar{s}}\|\mu_m - \beta_*\|_2}\kappa_{1,a_n}(\bar{s})\|\mu_m - \beta_*\|_2^2
\end{aligned}$$

given the definition of $\kappa_{1,a_n}(\bar{s})$ in (7).

Thus for $\mu_m \in \beta_* + \bar{\Theta}$, $\|\mu_m - \beta_*\|_2 > lM\epsilon_n/2$ and $Z \in \bar{\mathcal{E}}_n$, given the definition of $\mathcal{L}_{n,\beta}(Z)$ in (8), we have the following inequality:

$$\begin{aligned}
\frac{q_{n,\mu_m}(Z)}{q_{n,\beta_*}(Z)} &= \exp\left\{\triangledown \log q_{n,\beta_*}(Z)^T(\mu_m - \beta_*) + \mathcal{L}_{n,\mu_m}(Z)\right\} \\
&\leq \exp\left\{\frac{4\sqrt{n\bar{s}\log p}}{2}\|\mu_m - \beta_*\|_2 - \frac{1}{2}\frac{n\kappa_{1,a_n}(\bar{s})\|\mu_m - \beta_*\|_2^2}{1 + \sqrt{\bar{s}}\|\mu_m - \beta_*\|_2}\right\} \\
&\leq \exp\left\{\frac{8\sqrt{n\bar{s}\log p} - \big(n\kappa_{1,a_n}(\bar{s}) - 8\bar{s}\sqrt{n\log p}\big)\epsilon_n}{4\|\mu_m - \beta_*\|_2^{-1} + 4\sqrt{\bar{s}}} - \frac{1}{4}\frac{n\kappa_{1,a_n}(\bar{s})\|\mu_m - \beta_*\|_2^2}{1 + \sqrt{\bar{s}}\|\mu_m - \beta_*\|_2}\right\} \\
&\leq \exp\left\{\frac{8\sqrt{n\bar{s}\log p} - \frac{1}{2}n\kappa_{1,a_n}(\bar{s})\frac{16}{\kappa_{1,a_n}(\bar{s})}\sqrt{\frac{\bar{s}\log p}{n}}}{4\|\mu_m - \beta_*\|_2^{-1} + 4\sqrt{\bar{s}}} - \frac{1}{4}\frac{n\kappa_{1,a_n}(\bar{s})\|\mu_m - \beta_*\|_2^2}{1 + \sqrt{\bar{s}}\|\mu_m - \beta_*\|_2}\right\}
\end{aligned}$$

$$\leq \exp\left\{-\frac{1}{4}\frac{n\kappa_{1,a_n}(\bar{s})(lM\epsilon_n/2)^2}{1+\sqrt{\bar{s}}(lM\epsilon_n/2)}\right\},$$

under the conditions of $\epsilon_n \geq \dfrac{16}{\kappa_{1,a_n}(\bar{s})}\sqrt{\dfrac{\bar{s}\log p}{n}}$ and $n\kappa_{1,a_n}(\bar{s}) - 16\bar{s}\sqrt{n\log p} > 0$. The later one imposes a restriction on the sample size $n \geq \left(16\bar{s}\sqrt{\log p}/\kappa_{1,a_n}(\bar{s})\right)^2$. Since $\bar{s} \overset{\text{def}}{=} Ln\bar{\epsilon}_n^2/\log p$, $\epsilon_n \geq \bar{\epsilon}_n$ and $\sqrt{\bar{s}}lM\epsilon_n > \sqrt{\bar{s}}M\epsilon_n > 2$, Under the assumption that $\kappa_{1,a_n}(\bar{s})$ remains bounded away, the exponent above is up to a constant bounded by

$$\frac{8\sqrt{n\bar{s}\log p}\,lM\epsilon_n(\sqrt{\bar{s}}lM\epsilon_n/2)}{1+\sqrt{\bar{s}}(lM\epsilon_n/2)} \geq \frac{8\sqrt{n\bar{s}\log p}\,lM\epsilon_n}{(\sqrt{\bar{s}}\,lM\epsilon_n/2)^{-1}+1} \geq 4\sqrt{n\bar{s}\log p}\,lM\epsilon_n = c_4lMn\epsilon_n^2.$$

for $c_4 = 4\sqrt{L}$. Therefore, the right hand side of (29) can be bounded by

$$\mathcal{H}_{1/2}\big(q_{n,\beta_*}(Z), Q(Z)\big)^2 \leq \sum_m \alpha_m \exp\left\{-\frac{1}{4}\frac{n\kappa_{1,a_n}(\bar{s})(lM\epsilon_n/2)^2}{1+\sqrt{\bar{s}}(lM\epsilon_n/2)}\right\} \leq \sum_m \alpha_m \exp\{-c_4lMn\epsilon_n^2\}.$$

Thus we conclude that

$$\mathrm{E}_{\beta_*}\left[\phi_{\beta_k}(Z)\right] \leq \exp\left\{-\frac{1}{2}c_4lMn\epsilon_n^2\right\} \tag{30}$$

$$\sup_{Q\in\mathcal{P}_\beta}\int_{\mathcal{Z}^{(n)}}\big(1-\phi_{\beta_k}(Z)\big)Q(Z)dZ \leq \exp\left\{-\frac{1}{2}c_4lMn\epsilon_n^2\right\}. \tag{31}$$

Now we set the test function to be $\phi(Z) = 1\mathrm{l}_{Z\in\bar{\mathcal{E}}_n}\sup_{l\geq 1}\max_{\beta_k\in S_l}\phi_{\beta_k}(Z) + 1\mathrm{l}_{Z\in\bar{\mathcal{E}}_n^c}$. For any $\beta \in \beta_* + \bar{\Theta}$ such that $\|\beta-\beta_*\|_2 > M\epsilon_n$, it will be within the distance of $M\epsilon_n/2$ of a point in S_l for some $l \geq 1$. Hence by (31),

$$\sup_{\beta\in C_n}\mathrm{E}_\beta\left[1-\phi(Z)\right] \leq \int_{\mathcal{Z}^{(n)}}\big(1-\phi_\beta(Z)\big)1\mathrm{l}_{Z\in\bar{\mathcal{E}}_n}q_{n,\beta}(Z)dZ \leq \exp\left\{-\frac{1}{2}c_4Mn\epsilon_n^2\right\}. \tag{32}$$

Since the size of the defined set S_l is upper bounded by $D_l \overset{\text{def}}{=} D(lM\epsilon_n/2, B_p\big(\bar{\Theta}, (l+1)M\epsilon_n\big)$, the number of points within $B_p\big(\bar{\Theta}, (l+1)M\epsilon_n\big)$ that are separated by $\mathcal{L}_2$ distance of $lM\epsilon_n/2$. Then by (30) and the test function $\phi(Z)$,

$$\mathrm{E}_{\beta_*}\left[\phi(Z)\right] \leq \mathrm{P}^*\big(Z\in\bar{\mathcal{E}}_n^c\big) + \sum_{l\geq 1}D_l\exp\left\{-\frac{1}{2}c_4lMn\epsilon_n^2\right\}. \tag{33}$$

By Hoeffding's inequality and $|\triangledown\log q_{n,\beta_*}(Z)^T(\beta-\beta_*)| \leq \|\log q_{n,\beta_*}(Z)\|_\infty\|\beta-\beta_*\|_1$,

$$\begin{aligned}
\mathrm{P}^*\big(Z\in\bar{\mathcal{E}}_n^c\big) &\leq \mathrm{P}^*\left(\|\log q_{n,\beta_*}(Z)\|_\infty > \frac{4\sqrt{n\log p}}{2}\right)\\
&\leq \mathrm{P}^*\left(\max_{1\leq j\leq p}\left|\sum_{i=1}^n\big(z_i - g'(X_i^T(\beta-\beta_*))\big)x_{ij}\right| > \frac{4\sqrt{n\log p}}{2}\right)\\
&\leq 2\exp\left(\log p - \frac{16n\log p}{8n}\right) \leq 2/p.
\end{aligned} \tag{34}$$

For continuous shrinkage prior, no exact zero will be generated on β. Therefore, a thresholding policy is applied. When $|\beta_j| < a_n$ where $a_n < \epsilon_n/p$, the coefficient is treated as zero. For all the coordinates $j \notin \xi_*$ such that $\beta_{*j} = 0$, the set $B_p\big(\bar{\Theta}, (l+1)M\epsilon_n\big)$ will cover all the points that $|\beta_j| < a_n < (l+1)M\epsilon_n$ around $\beta_{*j} = 0$ and the distance between these points and β_{*j} will be smaller than $lM\epsilon_n/2$. By the definition of $\bar{\Theta}$, although the shrinkage prior does not produce a strict sparsity on β, the $p-\bar{s}$ coordinates where $|\beta_j| < a_n$ are treated like zero in the point-mass prior. Using the arguments given in Example 7.1 of Ghosal et al. (2000) for a k-dimensional ball, $k \leq \bar{s}$, we obtain the bound $\sup_{l\geq 1}D_l \leq (24)^{\bar{s}}e^{\bar{s}\log(pe)+\bar{s}\log L_n} \leq \exp[cn\bar{\epsilon}_n^2]$ for some $c > 0$. Therefore, the last term on the right hand side of (33) can be bounded by

$$\begin{aligned}
\sum_{l\geq 1}D_l\exp\left\{-\frac{1}{2}c_4lMn\epsilon_n^2\right\} &\leq (24)^{\bar{s}}e^{\bar{s}\log(pe)}\sum_{l\geq 1}\exp\left\{-\frac{1}{2}c_4lMn\epsilon_n^2\right\}\\
&= (24)^{\bar{s}}e^{\bar{s}\log(pe)}\frac{\exp\{-\frac{1}{2}c_4Mn\epsilon_n^2\}}{1-\exp\{-\frac{1}{2}c_4Mn\epsilon_n^2\}}\\
&\leq 2\exp\left\{\bar{s}\log(24e) + \bar{s}\log p - 2\sqrt{L}Mn\epsilon_n^2\right\}\\
&\leq 2\exp\left\{\bar{s}\log(24e) + \bar{s}\log p - 2\times 16^2\sqrt{L}M\bar{s}\log p/\kappa_{1,a_n}(\bar{s})^2\right\}.
\end{aligned}$$

If M is large enough such that $\frac{2\times 16^2\sqrt{LM}}{\kappa_{1,a_n}(\bar{s})^2} > 1+\log(24e)$, we can conclude that

$$\sum_{l\geq 1} D_l \exp\left\{-\frac{1}{2}c_4 lMn\epsilon_n^2\right\} \leq 2/p. \tag{35}$$

The condition can be verified since $M > L'\kappa_{1,a_n}(\bar{s})^2$ for a sufficiently large constant $L' > 0$.

Combining (33)–(35), we obtain (25). Setting $c_1 = \frac{1}{2}c_4 M = 2\sqrt{LM}$ and using the bound in (32), we show that (26) holds. Thus we verify the existence of a test function with the stated properties.

Next we show that for sufficiently large n,

$$\mathrm{P}^*\left\{\int_{\mathbb{R}^p}\frac{q_{n,\beta}(Z)}{q_{n,\beta_*}(Z)}\Pi_\alpha(\beta)d\beta \geq e^{-c_2 n\epsilon_n^2}\right\} > 1-2/p \tag{36}$$

for some positive constant $c_2 < c_1$. This can be achieved by setting $M > c_2/2\sqrt{L}$. Now for all the $\beta \in \mathbb{R}^p$ and $Z \in \mathcal{E}_{n,0} \overset{\text{def}}{=} \left\{Z \in \mathcal{Z}^{(n)} : \|\triangledown \log q_{n,\beta_*}(Z)\|_\infty \leq 2\sqrt{n\log p}\right\}$, using $g^{(2)}(z) \leq 1/4$, we have that

$$\begin{aligned}\int_{\mathbb{R}^p}\frac{q_{n,\beta}(Z)}{q_{n,\beta_*}(Z)}\Pi_\alpha(\beta)d\beta &= \int_{\mathbb{R}^p}\exp\left\{\triangledown\log q_{n,\beta_*}(Z)^T(\beta-\beta_*)+\mathcal{L}_{n,\beta}(Z)\right\}\Pi_\alpha(\beta)d\beta\\ &\geq \int_{\mathbb{R}^p}\exp\left\{-\|\triangledown\log q_{n,\beta_*}(Z)\|_\infty\|\beta-\beta_*\|_1-\frac{n}{8}\|X(\beta-\beta_*)/\sqrt{n}\|_2^2\right\}\Pi_\alpha(\beta)d\beta\\ &\geq \int_{\mathbb{R}^p}\exp\left\{-2\sqrt{n\log p}\|\beta-\beta_*\|_1-\frac{n}{8}\|\beta-\beta_*\|_1^2\right\}\Pi_\alpha(\beta)d\beta,\end{aligned}$$

since $\left|\triangledown\log q_{n,\beta_*}(Z)^T(\beta-\beta_*)\right| \leq \|\triangledown\log q_{n,\beta_*}(Z)\|_\infty\|\beta-\beta_*\|_1$ and

$$\|X(\beta-\beta_*)\|_2^2 = \sum_{i=1}^n\left(X_i^T(\beta-\beta_*)\right)^2 \leq n\left(\|X\|_\infty\|\beta-\beta_*\|_1\right)^2.$$

Therefore, for a sufficiently small constant $\eta > 0$,

$$\int_{\mathbb{R}^p}\frac{q_{n,\beta}(Z)}{q_{n,\beta_*}(Z)}\Pi_\alpha(\beta)d\beta \geq \exp\left\{-4\eta\sqrt{n\log p}\epsilon_n-\eta^2 n\epsilon_n^2/2\right\}\mathrm{P}_\alpha\left(\{\|\beta-\beta_*\|_1 < 2\eta\epsilon_n\}\right). \tag{37}$$

From the proof of Theorem A.2 in Song and Liang (2017), the event $\{\|\beta-\beta_*\|_1 < 2\eta\epsilon_n\}$ contains the event

$$\left\{\beta_j \in \left[\beta_{*j}-\frac{\eta\epsilon_n}{s_*}, \beta_{*j}+\frac{\eta\epsilon_n}{s_*}\right] \text{ for all } j\in\xi_*\right\} \cap \left\{\sum_{j\notin\xi_*}|\beta_j| \leq \eta\epsilon_n\right\}.$$

For the prior distribution $\Pi_\alpha(\beta) = \prod_{j=1}^p \pi_\alpha(\beta_j)$ satisfying conditions (3)–(5), we have

$$\mathrm{P}_\alpha\left(\sum_{j\notin\xi_*}|\beta_j| \leq \eta\epsilon_n\right) = \int_{\|\beta\|_1\leq\eta\epsilon_n}\prod_{j=1}^{p-s_*}\pi_\alpha(\beta_j)d\beta_1\cdots d\beta_{p-s_*} \geq e^{-c_5 n\epsilon_n^2}, \tag{38}$$

and since $|\beta_{*j}|-\eta\epsilon_n/s_* \geq -E$, $|\beta_{*j}|+\eta\epsilon_n/s_* \leq E$,

$$\mathrm{P}_\alpha\left(\left\{\beta_j \in \left[\beta_{*j}-\frac{\eta\epsilon_n}{s_*}, \beta_{*j}+\frac{\eta\epsilon_n}{s_*}\right] \text{ for all } j\in\xi_*\right\}\right) \geq \frac{2\eta\epsilon_n}{s_*}\cdot\inf_{\beta\in[-E,E]}\pi_\alpha(\beta) \geq e^{-c_6 n\epsilon_n^2} \tag{39}$$

for some constants $c_5 > 0$ and $c_6 > 0$. Therefore, given (38) and (39),

$$\begin{aligned}\mathrm{P}_\alpha\left(\{\|\beta-\beta_*\|_1 < 2\eta\epsilon_n\}\right) \geq &\mathrm{P}_\alpha\left(\left\{\beta_j \in \left[\beta_{*j}-\frac{\eta\epsilon_n}{s_*}, \beta_{*j}+\frac{\eta\epsilon_n}{s_*}\right] \text{ for all } j\in\xi_*\right\}\right)\\ &\times\ \mathrm{P}_\alpha\left(\sum_{j\notin\xi_*}|\beta_j| \leq \eta\epsilon_n\right) \geq e^{-(c_5+c_6)n\epsilon_n^2}.\end{aligned} \tag{40}$$

Combining the results in (37) and (40) with the constraints $s_*\log p \prec n\epsilon_n^2$, we conclude that, for $c_2 \overset{\text{def}}{=} 4\eta/\sqrt{s_*}+\eta^2/2+c_5+c_6$ and $\epsilon_n \leq \sqrt{(s_*\log p)/n}$,

$$\int_{\mathbb{R}^p}\frac{q_{n,\beta}(Z)}{q_{n,\beta_*}(Z)}\Pi_\alpha(\beta)d\beta \geq \exp\left\{-\left(4\eta/\sqrt{s_*}+\eta^2/2\right)n\epsilon_n^2-c_2'n\epsilon_n^2\right\} = e^{-c_2 n\epsilon_n^2}. \tag{41}$$

Since (41) holds for all $Z \in \mathcal{E}_{n,0}$, we have that

$$\begin{aligned}
&\mathrm{P}^*\left\{\int_{\mathbb{R}^p} \frac{q_{n,\beta}(Z)}{q_{n,\beta_*}(Z)} \Pi_\alpha(\beta)d\beta \geq e^{-c_2 n\epsilon_n^2}\right\} \\
&\geq 1 - \mathrm{P}^*(Z \notin \mathcal{E}_{n,0}) \\
&\geq 1 - \mathrm{P}^*\left(\max_{1\leq j\leq p}\Big|\sum_{i=1}^n (Z_i - g'(X_i^T(\beta-\beta_*)))x_{ij}\Big| > 2\sqrt{n\log p}\right) \\
&\geq 1 - 2\exp\left(\log p - \frac{16n\log p}{8n}\right) = 1 - 2/p,
\end{aligned}$$

leading to (36).

Finally, for $B_n = \{\text{At least } \bar{s} \text{ entries of } |\beta| \text{ is greater than } a_n\}$ with $\bar{s} = Ln\bar{\epsilon}_n^2/\log p$, we show that the probability $\mathrm{P}_\alpha(B_n) \leq \exp(-c_3 n\epsilon_n^2)$ for constant $c_3 > 0$. Let $N = \#\big\{j : |\beta_j| \geq a_n\big\}$. Note that $N \sim \mathrm{Bin}(p, \nu_n)$, where $\nu_n = \int_{|\beta|\geq a_n} \pi_\alpha(\beta)d\beta$, so $\mathrm{P}_\alpha(B_n) = \mathrm{P}_\alpha\big(N \geq \bar{s}\big)$. By the condition (3) on a shrinkage prior, $\nu_n \asymp p^{-(1+\mu)}$ for $\mu > 0$ and Lemma A.3 of Song and Liang (2017), we have

$$\mathrm{P}_\alpha(B_n) \leq 1 - \Phi\Big(\sqrt{2p\cdot H(\nu_n,(\bar{s}-1)/p)}\Big) \leq \frac{\exp\big\{-p\cdot H(\nu_n,(\bar{s}-1)/p)\big\}}{\sqrt{2\pi}\sqrt{2p\cdot H(\nu_n,(\bar{s}-1)/p)}},$$

where $p\cdot H(\nu_n,(\bar{s}-1)/p) = (\bar{s}-1)\log\big(\frac{\bar{s}-1}{p\nu_n}\big) + (p-\bar{s}+1)\log\big(\frac{p-\bar{s}+1}{p-p\nu_n}\big)$ and it is sufficient to show that $p\cdot H(\nu_n,(\bar{s}-1)/p) \gtrsim n\bar{\epsilon}_n^2$. Since $1/(p\nu_n) \geq p^{\nu_n}$ and $\bar{s} = Ln\epsilon_n^2/\log(p)$, then we have $(\bar{s}-1)\log(\dfrac{\bar{s}-1}{p\nu_n}) \asymp n\epsilon_n^2$ and $(p-\bar{s}+1)\log(\dfrac{p-\bar{s}+1}{p-p\nu_n}) \prec n\epsilon_n^2$. Therefore, for a constant c_3, $\mathrm{P}_\alpha(B_n) < e^{-c_3 n\epsilon_n^2}$.

Combining (25), (26), (36) and $\mathrm{P}_\alpha(B_n) < e^{-c_3 n\bar{\epsilon}_n^2}$ and using Lemma A.1, we conclude that there exist constants c_1, c_2 and c_3, such that

$$\mathrm{P}^*\left(\mathrm{P}_\alpha\big(C_n \cup B_n|Z,X\big) \geq \frac{e^{-c_3 n\bar{\epsilon}_n^2} + e^{-c_1 n\bar{\epsilon}_n^2}}{\delta_n e^{-c_2 n\bar{\epsilon}_n^2}}\right) \leq \delta_n + 6/p,$$

where B_n and C_n are as defined in the theorem, δ_n is any positive sequence and the constants c_1, c_2, c_3 are such that $c_2 < \min(c_1, c_3)$. This can be achieved by setting η small so that $c_2 = 4\eta/\sqrt{s_*} + \eta^2/2 + c_5 + c_6$ is smaller than c_3 and by setting M large so that $M > c_2/2\sqrt{L}$. Then with $c_0 \overset{\text{def}}{=} \frac{1}{2}(c_2 - \min(c_1, c_3))$ and $\delta_n \overset{\text{def}}{=} e^{-c_0 n\bar{\epsilon}_n^2}$, we have

$$\mathrm{P}^*\big(\mathrm{P}_\alpha(C_n \cup B_n|Z,X) \geq \exp\{-c_0 n\bar{\epsilon}_n^2\}\big) \leq \exp\{-c_0 n\bar{\epsilon}_n^2\} + 6/p \leq 7/p,$$

thus establishing the result.

References

Armagan, A., Dunson, D.B., Lee, J., 2013a. Generalized double Pareto shrinkage. Statist. Sinica 23 (1), 119.

Armagan, A., Dunson, D.B., Lee, J., Bajwa, W.U., Strawn, N., 2013b. Posterior consistency in linear models under shrinkage priors. Biometrika 100 (4), 1011–1018.

Atchadé, Y.F., 2017. On the contraction properties of some high-dimensional quasi-posterior distributions. Ann. Statist. 45, 2248–2273.

Belitser, E., Ghosal, S., 2019. Empirical Bayes oracle uncertainty quantification for regression. Ann. Stat. preprint.

Belitser, E., Nurushev, N., 2020. Needles and straw in a haystack: robust confidence for possibly sparse sequences. Bernoulli 26 (1), 191–225.

Bernardo, J., Burger, J., Smith, A., 1998. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems.

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

Bhattacharya, A., Pati, D., Pillai, N.S., Dunson, D.B., 2015. Dirichlet-Laplace priors for optimal shrinkage. J. Amer. Stat. Assoc. 110 (512), 1479–1490.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2009. Handling sparsity via the Horseshoe. In: AISTATS, Vol. 5. pp. 73–80.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2010. The Horseshoe estimator for sparse signals. Biometrika 97 (2), 465–480.

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., 2015. Bayesian linear regression with sparse priors. Ann. Statist. 43 (5), 1986–2018.

Castillo, I., van der Vaart, A., 2012. Needles and straw in a haystack: posterior concentration for possibly sparse sequences. Ann. Statist. 40 (4), 2069–2101.

George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. J. Amer. Statist. Assoc. 88 (423), 881–889.

Ghosal, S., 1997. Normal approximation to the posterior distribution for generalized linear models with many covariates. Math. Methods Statist. 6 (3), 332–348.

Ghosal, S., 1999. Asymptotic normality of posterior distributions in high-dimensional linear models. Bernoulli 5 (2), 315–331.

Ghosal, S., 2000. Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. J. Multivariate Anal. 74 (1), 49–68.

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

Ghosal, S., van der Vaart, A., 2017. Fundamentals of Nonparametric Bayesian Inference, Vol. 44. Cambridge University Press.

Green, P.J., 1995. Reversible Jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82 (4), 711–732.

Griffin, J.E., Brown, P.J., 2010. Inference with Normal-Gamma prior distributions in regression problems. Bayesian Anal. 5 (1), 171–188.

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

Jiang, W., 2007. Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. Ann. Statist. 35 (4), 1487–1511.

Martin, R., Mess, R., Walker, S.G., 2017. Empirical Bayes posterior concentration in sparse high-dimensional linear models. Bernoulli 23 (3), 1822–1847.

Norets, A., Pati, D., 2017. Adaptive Bayesian estimation of conditional densities. Econom. Theory 33 (4), 980–1012.

Park, T., Casella, G., 2008. The Bayesian LASSO. J. Amer. Stat. Assoc. 103 (482), 681–686.

Polson, N.G., Scott, J.G., Windle, J., 2013. Bayesian inference for logistic models using Pólya-Gamma latent variables. J. Amer. Statist. Assoc. 108 (504), 1339–1349.

Shen, W., Ghosal, S., 2016. Adaptive Bayesian density regression for high-dimensional data. Bernoulli 22 (1), 396–420.

Song, Q., Liang, F., 2017. Nearly optimal Bayesian shrinkage for high dimensional regression. arXiv arXiv:1712.08964.

Van Der Pas, S., Kleijn, B., Van Der Vaart, A., 2014. The Horseshoe estimator: posterior concentration around nearly black vectors. Electron. J. Stat. 8 (2), 2585–2618.

Yang, Y., Tokdar, S.T., 2015. Minimax-optimal nonparametric regression in high dimensions. Ann. Statist. 43 (2), 652–674.