Bayesian Modeling of Sparse High–Dimensional Data using Divergence Measures

Gyuhyeong Goh and Dipak K. Dey

Abstract

We introduce a novel divergence–based–approach, called *Bregman divergence*, to model sparse high–dimensional regression problems. We further introduce a new prior which induces a new version of the (approximate) adaptive lasso in a Bayesian framework. Unlike the original adaptive lasso in which the weights need to be pre-specified prior to the estimation, in our approach the coefficient estimates are directly used as the weights. In addition, due to the generality of the Bregman divergence, the proposed model is easily extended to generalized linear models as well as the group lasso.

KEY WORDS: Bayesian lasso; Bregman divergence; GD prior; Sparse high–dimensional data.

Gyuhyeong Goh is Ph.D. Student, Department of Statistics, University of Connecticut, 215 Glenbrook Road, Storrs, CT 06269–4120 (E–mail: gyuhyeong.goh@uconn.edu). Dipak K. Dey is Board of Trustees Distinguished Professor, Department of Statistics, University of Connecticut, 215 Glenbrook Road, Storrs, CT 06269–4120 (E–mail: dipak.dey@uconn.edu).

1 Introduction

In various fields of scientific research, the following linear regression model has been most commonly used:

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

where $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ is the n–dimensional response vector, $\boldsymbol{X}$ is the $n \times p$ predictor matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the unknown coefficient vector (which is of our interest), and $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}_n, \sigma^2 \boldsymbol{I}_n)$ is a n–dimensional noise vector with known σ^2. In many practical situations, we encounter that the number of coefficients p is very large or larger than the sample size n. For example, in gene expression data, the number of genes is larger (even much larger) than the number of subjects. Let $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^{\mathrm{T}}$ be the true value of $\boldsymbol{\beta}$, i.e., $\boldsymbol{y}|\boldsymbol{\beta}^* \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}^*, \sigma^2 \boldsymbol{I}_n)$. In such situation ($p > n$), the assumption that the true coefficient vector $\boldsymbol{\beta}^*$ is sparse (i.e., many components of $\boldsymbol{\beta}^*$ are zero) is necessarily required, unless $\boldsymbol{\beta}^*$ is not identifiable. Fortunately, the necessary assumption is realistic, because many researchers are interested in finding a small number of predictors that are significantly related to the response variable. For instance, the genomic study investigates the numerous genes to find only a few genes that determines a certain phenotype. Hence, throughout this paper, the number of non–zero coefficients, $p^* = \dim\left(\{\beta_j^* : \beta_j^* \neq 0\}\right)$, is assumed to be much less than the sample size, that is $p^* \ll n$, while $p \approx n$ or $p > n$.

For such sparse high–dimensional setting, a penalized likelihood estimation (PLE) is the most popular technique to estimate the unknown parameter $\boldsymbol{\beta}$. In the PLE, the estimate is

defined as

$$\hat{\boldsymbol{\beta}}_{\text{PLE}} = \arg\min_{\boldsymbol{\beta}} \left[\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \|^2 + \text{Pe}_\lambda(\boldsymbol{\beta}) \right], \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm such that $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^{\text{T}}\boldsymbol{x}}$ and $\text{Pe}_\lambda(\cdot)$ is a deterministic penalty function with a regularization parameter λ controlling the degree of penalization. For instance, Tibshirani (1996) proposed the *lasso* estimate using the ℓ_1 penalty as follows:

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left[\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \|^2 + \lambda \|\boldsymbol{\beta}\|_1 \right], \tag{3}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ and $\lambda > 0$. In fact, from a Bayesian viewpoint, the PLE method can be viewed as finding the *maximum a posteriori* (MAP) estimate as follows:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{MAP}} &= \arg\max_{\boldsymbol{\beta}} \pi(\boldsymbol{\beta}|\boldsymbol{y}) \\ &= \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}), \end{aligned}$$

where $f(\boldsymbol{y}|\boldsymbol{\beta})$ and $\pi(\boldsymbol{\beta})$ are respectively the likelihood function and the prior density function such that

$$f(\boldsymbol{y}|\boldsymbol{\beta}) \propto \exp\left\{ - \| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \|^2 \right\} \quad \text{and} \quad \pi(\boldsymbol{\beta}) \propto \exp\left\{ -\text{Pe}_\lambda(\boldsymbol{\beta}) \right\}. \tag{4}$$

From the aforementioned perspective, Park and Casella (2008) introduced the *Bayesian lasso* using the Laplace prior, which corresponds to the ℓ_1 penalty in the lasso. For more details and examples of the relationship between Bayesian modeling and PLE, see Kyung et al. (2010).

In this paper, we look at the high-dimensional problem from the Bayesian perspective based on divergence measures. To generalize and unify many existing methods, we use *Bregman divergence* (Bregman, 1967) as a major ingredient. Let $\phi : \Omega \to \mathbb{R}$ be a strictly

convex function on a convex set $\Omega \subseteq \mathbb{R}^m$, assumed to be nonempty and differentiable. Then for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$ the Bregman divergence with respect to ϕ is defined as

$$\mathrm{BD}_\phi(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - (\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}} \nabla\phi(\boldsymbol{y}), \tag{5}$$

where $\nabla\phi$ represents the gradient vector of ϕ. The Bregman divergence is indeed a general class of loss functions that includes squared Euclidean distance, Kullback-Leibler (KL) divergence, Itakura-Saito distance (Itakura and Saito, 1970), and Mahalanobis distance as special cases. Due to its generality, many applications of the Bregman divergence have played a key role in recent advances in machine learning (Banerjee et al., 2005; Kulis et al., 2009; Vemuri et al., 2011). Here, we turn the spotlight on its applications in statistical modeling. Our approach is motivated by the duality property between *loss function* (or divergence measure) and *probability density function*; the loss function can be viewed as the negative of the log likelihood function (Bernardo and Smith, 1994; Mallick et al., 2005). For example, if $\phi(\boldsymbol{x}) = ||\boldsymbol{x}||^2$, then the likelihood function in (4) can be expressed as

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\beta}) &\propto \exp\left\{-||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2\right\} \\ &= \exp\left[-\left\{||\boldsymbol{y}||^2 - ||\boldsymbol{X}\boldsymbol{\beta}||^2 - (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(2\boldsymbol{X}\boldsymbol{\beta})\right\}\right] \\ &= \exp\left\{-\mathrm{BD}_\phi(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta})\right\}. \end{aligned}$$

While the Bregman divergence induces a large class of density functions, the prior $\pi(\boldsymbol{\beta}) \propto \exp\{-\mathrm{Pe}_\lambda(\boldsymbol{\beta})\}$ could not be expressed as the form of Bregman divergence because the penalty function $\mathrm{Pe}_\lambda(\cdot)$ is usually very sharp (non–differentiable) to prompt sparse solutions. To overcome the drawback, we introduce a new penalty function which is strictly convex and differentiable. By the convexity and smoothness (differentiability), the new penalty can be

expressed as the Bregman divergence. In addition, it induces approximate sparse solution; many coefficient estimates are approximately zero.

The outline of the remainder of the paper is as follows. In Section 2, we discuss the relationship between prior distributions and divergence measures in the sparse high–dimensional setting. In Section 3, we introduce a new prior along with important and attractive properties. In Section 4, we propose a new divergence–based–approach to the sparse high–dimensional data using hierarchical Bayesian modeling. In addition, many extensions are introduced as well as several iterative algorithms to find MAP estimates. In Section 5 and 6, respectively, a Bayesian screening method is introduced and the asymptotic properties of the proposed method are discussed. Some illustrative examples with simulation studies are described in Section 7. Section 8 offers some concluding remarks and summarizes the salient features of our approach.

2 Prior Elicitation with Divergence Measures

In Bayesian modeling, the use of prior distribution enables us to reflect our prior knowledge in the parameter estimation procedure. As already mentioned, in the sparse high–dimensional problems, the sparsity is presumed in order to guarantee identifiability of the parameter estimate. In order to elicit the prior distribution based on the sparsity, we propose to construct the prior density function using the duality property between divergence measure and probability density function as follows:

$$\pi(\boldsymbol{\beta}) \propto \exp\left\{-d\left(\boldsymbol{\beta}, \mathbf{0}_p\right)\right\}, \tag{6}$$

where $\mathbf{0}_p = (0, \ldots, 0)^{\mathrm{T}}$ is the p–dimensional zero vector and $d(\cdot, \cdot)$ is a divergence measure. Since the divergence measure, $d(\boldsymbol{\beta}, \mathbf{0}_p)$, represents a dissimilarity between the coefficient vector and the zero vector, the probability density induced by the prior density in (6) has the highest density at $\mathbf{0}_p$ and decreases as $\boldsymbol{\beta}$ moves away from $\mathbf{0}_p$. To prompt many sparse solutions for $\boldsymbol{\beta}$, we need to consider an appropriate divergence measure that induces a zero concentrated prior in which the most of probabilities are concentrated in a neighborhood of zero. Let us consider the equal weighted Manhattan distance (or ℓ_1 distance) as

$$d_{\mathrm{M}}(\boldsymbol{\beta}, \mathbf{0}_p) = \sum_{j=1}^{p} \lambda |\beta_j - 0|, \tag{7}$$

where $\lambda > 0$ and the degree of the concentration can be controlled by λ. In this case, our approach is equivalent to the lasso methods (Tibshirani, 1996; Park and Casella, 2008). However, the prior induced by (7) could bring about biased estimates of non-zero coefficients in $\mathcal{B} = \{\beta_j^* : \beta_j^* \neq 0\}$, because the prior information for the non–zero coefficients is mis–reflected; in fact, we should assign less probability densities at zero for the non–zero coefficients because we already knew that the coefficients are not zero. To remedy the bias problem, we can consider the distinctively weighted Manhattan distance as

$$d_{\mathrm{WM}}(\boldsymbol{\beta}, \mathbf{0}_p) = \sum_{j=1}^{p} \lambda_j |\beta_j - 0|, \tag{8}$$

where $\lambda_j > 0$. It is appropriate to assign large values to λ_j's for zero coefficients and small values for non-zero coefficients. Note that, by taking $\lambda_j = \lambda \hat{w}_j$, it reduces to the *adaptive lasso* penalty (Zou, 2006) as:

$$d_{\mathrm{WM}}(\boldsymbol{\beta}, \mathbf{0}_p) = \sum_{j=1}^{p} \lambda \hat{w}_j |\beta_j - 0|, \tag{9}$$

where $\lambda > 0$ and $\hat{w}_j$'s are the data–driven weights such that $\hat{w}_j = 1/|\hat{\beta}_j|^\gamma$, where $\hat{\beta}_j$ is $\sqrt{n}$–consistent estimator for β_j for $j = 1, \ldots, p$ and $\gamma > 0$. Interestingly, in a Bayesian framework, the weights can be elicited by assigning a hyperprior for each λ_j in (8). Define $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$, where $\lambda_j > 0$ for $j = 1, \ldots, p$. The idea of assigning the hyperprior for the regularization parameter was originally proposed by Park and Casella (2008). They assigned the following gamma prior to $\boldsymbol{\lambda}$:

$$\pi(\boldsymbol{\lambda}) \propto \prod_{j=1}^{p} \lambda_j^{a_0 - 1} \exp\left(-b_0 \lambda_j\right), \tag{10}$$

where $a_0 > 0$ and $b_0 > 0$. Let $\frac{\boldsymbol{a}_0}{\boldsymbol{b}_0} = (\frac{a_0}{b_0}, \ldots, \frac{a_0}{b_0})^T$ be the p–dimensional hyper–parameter vector. Define the weighted Itakura–Saito distance between $\boldsymbol{\lambda}$ and $\frac{\boldsymbol{a}_0}{\boldsymbol{b}_0}$ as

$$d_{\mathrm{IS}}\left(\boldsymbol{\lambda}, \frac{\boldsymbol{a}_0}{\boldsymbol{b}_0}\right) = \sum_{j=1}^{p} w \left\{ \frac{\lambda_j}{a_0/b_0} - \log\left(\frac{\lambda_j}{a_0/b_0}\right) + 1 \right\}. \tag{11}$$

By taking $w = a_0$, we can easily show that the gamma prior can be expressed as

$$\pi(\boldsymbol{\lambda}) \propto \exp\left\{ -d_{\mathrm{IS}}\left(\boldsymbol{\lambda}, \frac{\boldsymbol{a}_0}{\boldsymbol{b}_0}\right) \right\}. \tag{12}$$

Hence, in the gamma prior, $\boldsymbol{\lambda}$ attains the highest probability density at $\frac{\boldsymbol{a}_0}{\boldsymbol{b}_0}$. The following lemma tells us the meaning of assigning such gamma prior for $\boldsymbol{\lambda}$ in (8).

Lemma 1. *Let $f(\boldsymbol{y}|\boldsymbol{\beta})$ be a likelihood function which does not depend on $\boldsymbol{\lambda}$ and $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda})$ be a prior density function of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ such that*

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{\lambda}) &\propto \pi(\boldsymbol{\beta}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda}) \\ &\propto \prod_{j=1}^{p} \lambda_j \exp\left\{-\lambda_j |\beta_j|\right\} \times \prod_{j=1}^{p} \lambda_j^{a_0 - 1} \exp\left(-b_0 \lambda_j\right), \end{aligned} \tag{13}$$

where $a_0 > 0$ and $b_0 > 0$. Suppose that $\hat{\boldsymbol{\beta}}$ is the MAP estimate for $\boldsymbol{\beta}$. Then, the MAP

estimate of $\boldsymbol{\lambda}$ *is given as*

$$\hat{\boldsymbol{\lambda}} = \left(\frac{a_0}{b_0 + |\hat{\beta}_1|}, \ldots, \frac{a_0}{b_0 + |\hat{\beta}_p|} \right). \tag{14}$$

The proof of lemma 1 can be easily shown by differentiating Eq. (13) with respect to λ_j's, letting them be zero, and finding the solutions of the equations. It is worth noting that $\hat{\boldsymbol{\lambda}}$ in (14) is a weight function, which is decreasing in $\hat{\boldsymbol{\beta}}$, and it is analogous to the adaptive lasso when $\gamma = 1$, $a_0 = \lambda$ and $b_0 \downarrow 0$. While the adaptive lasso requires to specify the weights a priori to the estimation, the Bayesian method uses the MAP estimate $\hat{\boldsymbol{\beta}}$ as the weight. Hence the weight is automatically determined in the estimation procedure, which is a very important feature of our proposed methodology.

3 GD prior

One of advantages of Bayesian modeling is that interval estimation, called credible interval (CI), can be easily constructed based on the posterior distribution. In general, constructing CI is accomplished by using Markov Chain Monte Carlo (MCMC) samples from the full posterior distribution. However, in the high–dimensional setting, the MCMC technique is often computationally inefficient; it requires an enormous amount of time to obtain the stable MCMC samples. As an alternative, using the Laplace approximation, the following approximation to the posterior density $\pi(\boldsymbol{\beta}|\boldsymbol{y})$ could be considered to obtain an approximate CI:

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) \approx \boldsymbol{g}\left(\boldsymbol{\beta} \ : \ \hat{\boldsymbol{\beta}}, \ V(\hat{\boldsymbol{\beta}})\right), \tag{15}$$

where $\boldsymbol{g}(\cdot\ ; \hat{\boldsymbol{\beta}}, \ V(\hat{\boldsymbol{\beta}}))$ denotes the multivariate normal (or Gaussian) density function with mean vector $\hat{\boldsymbol{\beta}}$ and covariance matrix $V(\hat{\boldsymbol{\beta}})$, herein $\hat{\boldsymbol{\beta}}$ is the posterior mode and $V(\hat{\boldsymbol{\beta}})$ is the

inverse of the negative Hessian of $\log\{\pi(\boldsymbol{\beta}|\boldsymbol{y})\}$ evaluated at $\hat{\boldsymbol{\beta}}$. Unfortunately, the Laplace approximation is not applicable to ℓ_1 distance–based–approaches such as the lasso methods, because the corresponding prior density is not differentiable. To satisfy the differentiability, we define a new distance as follows:

$$d_c(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{p} \lambda_j \frac{c(x_j - y_j)^2}{\sqrt{1 + c^2(x_j - y_j)^2}}, \tag{16}$$

where $c > 0$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$. Using the defined distance, we introduce a new prior for $\boldsymbol{\beta}$, called Gaussian and double exponential (GD) prior, as follows:

$$\pi_{\mathrm{GD}}(\boldsymbol{\beta}) \;\propto\; \exp\{-d_c(\boldsymbol{\beta}, \mathbf{0}_p)\} = \exp\left\{-\sum_{j=1}^{p} \lambda_j \frac{c\beta_j^2}{\sqrt{1 + c^2\beta_j^2}}\right\}, \tag{17}$$

where $c > 0$ and $\lambda_j > 0$. The GD prior has many attractive properties which are given as follows:

P1. The GD prior is proper due to the following fact:

$$\exp\left\{-\lambda \frac{c\beta^2}{\sqrt{1 + c^2\beta^2}}\right\} < \exp\left\{-\lambda \frac{c\beta^2}{\sqrt{1 + c^2}}\right\} \mathbf{1}\{\beta^2 \le 1\} + \exp\left\{-\lambda \frac{c|\beta|}{\sqrt{1 + c^2}}\right\} \mathbf{1}\{\beta^2 > 1\},$$

where $\lambda > 0$, $c > 0$ and $\mathbf{1}\{A\}$ denotes the indicator function for the set A.

P2. The negative of the log GD prior is differentiable and strictly convex; the proof to check the convexity and differentiability is straightforward.

P3. For given λ_j's, if c is sufficiently large, then the GD prior is analogous to the Laplace (or double exponential) prior due to the fact that, for given $\lambda(> 0)$,

$$\lambda \frac{c\beta^2}{\sqrt{1 + c^2\beta^2}} \longrightarrow \lambda|\beta| \quad \text{as} \quad c \to \infty. \tag{18}$$

P4. Define $\lambda = \frac{1}{2\sigma^2 c}$, if c is sufficiently small, then the GD prior is analogous to the Gaussian prior as follows:

$$\lambda \frac{c\beta^2}{\sqrt{1+c^2\beta^2}} = \frac{1}{2\sigma^2 c} \frac{c\beta^2}{\sqrt{1+c^2\beta^2}} \longrightarrow \frac{1}{2\sigma^2}\beta^2 \quad \text{as} \quad c \downarrow 0. \tag{19}$$

One may recognize that the name "GD prior" comes from the last two properties. To illustrate the aforementioned properties, we draw the plots of the univariate density function of GD prior corresponding to $c = 0.01, 0.1, 1, 10, 100$ for given $\lambda = 1$ in Fig.1. Note that when $c = 100$, it is analogous to the Laplace density function (Fig.1a), but the peak is indeed smooth (Fig.1b). Hence, the GD prior with a sufficiently large c produces an approximation of the Bayesian lasso. While the posterior density in the Bayesian lasso is non-differentiable, the posterior induced by the GD prior is differentiable.

An extension to the multivariate GD prior is easily obtained as follows:

$$\pi_{\text{mGD}}(\boldsymbol{\beta}) \quad \propto \quad \exp\left\{-\lambda \frac{c\boldsymbol{\beta}^{\text{T}}\boldsymbol{W}\boldsymbol{\beta}}{\sqrt{1+c^2\boldsymbol{\beta}^{\text{T}}\boldsymbol{W}\boldsymbol{\beta}}}\right\}, \tag{20}$$

where $c > 0$, $\lambda > 0$, and $\boldsymbol{W}$ is a nonnegative–definite matrix. Note that the multivariate GD prior is proper, differentiable, and strictly convex as in P1 and P2. Furthermore, P3 and P4 are naturally extended to multivariate Gaussian prior and multivariate Laplace prior, respectively.

4 Hierarchical Bayesian Modeling using Bregman Divergence

In this section, we introduce a novel approach to sparse high–dimensional problems based on the Bregman divergence with certain convex functions in a hierarchical Bayesian framework. Since the Bregman divergence induces smooth (differentiable) loss functions, all parame-

(a) $\beta \in (-10, 10)$ (b) $\beta \in (-0.1, 0.1)$

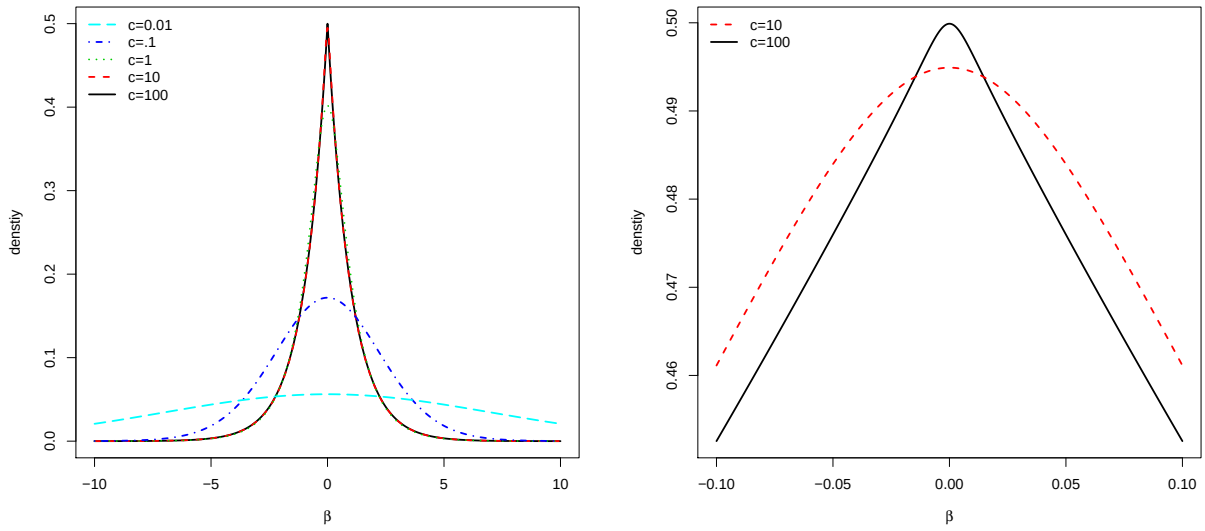

Figure 1: Plots of GD prior for $c = 0.01, 0.1, 1, 10, 100$ (a) and augmented plots of GD prior for $c = 10, 100$ (b).

ter estimates including the regularization parameters can be easily obtained via Newton–Raphson type algorithm. In addition, we extend the proposed method to generalized linear models (GLMs) as well as in the group lasso scenario (Yuan and Lin, 2007).

4.1 Linear Regression Models

Recall the linear regression model in (1):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n), \tag{21}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^{\mathrm{T}}$, and $\boldsymbol{x}_i = (x_{i,1} \ldots, x_{i,p})^{\mathrm{T}}$. For the sake of simplicity, we assume $\sigma^2 = 1$. Recall that, using the duality property, the

likelihood function can be expressed as

$$f(\boldsymbol{y}|\boldsymbol{\beta}) \propto \exp\left\{-\mathrm{BD}_{\phi_f}\left(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}\right)\right\}, \tag{22}$$

where $\mathrm{BD}_{\phi_f}(\cdot,\cdot)$ is the Bregman divergence with $\phi_f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$. Using the GD prior in (17) for $\boldsymbol{\beta}$ and gamma prior for $\boldsymbol{\lambda}$, we define the following prior for $(\boldsymbol{\beta}, \boldsymbol{\lambda})$:

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}) \;\propto\; \prod_{j=1}^{p} \lambda_j^{a_0} \exp\left\{-\lambda_j\left(b_0 + \frac{c\beta_j^2}{\sqrt{1+c^2\beta_j^2}}\right)\right\}, \tag{23}$$

for $c > 0$, $a_0 > 0$, and $b_0 > 0$. Again, using the duality of the Bregman divergence, the prior density can be expressed as

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}) \;\propto\; \exp\left\{-\mathrm{BD}_{\phi_{\pi_1}}(\boldsymbol{\beta}, \mathbf{0}_p) - \mathrm{BD}_{\phi_{\pi_2}}\left(\boldsymbol{\lambda}, \frac{\boldsymbol{a}_0}{\boldsymbol{b}_0}\right)\right\}, \tag{24}$$

where $\phi_{\pi_1}(\boldsymbol{x}) = \sum_{j=1}^{p} \lambda_j \frac{cx_j^2}{\sqrt{1+c^2x_j^2}}$ and $\phi_{\pi_2}(\boldsymbol{x}) = -\sum_{j=1}^{p} a_0 \log(x_j)$. From (22) and (24), the posterior density function is given as follows:

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{y}) \;&\propto\; f(\boldsymbol{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}) \\
&\propto\; \exp\left\{-\mathrm{BD}_{\phi_f}\left(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}\right) - \mathrm{BD}_{\phi_{\pi_1}}(\boldsymbol{\beta}, \mathbf{0}_p) - \mathrm{BD}_{\phi_{\pi_2}}\left(\boldsymbol{\lambda}, \frac{\boldsymbol{a}_0}{\boldsymbol{b}_0}\right)\right\},
\end{aligned} \tag{25}$$

where $\phi_f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$, $\phi_{\pi_1}(\boldsymbol{x}) = \sum_{j=1}^{p} \lambda_j \frac{cx_j^2}{\sqrt{1+c^2x_j^2}}$ and $\phi_{\pi_2}(\boldsymbol{x}) = -\sum_{j=1}^{p} a_0 \log(x_j)$. Note that the proposed model is approximately equal to the Bayesian lasso for large enough c and the Bayesian regression model with non-informative (or flat) prior for small enough c. Using the fact that the full posterior distribution is differentiable, the MAP estimates of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ can be easily obtained by the following iterative algorithm:

Algorithm 1 (Linear Regression Model)**.** *Set initial values* $\left(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\lambda}}^{(0)}\right)$.

1. *Update*

$$\hat{\lambda}_j^{(t+1)} \leftarrow a_0 \left(b_0 + \frac{c\left(\hat{\beta}_j^{(t)}\right)^2}{\sqrt{1 + c^2\left(\hat{\beta}_j^{(t)}\right)^2}} \right)^{-1}, \qquad j = 1, \ldots, p,$$

$$\hat{\boldsymbol{\beta}}^{(t+1)} \leftarrow \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + diag_p \left\{ \hat{\lambda}_j^{(t+1)} \frac{2c + c^3\left(\hat{\beta}_j^{(t)}\right)^2}{\left(1 + c^2\left(\hat{\beta}_j^{(t)}\right)^2\right)^{\frac{3}{2}}} \right\} \right]^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y},$$

where $diag_p\{d_j\}$ *denotes the* $(p \times p)$ *diagonal matrix with the* j^{th} *diagonal element* d_j.

2. *Repeat until convergence.*

Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})$ be the MAP estimates. Since $\hat{\boldsymbol{\beta}}$ is the posterior mode and the posterior is twice differentiable, using the Laplace approximation, we can derive the following approximation to the conditional posterior density of $\boldsymbol{\beta}$ given $\hat{\boldsymbol{\lambda}}$, $\pi(\boldsymbol{\beta}|\boldsymbol{y}, \hat{\boldsymbol{\lambda}})$, as follows:

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}, \hat{\boldsymbol{\lambda}}) \approx \boldsymbol{g}\left(\boldsymbol{\beta} \ : \ \hat{\boldsymbol{\beta}}, \ V_\lambda(\hat{\boldsymbol{\beta}})\right), \tag{26}$$

where $V_\lambda(\hat{\boldsymbol{\beta}})$ is the inverse of the negative Hessian of $\log\left\{\pi(\boldsymbol{\beta}|\boldsymbol{y}, \hat{\boldsymbol{\lambda}})\right\}$ evaluated at $\hat{\boldsymbol{\beta}}$. Hence the approximate $100 \times (1-\alpha)\%$ credible set for $\boldsymbol{\beta}$ can be easily constructed using (26).

4.2 Extensions of Likelihood Function

Banerjee et al. (2005) showed that every distribution in the natural exponential family corresponds to a unique and distinct Bregman divergence (one-to-one mapping). Due to the bijection between the natural exponential family and a class of Bregman divergences, we can easily extend the proposed method to GLMs. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the likelihood function for a natural exponential family, which is given as

$$f(\boldsymbol{y}|\boldsymbol{\theta}) \propto \exp\left\{\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{y} - \xi(\boldsymbol{\theta})\right\}, \tag{27}$$

where $\xi(\cdot)$ is a differentiable and strictly convex function. Let ϕ be a conjugate of ξ. By the Legendre transformation, ξ can be expressed as

$$\xi(\boldsymbol{\theta}) = \langle \nabla\xi(\boldsymbol{\theta}), \boldsymbol{\theta} \rangle - \phi\left(\nabla\xi(\boldsymbol{\theta})\right). \tag{28}$$

By (28) and the fact that $\boldsymbol{\theta} = \nabla\phi\left(\nabla\xi(\boldsymbol{\theta})\right)$, the likelihood function in (27) can be expressed as

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\theta}) &\propto \exp\left\{\langle \boldsymbol{y} - \nabla\xi(\boldsymbol{\theta}), \boldsymbol{\theta} \rangle + \phi\left(\nabla\xi(\boldsymbol{\theta})\right)\right\} \\ &\propto \exp\left\{-\mathrm{BD}_{\phi}(\boldsymbol{y}, \boldsymbol{\mu})\right\}, \end{aligned} \tag{29}$$

where $\boldsymbol{\mu} = \nabla\xi(\boldsymbol{\theta})$. Note that $\boldsymbol{\mu} = E_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y})$ due to the property of the natural exponential family. Hence the extension to GLMs is accomplished by switching (22) with the following likelihood function:

$$f(\boldsymbol{y}|\boldsymbol{\beta}) \propto \exp\left\{-\mathrm{BD}_{\phi_f}\left(\boldsymbol{y}, \boldsymbol{h}\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right)\right\}, \tag{30}$$

where ϕ_f is a differentiable and strictly convex function and $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^n$ is a link function such that $\boldsymbol{h}(\cdot)$ is differentiable and $\boldsymbol{h}^{-1}(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}$. Note that the linear regression model is a special case of this extension when $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}$ and $\phi_f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$. Table 1 displays some examples for the bijection between the natural exponential family and the Bregman divergence.

Wedderburn (1974) discussed a connection between the log likelihood for exponential family and the quasi–likelihood function. Similarly, using the bijection between the exponential family and Bregman divergence, we have the following lemma, and it will be used later to prove some theoretical properties of our methods.

Table 1: Examples of the Bregman divergences generated by some convex functions and related distributions in the exponential family.

$\phi(\boldsymbol{x})$	$\mathrm{BD}_\phi(\boldsymbol{x}, \boldsymbol{\mu})$	Distribution
$\sum_{i=1}^n \left\{\frac{1}{2\sigma^2} x_i^2\right\}$	$\sum_{i=1}^n \left\{\frac{1}{2\sigma^2}(x_i - \mu_i)^2\right\}$	Gaussian
$\sum_{i=1}^n \{x_i \log x_i\}$	$\sum_{i=1}^n \left\{x_i \log\left(\frac{x_i}{\mu_i}\right) - (x_i - \mu_i)\right\}$	Poisson
$\sum_{i=1}^n \{-\log x_i\}$	$\sum_{i=1}^n \left\{\frac{x_i}{\mu_i} - \log\left(\frac{x_i}{\mu_i}\right) - 1\right\}$	Exponential
$\sum_{i=1}^n \{x_i \log x_i + (1-x_i)\log(1-x_i)\}$	$\sum_{i=1}^n \left\{x_i \log\left(\frac{x_i}{\mu_i}\right) + (1-x_i)\log\left(\frac{1-x_i}{1-\mu_i}\right)\right\}$	Bernoulli

Lemma 2. *If $\boldsymbol{y}$ is a random sample from natural exponential family, then the corresponding Bregman divergence is its negative quasi–likelihood function.*

4.3 Extensions of Prior Distribution

In addition, the divergence–based–approach can be easily extended to the (approximate) group lasso (Yuan and Lin, 2007) by introducing a new convex function in the Bregman divergence at the prior stage. Suppose that the p predictors can be divided into K groups. Without loss of generality, we assume that from the first, p_k predictors are in the k^{th} group, say $\boldsymbol{\beta}_k$, for $k = 1, \ldots, K$, i.e., $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_{p_1})^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (\beta_{p_1+1}, \ldots, \beta_{p_1+p_2})^{\mathrm{T}}$, $\ldots$, $\boldsymbol{\beta}_K = (\beta_{\sum_{j=1}^{K-1} p_j + 1}, \ldots, \beta_{\sum_{j=1}^{K} p_j})^{\mathrm{T}}$, where $\sum_{j=1}^k p_j = p$. In the group lasso set–up, the prior is defined as

$$\pi_{\text{G–lasso}}(\boldsymbol{\beta}) \propto \exp\left\{-\sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|\right\}, \tag{31}$$

where $\lambda_k > 0$. As an alternative, using the multivariate GD prior in (20), we define a new prior as follows:

$$\pi_{\mathrm{mGD}}(\boldsymbol{\beta}) \propto \exp\left\{-\sum_{k=1}^{K} \lambda_k \frac{c\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k}{\sqrt{1+c^2\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k}}\right\}, \tag{32}$$

where $c > 0$. Note that

$$\sum_{k=1}^{K} \lambda_k \frac{c\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k}{\sqrt{1+c^2\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k}} \longrightarrow \sum_{k=1}^{K} \lambda_k \sqrt{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k} = \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|,$$

as $c \to \infty$. Hence the new prior is an approximation of the group lasso prior when c is sufficiently large, and it is easy to show that $\pi_{\mathrm{mGD}}(\boldsymbol{\beta})$ can be expressed in terms of Bregman divergence with the convex function

$$\phi_{\pi_1}(\boldsymbol{x}) = \sum_{k=1}^{K} \lambda_k \frac{c\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{x}_k}{\sqrt{1+c^2\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{x}_k}}. \tag{33}$$

Consequently, our proposed model in (25) is then extended to the (approximate) Bayesian group lasso by using (33).

4.4 Extensions of Algorithm

In this section, we discuss general iterative algorithms to find the MAP estimate for the Bregman divergence–based–model. Using the link function $\boldsymbol{h}(\cdot)$ in (30), the full posterior density in (25) can be generalized as

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{y}) \propto \exp\left\{-\mathrm{BD}_{\phi_f}\left(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta})\right) - \mathrm{BD}_{\phi_{\pi_1}}(\boldsymbol{\beta}, \mathbf{0}_p) - \mathrm{BD}_{\phi_{\pi_2}}(\boldsymbol{\lambda}, \boldsymbol{\tau})\right\}, \tag{34}$$

where $\boldsymbol{\tau}$ is a p–dimensional constant vector and ϕ_f, ϕ_{π_1} and ϕ_{π_2} are differentiable and strictly convex, but unspecified. Due to the smoothness (or differentiability) of the Bregman divergence and the link function, the MAP estimates of (34), say $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})$, can be easily obtained by the following Newton–Raphson type algorithm:

Algorithm 2. *Set initial values* $\left(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\lambda}}^{(0)}\right)$.

1. *Update*

$$\begin{aligned}
\hat{\boldsymbol{\lambda}}^{(t+1)} &\leftarrow \arg\min_{\boldsymbol{\lambda}} \Psi(\hat{\boldsymbol{\beta}}^{(t)}, \boldsymbol{\lambda}), \\
\hat{\boldsymbol{\beta}}^{(t+1)} &\leftarrow \hat{\boldsymbol{\beta}}^{(t)} - \left[\frac{\partial^2 \Psi(\boldsymbol{\beta}, \hat{\boldsymbol{\lambda}}^{(t+1)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(t)}}\right]^{-1} \left(\frac{\partial \Psi(\boldsymbol{\beta}, \hat{\boldsymbol{\lambda}}^{(t+1)})}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(t)}}\right),
\end{aligned}$$

where $\Psi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = BD_{\phi_f}\left(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta})\right) + BD_{\phi_{\pi_1}}(\boldsymbol{\beta}, \mathbf{0}) + BD_{\phi_{\pi_2}}(\boldsymbol{\lambda}, \boldsymbol{\tau})$.

2. *Repeat until convergence.*

For GD priors, however, when c is very large, the convergence of the above algorithm could be very slow around the neighborhood of the posterior mode due to the fact that the curvature around the peak is very small; see Fig 1b when $c = 100$. In such cases, to avoid the slow convergence, we consider the following $\boldsymbol{\epsilon}$–neighborhood of MAP estimate:

$$\hat{\boldsymbol{\beta}}_\epsilon \quad \text{such that} \quad \frac{\partial \pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{y})}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\epsilon} = \boldsymbol{\epsilon}, \tag{35}$$

where $\boldsymbol{\epsilon} > 0$. Note that $\hat{\boldsymbol{\beta}}_\epsilon \longrightarrow \hat{\boldsymbol{\beta}}$ as $\boldsymbol{\epsilon} \downarrow \mathbf{0}$. Using the Newton–Raphson method, the $\boldsymbol{\epsilon}$–neighborhood of MAP estimate can be obtained via the following algorithm:

Algorithm 3. *Set initial values* $\left(\hat{\boldsymbol{\beta}}_\epsilon^{(0)}, \hat{\boldsymbol{\lambda}}^{(0)}\right)$.

1. *Update*

$$\begin{aligned}
\hat{\boldsymbol{\lambda}}^{(t+1)} &\leftarrow \arg\min_{\boldsymbol{\lambda}} \Psi(\hat{\boldsymbol{\beta}}_\epsilon^{(t)}, \boldsymbol{\lambda}), \\
\hat{\boldsymbol{\beta}}_\epsilon^{(t+1)} &\leftarrow \hat{\boldsymbol{\beta}}^{(t)} - \left[\frac{\partial^2 \Psi(\boldsymbol{\beta}, \hat{\boldsymbol{\lambda}}^{(t+1)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\epsilon^{(t)}}\right]^{-1} \left(\frac{\partial \Psi(\boldsymbol{\beta}, \hat{\boldsymbol{\lambda}}^{(t+1)})}{\partial \boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\epsilon^{(t)}} - \boldsymbol{\epsilon}\right),
\end{aligned}$$

where $\Psi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = BD_{\phi_f}\left(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta})\right) + BD_{\phi_{\pi_1}}(\boldsymbol{\beta}, \mathbf{0}) + BD_{\phi_{\pi_2}}(\boldsymbol{\lambda}, \boldsymbol{\tau})$.

2. Repeat until convergence.

Note that

$$\left[H(\hat{\boldsymbol{\beta}}_\epsilon) + H(\hat{\boldsymbol{\beta}}_{-\epsilon})\right]^{-1} \left(H(\hat{\boldsymbol{\beta}}_\epsilon)\hat{\boldsymbol{\beta}}_\epsilon + H(\hat{\boldsymbol{\beta}}_{-\epsilon})\hat{\boldsymbol{\beta}}_{-\epsilon}\right) \longrightarrow \hat{\boldsymbol{\beta}}, \tag{36}$$

as $\boldsymbol{\epsilon} \downarrow \mathbf{0}$, where $H(\boldsymbol{\beta}_0)$ is the negative Hessian of $\log\left\{\pi(\boldsymbol{\beta}|\boldsymbol{y}, \hat{\boldsymbol{\lambda}})\right\}$ evaluated at $\boldsymbol{\beta}_0$. Hence the MAP estimate can be approximated by

$$\hat{\boldsymbol{\beta}}_H = \left[H(\hat{\boldsymbol{\beta}}_\epsilon) + H(\hat{\boldsymbol{\beta}}_{-\epsilon})\right]^{-1} \left(H(\hat{\boldsymbol{\beta}}_\epsilon)\hat{\boldsymbol{\beta}}_\epsilon + H(\hat{\boldsymbol{\beta}}_{-\epsilon})\hat{\boldsymbol{\beta}}_{-\epsilon}\right), \tag{37}$$

where $\boldsymbol{\epsilon} \approx \mathbf{0}$. Note that $\hat{\boldsymbol{\beta}}_H$ is a weighted mean of $\hat{\boldsymbol{\beta}}_\epsilon$ and $\hat{\boldsymbol{\beta}}_{-\epsilon}$ by their curvatures.

5 Ultra–High–Dimensional Problems

When the number of coefficients is much larger than the sample size, that is $p \gg n$, it is reasonable to conduct an appropriate pre–screening procedure such as Sure independence screening (Fan and Lv, 2008; Fan and Song, 2010) to reduce the ultra–high–dimensionality. Let $\mathcal{M}^* = \{j : \beta_j^* \neq 0\}$ be the index set of the true non-zero coefficients. Consider the following marginal likelihood function:

$$f(\boldsymbol{y}|\beta_j) = \exp\left\{-\mathrm{BD}_{\phi_f}(\boldsymbol{y}, h(X_j\beta_j))\right\}, \tag{38}$$

where X_j denotes the j^{th} column of $\boldsymbol{X}$. Define the maximum marginal likelihood estimator (MMLE) as

$$\tilde{\beta}_j^M = \arg\min_{\beta_j} \left[\frac{1}{n}\mathrm{BD}_{\phi_f}(\boldsymbol{y}, h(X_j\beta_j))\right], \qquad \text{for } j = 1, \ldots, p. \tag{39}$$

Let $\tilde{\mathcal{M}}_{t_n} := \{1 \leq j \leq p : |\tilde{\beta}_j^M| \geq t_n\}$ be the index set of selected variables based on the MMLE and a deterministic threshold t_n. Fan and Song (2010) have shown that if the partial

orthogonality assumption (i.e., $\{X_j : j \notin \mathcal{M}^*\}$ is independent of $\{X_i : i \in \mathcal{M}^*\}$) holds under some general conditions, then, for sufficiently large n, there exists the threshold t_n such that

$$\tilde{\mathcal{M}}_{t_n} \overset{a.s.}{=} \mathcal{M}^*. \tag{40}$$

Similarly, we consider a Bayesian screening method using the following marginal MAP estimates:

$$\hat{\beta}_j^M = \arg\min_{\beta_j} \left\{ \mathrm{BD}_{\phi_f}\left(\boldsymbol{y}, h(X_j\beta_j)\right) + \mathrm{BD}_{\phi_{\pi_1}}(\beta_j, 0) \right\}, \tag{41}$$

where $\phi_{\pi_1}(x_j) = \lambda_j \frac{cx_j^2}{\sqrt{1+c^2x_j^2}}$ for $j = 1, \ldots, p$. Define $\hat{\mathcal{M}}_{z_n} := \{1 \le j \le p : |\hat{\beta}_j^M| \ge z_n\}$, where z_n denotes a predefined threshold. From (40), it is easy to verify that, for sufficiently large n and small c, there exists a threshold z_n such that

$$\hat{\mathcal{M}}_{z_n} \overset{a.s.}{=} \mathcal{M}^*. \tag{42}$$

Note that if $c \downarrow 0$, then $\left|\hat{\beta}_j^M - \tilde{\beta}_j^M\right| \to 0$ for all j.

6 Bayesian Oracle Properties

In this section, we discuss the Bayesian oracle properties (i.e., posterior consistency and normality) of our main model given as

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}_n) \propto \exp\left\{ -\mathrm{BD}_{\phi_f}\left(\boldsymbol{y}_n, \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta})\right) - \mathrm{BD}_{\phi_{\pi_1}}(\boldsymbol{\beta}, \mathbf{0}_{p_n}) \right\}, \tag{43}$$

where ϕ_f is a differentiable and strictly convex function, $\phi_{\pi_1}(\boldsymbol{x}) = \sum_{j=1}^{p_n} \lambda_j \frac{cx_j^2}{\sqrt{1+c^2x_j^2}}$, $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^n$ is a one–to–one and differentiable link function, $\boldsymbol{y}_n = (y_1, \ldots, y_n)^{\mathrm{T}}$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_n})^{\mathrm{T}}$, where p_n means the dimension of the parameter space which is allowed to grow with the sample size n. However, we assume that the true number of non–zero coefficients

p^* is fixed. We further assume that the true parameter $\boldsymbol{\beta}^*$ is a interior point of compact and convex set $\mathcal{B}^* \in \mathbb{R}^{p_n}$. Our results rely on the following assumptions:

A1. The fisher information $\mathcal{I}_n(\boldsymbol{\beta}) = E_{Y_n|\boldsymbol{\beta}}\left[\frac{\partial^2 \mathrm{BD}_{\phi_f}(\boldsymbol{y}_n, \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\right]$ is finite and positive definite at $\boldsymbol{\beta}^*$, and $\sup_{\boldsymbol{\beta}\in\mathcal{B}^*} ||\mathcal{I}_n(\boldsymbol{\beta})^{1/2}\boldsymbol{x}|| < \infty$ subject to $||\boldsymbol{x}|| = 1$.

A2. There exists a positive constant K such that

$$\left|\mathrm{BD}_{\phi_f}\left(y, \boldsymbol{h}\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_1\right)\right) - \mathrm{BD}_{\phi_f}\left(y, \boldsymbol{h}\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_2\right)\right)\right| \mathbf{1}\{y \in \Omega_n\} \leq K\|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_1 - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_2\|\mathbf{1}\{y \in \Omega_n\},$$

for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{B}^*$, where $\Omega_n = \{y : |y| \leq K_n\}$ for some sufficiently large positive constant K_n. In addition, there exists $\delta_1 > 0$ such that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}^*, \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|\leq\delta_1} \left|E_{\boldsymbol{Y}_n|\boldsymbol{\beta}^*}\left[\mathrm{BD}_{\phi_f}\left(\boldsymbol{y}_n, \boldsymbol{h}\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right) - \mathrm{BD}_{\phi_f}\left(\boldsymbol{y}_n, \boldsymbol{h}\left(\boldsymbol{X}\boldsymbol{\beta}^*\right)\right)\right]\mathbf{1}\{y \notin \Omega_n\}\right| \leq o\left(\frac{p_n}{n}\right).$$

A3. $\mathrm{BD}_{\phi_f}(\boldsymbol{y}_n, h(\boldsymbol{X}\boldsymbol{\beta}))$ is convex in $\boldsymbol{\beta}$. Furthermore, there exists a positive constant M such that

$$E_{\boldsymbol{Y}_n|\boldsymbol{\beta}^*}\left|\mathrm{BD}_{\phi_f}\left(\boldsymbol{y}_n, \boldsymbol{h}\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right) - \mathrm{BD}_{\phi_f}\left(\boldsymbol{y}_n, \boldsymbol{h}\left(\boldsymbol{X}\boldsymbol{\beta}^*\right)\right)\right| \geq M\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2,$$

for all $\boldsymbol{\beta} \in \mathcal{B}^*$.

A4. $\sup_{1\leq j\leq p_n} |\beta_j^*| < \infty$, $\sup_{1\leq i\leq n} E_{Y_i|\boldsymbol{\beta}^*}|Y_i| < \infty$, and

$$0 < \Lambda_{\min}^* < \liminf_n \Lambda_{\min}/\sqrt{n} \leq \limsup_n \Lambda_{\max}/\sqrt{n} < \Lambda_{\max}^* < \infty,$$

where $\Lambda_{\min}$ and $\Lambda_{\max}$ denote the smallest and the largest singular values of $\boldsymbol{X}$, respectively.

A5. $p_n = o(n)$.

A6. $p_n^4 \log p_n = o(n)$.

Note that the assumptions are similar to that of Ghosal (1997) and Fan and Song (2010) and they are satisfied in many cases such as linear regression, logistic regression, Poisson regression, etc. For notational convenience, "P_{β^*}–almost surely" and "P_{β^*}–in probability", respectively, are written as "almost surely" and "in probability".

First, we show that the posterior of $\boldsymbol{\beta}$ is strongly consistent.

Theorem 1 (Consistency)**.** *Under the assumptions A1–A5, we have*

$$\Pi\left(\left\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| > n^{-\kappa}\right\} \mid \boldsymbol{y}_n\right) \longrightarrow 0 \qquad \textit{almost surely,} \tag{44}$$

as $n \to \infty$ *, where* $0 < \kappa < \frac{1}{2}$ *and* $\Pi(\cdot|\boldsymbol{y}_n)$ *denotes the posterior probability of (43).*

If the prior in (43) satisfies the Lipschitz conditions, we can directly quote Theorem 2.1 of Ghosal (1997) to discuss the posterior normality because our assumptions imply the necessary assumptions of the theorem. The following lemma ensures that our prior indeed satisfies the Lipschitz condition.

Lemma 3. *The independent GD prior satisfies the Lipschitz condition, that is, for some* $M,\ \delta_0,\ \eta_0 > 0$ *and for all* j*, we have*

$$\pi(\beta_j^*) > \eta_0 \quad \textit{and} \quad \left|\log \pi(\beta_j) - \log \pi(\beta_j^*)\right| \le M \left|\beta_j - \beta_j^*\right|, \tag{45}$$

whenever $\left|\beta_j - \beta_j^*\right| \le \delta_0$*, where* $\pi(\beta_j) \propto \exp\left\{-\frac{c\beta_j^2}{\sqrt{1+c^2\beta_j^2}}\right\}$.

Define $\boldsymbol{u} = \boldsymbol{B}_n^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$, where $\boldsymbol{B}_n = \left.\frac{\partial^2 \mathrm{BD}_{\phi_f}(\boldsymbol{y}_n, \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ and $\boldsymbol{B}_n^{1/2}$ is its positive definite square root matrix (i.e., $\boldsymbol{B}_n = \boldsymbol{B}_n^{1/2}\boldsymbol{B}_n^{1/2}$). Let $\pi(\boldsymbol{u}|\boldsymbol{y}_n)$ be the posterior density of $\boldsymbol{u}$. Then, the posterior normality can be argued by the following lemma.

Lemma 4 (Normality, Ghosal (1997)). *Under assumptions A1–A4 and A6,*

$$\int |\pi(\boldsymbol{u}|\boldsymbol{y}_n) - \boldsymbol{g}(\boldsymbol{u} : \boldsymbol{\mu}_n, \boldsymbol{I}_{p_n})|\, d\boldsymbol{u} \longrightarrow 0 \qquad \textit{in probability,} \tag{46}$$

where $\boldsymbol{\mu}_n = \boldsymbol{B}_n^{-1/2}\left[-\frac{\partial BD_{\phi_f}(\boldsymbol{y}_n, \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right]$.

7 Numerical Studies

In this section, we conduct numerical studies to compare our proposed GD method to the adaptive lasso (Zou, 2006). To obtain the adaptive lasso estimate, the *glmnet* package in R is used.

7.1 Simulation Studies

To validate our method, Monte Carlo experiments are conducted in this section. The adaptive lasso estimate is defined as

$$\hat{\boldsymbol{\beta}}_{\text{A–lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \parallel \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\parallel^2 + \lambda_0 \sum_{j=1}^{p} \tilde{w}_j |\beta_j| \right\}, \tag{47}$$

where $\lambda_0 > 0$, $\tilde{w}_j = 1/|\tilde{\beta}_j^{\text{ols}}|$, and $\tilde{\beta}^{\text{ols}}$ denotes the ordinary least square (ols) estimate. The parameter λ_0 is determined by 10–fold cross–validation in the adaptive lasso. For the GD method, the posterior of (25) is used with $c = (10)^4$, $a_0 = \lambda_0$ and $b_0 = 1$. Hence the MAP estimate is given as

$$\hat{\boldsymbol{\beta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \parallel \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\parallel^2 + \lambda_0 \sum_{j=1}^{p} \hat{w}_j \frac{\beta_j^2}{\sqrt{(0.1)^8 + \beta_j^2}} \right\}, \tag{48}$$

where $\hat{w}_j = 1 + \hat{\beta}_j^{\text{MAP}}$ and $\hat{\beta}_j^{\text{MAP}}$ denotes the j^{th} component of $\hat{\boldsymbol{\beta}}_{\text{MAP}}$. Note that $\frac{\beta_j^2}{\sqrt{(0.1)^8+\beta_j^2}} \approx |\beta_j|$.

In each simulation, we simulate 1,000 data sets from the model $y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i$ for $i = 1, \ldots, n$, where the components of $\boldsymbol{x}_i$ are standard normal with $corr(x_{ij_1}, x_{ij_2}) = (0.5)^{|j_1 - j_2|}$, $\epsilon \overset{iid}{\sim} \mathrm{N}(0,1)$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ such that

$$\beta_j \;=\; (-1)^{u_j}(\alpha_n + |z_j|)\mathbf{1}_{\{1 \le j \le p^*\}}, \tag{49}$$

for $j = 1, \ldots, p$, where $u_j \overset{iid}{\sim}$ Bernoulli(0.4), $z_j \overset{iid}{\sim} \mathrm{N}(0,1)$ and $\alpha_n = 4\log(n)/\sqrt{n}$ with $n = 50, 100, 150$ and $p = 49, 99, 149$ for varying p^* given in Table 2. To compare, mean squared errors (MSEs) are estimated via the Monte Carlo method from the 1,000 replications. The results in Table 2 show that the GD method always performs better than the adaptive lasso in this setting.

Table 2: Estimated MSEs for the Adaptive Lasso and the GD method.

(n, p, p^*)	Adaptive lasso	GD method
(50,49,3)	0.01340	0.00763
(100,99,5)	0.00559	0.00263
(150,149,8)	0.00404	0.00183
(50,49,5)	0.02164	0.01215
(100,99,10)	0.01095	0.00550
(150,149,15)	0.00713	0.00346
(50,49,10)	0.04813	0.02926
(100,99,20)	0.02299	0.01299
(150,149,30)	0.01496	0.00815

7.2 Colon Tissue Data

In this section, we compare the prediction (or classification) performance of the GD method to the adaptive lasso using *colon tissue data* (Alon et al., 1999); the data set is available at *plsgenomics* package in R. This data set consists of 62 samples with 2,000 genes. For the i^{th} observation, the response variable y_i is a binary outcome, indicating the status of colon tissue (normal or tumor) and the predictor vector $\boldsymbol{x}_i$ gives the expression levels of 2000 genes, here we standardize the expression levels of all genes. Define the success (=tumor) probability of i^{th} tissue as p_i =Probability(y_i = tumor). The link function $\boldsymbol{h}$ in (30) is defined as $p_i = \boldsymbol{h}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) = [1 + \exp(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})]^{-1}$ (i.e., logit link). We split the data randomly into two partitions of 31 training samples and 31 testing samples, so that the training set can be used to build a model and the test set can be used to measure its prediction performance. First, we conduct a pre–screening procedure to reduce the ultra–high–dimensionality using the proposed Bayesian screening method with $c = 0.1$ and $\lambda_j = 0.1$ for all j. In this set–up, our screening method is (approximately) equivalent to the MMLE method of Fan and Song (2010). Using the screening procedure we rank all genes by the magnitude of marginal estimates and select the top 100 genes (i.e., $p = 100$). In the GD method, we set $c = 100$, $a_0 = 3$, and $b_0 = 1$. In the adaptive lasso, the regularization parameter λ is determined by the 10–fold cross validation as in the previous simulation study, but the classical lasso estimate used as the weight. To predict the outcomes, $\hat{p}_i = 0.5$ is used as the threshold. Table 3 presents the contingency tables of both methods. According to Table 3, we know that the sensitivity (77.27%) and specificity (88.89%) of the GD method are higher than the adaptive lasso (72.73%, 77.78%). Consequently, it demonstrates that the classification

model based on the GD method performs better than the adaptive lasso.

Table 3: Contingency tables.

Methods	Adaptive lasso		GD method	
True\ Prediction	Normal	Tumor	Normal	Tumor
Normal	7	2	8	1
Tumor	6	16	5	17
Specificity vs Sensitivity (%)	77.78	72.73	88.89	77.27
Misclassification Rate (%)		25.81		19.35

8 Concluding Remarks

From a Bayesian perspective, we have proposed a divergence–based–approach to sparse high–dimensional data using the Bregman divergence. Furthermore, the GD prior has been introduced along with its attractive properties. The numerical studies demonstrate that the new method indeed performs better than the original adaptive lasso method. In addition, the proposed method has been extended to several directions such as GLMs and the group lasso.

The determination of tuning parameter, in the PLE method, is very important because it controls the sparsity of the parameter estimates. However, the hyper–parameter, which corresponds to the tuning parameter in PLE method, has been roughly determined in this paper because the coefficient estimate does not need to be exactly zero in our method due to the use of the credible interval. Nevertheless, the determination of hyper–parameter should

be investigated to obtain more accurate parameter estimate, which will be studied in future.

One of the advantages of divergence–based–approach is that many extensions can be easily accomplished by assigning new divergence measures in the likelihood function and/or prior density function. For example, our model can be adapted to multivariate regression models or sparse variance–covariance estimation by using Bregman matrix divergence (Kulis et al., 2009) in the likelihood function. Since the Gaussian process regression model or the wavelet model (Ray and Mallick, 2006) are also expressed in terms of the Bregman matrix divergence due to the duality property, functional data can be handled by this extended framework, which is a work in progress by the authors. In addition, due to the fact that the mixture of the GD prior and Gaussian prior is differentiable and strictly convex, the mixture can be used to extend our model to the elastic–net (Zou and Hastie, 2005) which is more powerful tool than the lasso methods for ultra–high–dimensional problem.

APPENDIX: PROOFS

Proof of lemma 2

Let $\boldsymbol{y}$ be a random sample from $f(\boldsymbol{y}|\boldsymbol{\theta}) \propto \exp\left\{\boldsymbol{\theta}^T\boldsymbol{y} - \xi(\boldsymbol{\theta})\right\}$. Then the corresponding Bregman divergence is given as

$$\mathrm{BD}_\phi(\boldsymbol{y}, \boldsymbol{\mu}) = \phi(\boldsymbol{y}) - \phi(\boldsymbol{\mu}) - \langle \boldsymbol{y} - \boldsymbol{\mu}, \nabla_\mu \phi(\boldsymbol{\mu}) \rangle \,, \tag{50}$$

where $\boldsymbol{\mu} = E_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y})$ and ∇_μ denotes the gradient with respect to $\boldsymbol{\mu}$. By differentiating (50) with respect to $\boldsymbol{\mu}$, we have

$$\frac{\partial \mathrm{BD}_\phi(\boldsymbol{y}, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = -\nabla^2_\mu \phi(\boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu}),$$

where ∇^2_μ denotes the Hessian matrix with respect to $\boldsymbol{\mu}$. The following fact completes the proof:

$$\nabla^2_\mu \phi(\boldsymbol{\mu}) \;=\; \frac{\partial \nabla_\mu \phi(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}^T} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}^{\mathrm{T}}} = \left[\frac{\partial \boldsymbol{\mu}^{\mathrm{T}}}{\partial \boldsymbol{\theta}}\right]^{-1} = \left[\nabla^2_\theta \xi(\boldsymbol{\theta})\right]^{-1} = \left[V_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y})\right]^{-1}.$$

Proof of lemma 3

Since $|\beta_j^*| < \infty$, it is straightforward to check the first argument. Hence, we only need to show the second argument. Suppose $\left|\beta_j - \beta_j^*\right| \leq \delta_0$. By the triangle inequality, it implies that $|\beta_j| \leq \delta_0 + \left|\beta_j^*\right|$. Due to the fact that $\frac{cx^2}{\sqrt{1+c^2x^2}} \leq cx^2$ for any $c > 0$, we have

$$\begin{aligned}
\left|\log \pi(\beta_j) - \log \pi(\beta_j^*)\right| &= \left| \frac{\lambda_j c \beta_j^{*2}}{\sqrt{1 + c^2 \beta_j^{*2}}} - \frac{\lambda_j c \beta_j^2}{\sqrt{1 + c^2 \beta_j^2}} \right| \\
&\leq \left|\lambda_j c \beta_j^{*2} - \lambda_j c \beta_j^2\right| \\
&\leq \lambda_j c \left|\beta_j^* + \beta_j\right| \left|\beta_j^* - \beta_j\right| \\
&\leq \lambda_j c (2\left|\beta_j^*\right| + \delta_0) \left|\beta_j^* - \beta_j\right|.
\end{aligned}$$

By taking $M = \lambda_j c(2\left|\beta_j^*\right| + \delta_0)$, this completes the proof.

Proof of theorem 1

The major technique of our proof is analogous to Armagan et al. (2013).

Let $\Phi_n(\boldsymbol{y}_n) := \mathbf{1}(\boldsymbol{y}_n \in \mathcal{C}_n)$ be a test function, where $\mathcal{C}_n = \{\boldsymbol{y}_n : \|\tilde{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}^*\| > \frac{n^{-\kappa}}{2}\}$ and $\tilde{\boldsymbol{\beta}}_Q = \arg\min_{\boldsymbol{\beta}} \left\{\mathrm{BD}_{\phi_f}\left(\boldsymbol{y}_n, \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta})\right)\right\}$. Note that $\tilde{\boldsymbol{\beta}}_Q$ is the quasi–maximum likelihood estimator (QMLE) due to Lemma 2. Hence, according to Theorem 1 of Fan and Song (2010), we have

$$E_{\boldsymbol{Y}_n|\boldsymbol{\beta}^*}\left[\Phi_n(\boldsymbol{y}_n)\right] = \Pr\left\{\mathcal{C}_n|\boldsymbol{\beta}^*\right\} \leq \exp\left(-c_4 n^{1-2\kappa}\right), \tag{51}$$

where c_4 is a positive constant. Define $\mathcal{B}_n = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| > n^{-\kappa}\}$ for given κ. Then, the posterior probability of $\mathcal{B}_n$ can be expressed as

$$\begin{aligned}\Pi\left(\mathcal{B}_n \mid \boldsymbol{y}_n\right) &= \frac{\int_{\mathcal{B}_n}\{f(\boldsymbol{y}_n|\boldsymbol{\beta})/f(\boldsymbol{y}_n|\boldsymbol{\beta}^*)\}\Pi(d\boldsymbol{\beta})}{\int_{\mathbb{R}^{p_n}}\{f(\boldsymbol{y}_n|\boldsymbol{\beta})/f(\boldsymbol{y}_n|\boldsymbol{\beta}^*)\}\Pi(d\boldsymbol{\beta})} \\ &\leq \Phi_n + \frac{(1-\Phi_n)J_{\mathcal{B}_n}}{J_n} \quad (:= I_1 + I_2/J_n),\end{aligned}$$

where $J_{\mathcal{B}_n} = \int_{\mathcal{B}_n}\{f(\boldsymbol{y}_n|\boldsymbol{\beta})/f(\boldsymbol{y}_n|\boldsymbol{\beta}^*)\}\Pi(d\boldsymbol{\beta})$ and $J_n = J_{\mathbb{R}^{p_n}}$. Now, we need to show $I_1 + I_2/J_n \to 0$ almost surely as $n \to \infty$. By (51) and Markov's inequality, for sufficiently large n, we have

$$\Pr\left\{I_1 \geq \exp(-c_4 n^{1-2\kappa}/2)|\boldsymbol{\beta}^*\right\} \leq E_{\boldsymbol{Y}_n|\boldsymbol{\beta}^*}\left(\Phi_n\right)\exp(c_4 n^{1-2\kappa}/2) \leq \exp(-c_4 n^{1-2\kappa}/2).$$

Since $\sum_{n=1}^{\infty}\exp(-c_4 n^{1-2\kappa}/2) < \infty$, by Borel–Cantelli lemma, we have

$$\Pr\left\{I_1 \geq \exp(-c_4 n^{1-2\kappa}/2) \text{ infinitely often} \mid \boldsymbol{\beta}^*\right\} = 0. \tag{52}$$

To investigate the behavior of I_2, it is worth noting that

$$\begin{aligned}E_{\boldsymbol{Y}_n|\boldsymbol{\beta}^*}\left(I_2\right) &= E_{\boldsymbol{Y}_n|\boldsymbol{\beta}^*}\left\{(1-\Phi_n)J_{\mathcal{B}_n}\right\} \\ &= E_{\boldsymbol{Y}_n|\boldsymbol{\beta}^*}\left\{\mathbf{1}\left(\boldsymbol{y}_n \notin \mathcal{C}_n\right)\int_{\mathcal{B}_n}\{f(\boldsymbol{y}_n|\boldsymbol{\beta})/f(\boldsymbol{y}_n|\boldsymbol{\beta}^*)\}\Pi(d\boldsymbol{\beta})\right\} \\ &= \int_{\mathcal{B}_n}\left\{\int_{\mathbb{R}^n}\mathbf{1}\left(\{\boldsymbol{y}_n : \|\tilde{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}^*\| \leq n^{-\kappa}/2\}\right)f(\boldsymbol{y}_n|\boldsymbol{\beta})d\boldsymbol{y}_n\right\}\Pi(d\boldsymbol{\beta}) \\ &\leq \int_{\mathcal{B}_n}\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\left\{\int_{\mathbb{R}^n}\mathbf{1}\left(\{\boldsymbol{y}_n : \|\tilde{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}^*\| \leq n^{-\kappa}/2\}\right)f(\boldsymbol{y}_n|\boldsymbol{\beta})d\boldsymbol{y}_n\right\}\Pi(d\boldsymbol{\beta}) \\ &= \Pi_n(\mathcal{B}_n)\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\left\{\int_{\mathbb{R}^n}\mathbf{1}\left(\{\boldsymbol{y}_n : \|\tilde{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}^*\| \leq n^{-\kappa}/2\}\right)f(\boldsymbol{y}_n|\boldsymbol{\beta})d\boldsymbol{y}_n\right\} \\ &\leq \sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\left\{\int_{\mathbb{R}^n}\mathbf{1}\left(\{\boldsymbol{y}_n : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| - \|\tilde{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}\| \leq n^{-\kappa}/2\}\right)f(\boldsymbol{y}_n|\boldsymbol{\beta})d\boldsymbol{y}_n\right\} \\ &= \int_{\mathbb{R}^n}\mathbf{1}\left(\{\boldsymbol{y}_n : \|\tilde{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}\| \geq n^{-\kappa}/2\}\right)f(\boldsymbol{y}_n|\boldsymbol{\beta})d\boldsymbol{y}_n \\ &= \int_{\mathbb{R}^n}\mathbf{1}\left(\{\boldsymbol{y}_n : \|\tilde{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}^*\| \geq n^{-\kappa}/2\}\right)f(\boldsymbol{y}_n|\boldsymbol{\beta}^*)d\boldsymbol{y}_n \\ &= \Pr\left\{\mathcal{C}_n|\boldsymbol{\beta}^*\right\} \leq \exp\left(-c_4 n^{1-2\kappa}\right).\end{aligned}$$

Hence, for sufficiently large n, we have

$$\Pr\left\{I_2 \geq \exp(-c_4 n^{1-2\kappa}/2)|\boldsymbol{\beta}^*\right\} \leq E_{\boldsymbol{Y}_n|\boldsymbol{\beta}^*}\left(I_2\right)\exp(c_4 n^{1-2\kappa}/2) \leq \exp(-c_4 n^{1-2\kappa}/2).$$

Since $\sum_{n=1}^{\infty}\exp(-c_4 n^{1-2\kappa}/2) < \infty$, hence by Borel–Cantelli lemma

$$\Pr\left\{I_2 \geq \exp(-c_4 n^{1-2\kappa}/2) \text{ infinitely often } | \ \boldsymbol{\beta}^*\right\} = 0. \tag{53}$$

To complete the proof, we need to show that $\exp(c_4 n^{1-2\kappa}/2)J_n \to \infty$ as $n \to \infty$. Hence, it is enough to show that, for $0 < c_5 < c_4/2$, $J_n \geq \exp\left(-c_5 n^{1-2\kappa}\right)$ for sufficiently large n. Define $\mathcal{D}_n := \left\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| < \frac{c_5 n^{1-2\kappa}}{2K\Lambda_{\max}}\right\}$. By the assumption A2, we have

$$\begin{aligned}
J_n &= \int_{\mathbb{R}^{p_n}} \exp\left\{-\mathrm{BD}_{\phi_f}\left(\boldsymbol{y}_n, \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta}^*)\right) + \mathrm{BD}_{\phi_f}\left(\boldsymbol{y}_n, \boldsymbol{h}(\boldsymbol{X}\boldsymbol{\beta})\right)\right\}\Pi(d\boldsymbol{\beta}) \\
&\geq \int_{\mathbb{R}^{p_n}} \exp\left(-K\|\boldsymbol{X}\|\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|\right)\Pi(d\boldsymbol{\beta}) \\
&\geq \int_{\mathcal{D}_n} \exp\left(-c_5 n^{1-2\kappa}/2\right)\Pi(d\boldsymbol{\beta}) \\
&\geq \exp\left(-c_5 n^{1-2\kappa}/2\right)\Pi(\mathcal{D}_n).
\end{aligned}$$

Using the fact that $\exp(-c\beta^2/\sqrt{1+c^2\beta^2}) \geq \exp(-|\beta|)$ for any $c > 0$ and the proof of Theorem 2 of Armagan et al. (2013), it is straightforward to show that there exists a positive constant c_6 such that $0 < c_6 < \frac{c_5}{2}$ and

$$\Pi\left(\left\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| < \frac{c_5 n^{1-2\kappa}}{2K\Lambda_{\max}}\right\}\right) \geq \exp\left\{-c_6 n^{1-2\kappa}\right\}. \tag{54}$$

This completes the proof.

REFERENCES

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999), "Broad patterns of gene expression revealed by clustering analysis of tumor

and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences USA*, 96, 6745–6750.

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

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005), "Clustering with Bregman Divergences," *Journal of Machine Learning Research*, 6, 1705–1749.

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, London: Wiley.

Bregman, L. M. (1967), "The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming," *USSR Computational Mathematics and Mathematical Physics*, 7, 200–217.

Fan, J., and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society, Series B*, 70, 849–911.

Fan, J., and Song, R. (2010), "Sure independence screening in generalized linear models with NP–dimensionality," *The Annals of Statistics*, 38, 3217–3841.

Ghosal, S. (1997), "Normal approximation to the posterior distribution for generalized linear models with many covariates," *Mathematical Methods of Statistics*, 6, 332–348.

Itakura, F., and Saito, S. (1970), "A statistical method for estimation of speech spectral denstity and formant frequencies," *Electronics and Communications in Japan*, 53, 36–43.

Kulis, B., Sustik, M. A., and Dhillon, I. S. (2009), "Low–Rank Kernel Learning with Bregman Matrix Divergences," *Journal of Machine Learning Research*, 10, 341–376.

Kyung, M., Gilly, J., Ghosh, M., and Casella, G. (2010), "Penalized Regression, Standard Errors, and Bayesian Lassos," *Bayesian Analysis*, 5, 369–412.

Mallick, B. K., Ghosh, D., and Ghosh, M. (2005), "Bayesian classification of tumours by using gene expression data," *Journal of the Royal Statistical Society, Series B*, 67, 219–234.

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

Ray, S., and Mallick, B. (2006), "Functional clustering by Bayesian wavelet methods," *Journal of the Royal Statistical Society, Series B*, 68, 305–332.

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

Vemuri, B. C., Liu, M., Amari, S. I., and Nielsen, F. (2011), "Total Bregman Divergence and Its Applications to DTI Analysis," *IEEE Transactions on Medical Imaging*, 30, 475–483.

Wedderburn, R. W. M. (1974), "Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method," *Biometrika*, 61, 439–447.

Yuan, M., and Lin, Y. (2007), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, 68, 49–67.

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

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