7-6-2015

Applications of Bregman Divergence Measures in Bayesian Modeling

Gyuhyeong Goh
University of Connecticut - Storrs, gyuhyeong.goh@gmail.com

Applications of Bregman Divergence Measures in Bayesian Modeling

Gyuhyeong Goh, Ph.D.

University of Connecticut, 2015

ABSTRACT

This dissertation has mainly focused on the development of statistical theory, methodology, and application from a Bayesian perspective using a general class of divergence measures (or loss functions), called Bregman divergence. Many applications of Bregman divergence have played a key role in recent advances in machine learning. My goal is to turn the spotlight on Bregman divergence and its applications in Bayesian modeling. Since Bregman divergence includes many well-known loss functions such as squared error loss, Kullback-Leibler divergence, Itakura-Saito distance, and Mahalanobis distance, the theoretical and methodological development unify and extend many existing Bayesian methods. The broad applicability of both Bregman divergence and Bayesian approach can handle diverse types of data such as circular data, high-dimensional data, multivariate data and functional data. Furthermore, the developed methods are flexible to be applied to real applications in various scientific fields including biology, physical sciences, and engineering.

Applications of Bregman Divergence Measures in Bayesian Modeling

Gyuhyeong Goh

B.S., Statistics, Kyungpook National University, Daegu, South Korea, 2008

M.S., Statistics, University of Connecticut, CT, USA, 2013

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2015

APPROVAL PAGE

Doctor of Philosophy Dissertation

Applications of Bregman Divergence Measures in Bayesian Modeling

Presented by

Gyuhyeong Goh, B.S. Statistics, M.S. Statistics

Major Advisor _____
Dipak K. Dey

Associate Advisor _____
Ming-Hui Chen

Associate Advisor _____
Xiaojing Wang

University of Connecticut

2015

ACKNOWLEDGMENTS

First of all, I thank my God through Jesus Christ.

I would like to express my sincere appreciation to my Major Advisor, Professor Dipak K. Dey. Without his guidance, understanding, patience, and mentoring, I could not finish this long journey. He is always there for me and share not only his knowledge, but also his wisdom.

I would also like to thank my Associate Advisor, Professor Ming-Hui Chen. His tremendous works on MCMC computation inspire me with a lot of ideas in this dissertation. A short chat with him always helps me to overcome the struggles that I have faced for a long time.

Sincere thanks to my Associate Advisor, Professor Xiaojing Wang. Her kind support is crucial to complete this dissertation.

Special thanks to Professor Kun Chen, who gives me a lot of help and knowledge about sparse and low-rank regression models. I keenly appreciate his hearty support.

Lastly, thanks to all faculty members and graduate students. This research was partially supported by Elizabeth Macfarlane Fellowship from Department of Statistics, University of Connecticut.

To my Creator, my Savior, my Father and my Friend, Jesus Christ

Contents

Chapter 1

Introduction

The Bregman divergence is a general class of loss functions that includes squared error loss, Kullback-Leibler (KL) divergence, Itakura-Saito distance (Itakura and Saito, 1970), and Mahalanobis distance. The Bregman divergence was originally introduced by Bregman (1967). Later, Banerjee et al. (2005) discovered that a unified optimization algorithm is obtained for any loss function generated by the Bregman divergence. Since then, many applications of the Bregman divergence have played a key role in recent advances in machine learning, see Kulis et al. (2009) and Vemuri et al. (2011) for example.

Unlike in machine learning, the Bregman divergence has not been spotlighted in statistical theory, while some studies had showed its availability in statistical decision theory (Grünwald and Dawid, 2004; Gneiting et al., 2007; Gneiting and Raftery, 2007). In this dissertation, our goal is to turn the spotlight on the Bregman divergence and its applications in Bayesian modeling such as Bayesian model diagnostics, Bayesian predictive model selection, and simultaneous Bayesian estimation and

variable selection for various high-dimensional data including multivariate data and functional data.

We now introduce the major ingredient, *Bregman divergence*.

Definition 1.1. (Bregman Divergence)
Let $\psi : \Omega \to \mathbb{R}$ be a strictly convex function on a convex set $\Omega \subseteq \mathbb{R}^m$, assumed to be nonempty and differentiable. Then for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ the Bregman divergence with respect to ψ is defined as

$$BD_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\psi(\mathbf{y}) \rangle \,, \tag{1.1}$$

where $\nabla\psi$ represents the gradient vector of ψ.

The defined Bregman divergence in (1.1) can be interpreted as the difference between the value of the convex function at $\mathbf{x}$ and its first order Taylor expansion at $\mathbf{y}$, or equivalently the remainder term of the first order Taylor expansion of ψ at $\mathbf{y}$. The geometric significance of the Bregman divergence is illustrated in Figure 1.1. According to Figure 1.1, it is clear that the Bregman divergence is the ordinate distance between the value of the convex function at $\mathbf{x}$ and its tangent at $\mathbf{y}$.

Indeed, the Bregman divergence reduces to a well–known loss function according to the choice of the convex function ψ. For example, if $\psi(\mathbf{x}) = \mathbf{x^T W x}$, where $\mathbf{W}$ is a positive definite matrix, then the Bregman divergence is given as

$$\begin{aligned} BD_\psi(\mathbf{x}, \mathbf{y}) &= \mathbf{x^T W x} - \mathbf{y^T W y} - \langle \mathbf{x} - \mathbf{y}, 2\mathbf{W y} \rangle \\ &= (\mathbf{x} - \mathbf{y})^\mathbf{T} \mathbf{W} (\mathbf{x} - \mathbf{y}). \end{aligned}$$

Note that if $\mathbf{W}$ is assumed to be the inverse of the covariance matrix, then it is called

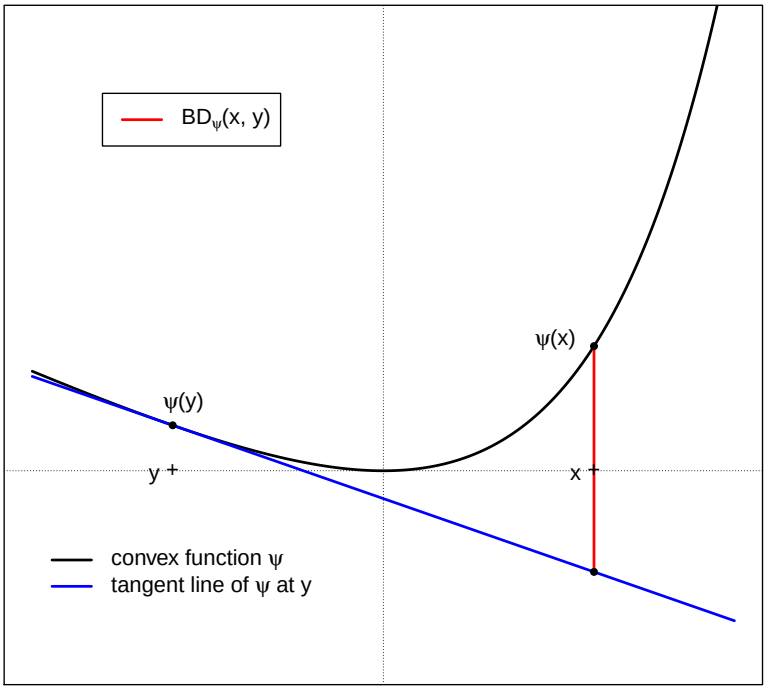

FIGURE 1.1: Geometrical illustration of Bregman divergence, $BD_\psi(x, y)$, with $\psi(x) = e^{cx} - cx - 1$, $c = 0.5$.

"Mahalanobis distance" between $\mathbf{x}$ and $\mathbf{y}$. If we assume that $\mathbf{W}$ is an identity matrix, i.e., $\mathbf{W} = \mathbf{I}$, then the Bregman divergence reduces to the squared Euclidean distance (or squared error loss) between $\mathbf{x}$ and $\mathbf{y}$ such that $BD_\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$, where $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\mathbf{T}\mathbf{x}}$; see Table 1.1 for more examples.

In addition, we define the functional Bregman divergence. Let (X, Ω, ν) be a σ-finite measure space and $f_1(x)$ and $f_2(x)$ be two non-negative measurable functions.

Definition 1.2. (Functional Bregman Divergence)

Let $\psi : (0, \infty) \to \mathbb{R}$ be a strictly convex and differentiable function. Then the

TABLE 1.1: Examples of the Bregman divergence generated by some convex functions, ψ's.

$\psi(\mathbf{x})$	Bregman Divergence	Loss Function
$\|\|\mathbf{x}\|\|^2$	$\|\|\mathbf{x}-\mathbf{y}\|\|^2$	Squared error loss
$\mathbf{x^T W x}$	$(\mathbf{x}-\mathbf{y})^\mathbf{T}\mathbf{W}(\mathbf{x}-\mathbf{y})$	Mahalanobis distance
$\sum_{i=1}^m x_i \log x_i$	$\sum_{i=1}^m \left\{ x_i \log\left(\frac{x_i}{y_i}\right) - x_i + y_i \right\}$	Kullback-Leibler divergence
$\sum_{i=1}^m -\log x_i$	$\sum_{i=1}^m \left\{ \frac{x_i}{y_i} - \log\left(\frac{x_i}{y_i}\right) - 1 \right\}$	Itakura-Saito distance
$\sum_{i=1}^m e^{cx_i}$	$\sum_{i=1}^m \left\{ e^{cy_i}\left(e^{c(x_i-y_i)} - c(x_i-y_i) - 1\right)\right\}$	Weighted Linex loss

functional Bregman divergence D_ψ is defined as

$$D_\psi(f_1, f_2) = \int \{\psi(f_1(x)) - \psi(f_2(x)) - (f_1(x) - f_2(x))\psi'(f_2(x))\}\, d\nu(x), \qquad (1.2)$$

where ψ' represents the derivative of ψ.

Similar to the standard Bregman divergence in (1.1), the functional Bregman divergence also has many nice properties and these aspects will be discussed in Chapter 2.

Lastly, we outline the contents of this dissertation as follows:

- In Chapter 2, we utilize the functional Bregman divergence to measure dissimilarity between posterior distributions in order to develop Bayesian model diagnostics including outlier detection and prior sensitivity analysis. The methodology is exemplified through a logistic regression and a circular data model.

- In Chapter 3, with the intention of generalizing and unifying various existing methods, we introduce a new model selection criterion generated by the Bregman divergence in view of Bayesian decision making. For calculation of the proposed criterion, we also develop a Monte Carlo estimator which significantly

eases the computational burden associated with our approach.

- In Chapter 4, using Bregman divergence, we introduce a novel divergence-based-approach to model sparse high-dimensional regression problems based on the fact that a penalty function in the penalized likelihood method can be viewed as a negative logarithm of prior density function. We further introduce a new prior which induces a new version of the (approximate) adaptive lasso in a Bayesian framework. Theoretically, the posterior consistency is established under a high-dimensional asymptotic regime.

- In Chapter 5, motivated by Chapter 4, we develop a Bayesian simultaneous dimension reduction and variable selection method in the framework of high-dimensional multivariate regression using the Frobenius norm and Kullback-Leibler divergence, which are special cases of Bregman divergence. The newly developed method enables simultaneous rank reduction, predictor selection, as well as response selection and therefore, using the method, we model a regulatory mechanism between Transcription factors, also called sequence-specific DNA binding proteins, and their target genes.

- In Chapter 6, using Bregman divergence, we propose a novel Bayesian clustering method for sparse functional data in reproducing kernel Hilbert space. The Bayesian computations (MCMC and MAP) are also discussed.

Chapter 2

Bayesian Model Diagnostics using Functional Bregman Divergence

2.1 Introduction

Since a Bayesian approach provides a feasible solution to fit complex models, it has been widely used in various fields of study. In general, it is important to perform appropriate model diagnostics after fitting statistical models. In the Bayesian framework, of course, model diagnostics is a crucial data analysis task. Bayesian model diagnostics can be classified into two categories: 1) detection of outliers and influential observations and 2) Bayesian robustness (also called Bayesian sensitivity analysis). In a Bayesian viewpoint, the posterior distribution contains all the information about the parameter, so that examining the posterior distribution of the parameter can provide a method of Bayesian model diagnostics, which is the primary goal in this chapter. One way to detect outliers and influential observations is to measure the

changes in the posterior distribution when the likelihood function is perturbed. For instance, Guttman and Peña (1988) and Peng and Dey (1995) proposed to measure the changes in the posterior distribution to determine the effect of a set of observations when they are deleted from the model. To perform Bayesian robustness or sensitivity analysis (Berger et al., 1988), an investigation of the change caused by a perturbation of a prior can be considered. Gelfand and Dey (1991) and Dey and Birmiwal (1994), for example, considered posterior robustness for different classes of contaminated or mixture of priors using divergence measures.

In this chapter, $\pi_\delta(\boldsymbol{\theta}|\mathbf{y})$ denotes a perturbed posterior distribution and $\pi(\boldsymbol{\theta}|\mathbf{y})$ denotes an unperturbed or full posterior distribution, where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathbf{T}}$ is a vector of independent observations and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a $p-$dimentional parameter vector such that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathbf{T}}$. Let $f_\delta(\mathbf{y}|\boldsymbol{\theta})$ and $\pi_\delta(\boldsymbol{\theta})$ be respectively a likelihood function and a prior distribution under some types of perturbations, then the perturbed posterior can be expressed as

$$\pi_\delta(\boldsymbol{\theta}|\mathbf{y}) \propto f_\delta(\mathbf{y}|\boldsymbol{\theta})\pi_\delta(\boldsymbol{\theta}). \tag{2.1}$$

Similarly, the unperturbed posterior can be expressed as

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{2.2}$$

where $f(\mathbf{y}|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ are respectively the likelihood and the prior without any perturbation. Hence, measuring a divergence between $\pi(\boldsymbol{\theta}|\mathbf{y})$ and $\pi_\delta(\boldsymbol{\theta}|\mathbf{y})$ can be directly applied to the Bayesian model diagnostics with perturbations on the likelihood or on

the prior or both.

If one decides to use the idea of the examination of a discrepancy in the posterior distributions for Bayesian model diagnostics, one will face a problem with a choice of an effective tool to measure the distance (divergence) between unperturbed and perturbed posteriors. In earlier literature, Johnson and Geisser (1983) used Kullback-Leibler divergence to develop model diagnostics for the detection of influential observations. Gelfand and Dey (1991) also proposed the use of Kullback-Leibler divergence in the context of measuring Bayesian robustness. Meanwhile, Dey and Birmiwal (1994) and Peng and Dey (1995) developed Bayesian model diagnostics for a general class of models using general divergence measure, f-divergence in Csiszár (1967).

Our aim, here, is to develop a generalized Bayesian model diagnostic tool that unifies the previous studies (Dey and Birmiwal, 1994; Peng and Dey, 1995) by introducing functional Bregman divergence (Csiszár, 1995; Grünwald and Dawid, 2004). Hence our method can be applied not only to detection of outlier and influential observation but also to Bayesian sensitivity analysis. Furthermore, the use of the functional Bregman divergence greatly benefits the generalization. First, the functional Bregman divergence provides various loss functions such as a half squared ($L^2/2$) Euclidean distance, Kullback-Liebler (KL) divergence, and Itakura-Saito (IS) divergence (Itakura and Saito, 1968). Hence we can measure the discrepancy between the posterior distributions using various measuring devices (loss functions). In addition, the functional Bregman divergence helps us to apply the proposed method to continuous as well as discrete cases. By the definition given in (1.2), the functional

Bregman divergence reduces to Bregman divergence (Bregman, 1967) using an appropriate measure such as a counting measure. The classical Bregman divergence has been widely used in machine learning problems (Banerjee et al., 2005; Taskar et al., 2006), but it assumes that the probability distribution is discrete. Therefore, our method can be used when the posterior distribution is continuous, discrete, or even mixtures. In general, posterior distributions are continuous, thus we assume that they are continuous. More details about functional Bregman divergence will be discussed in Section 2.2.

The outline of the rest of the chapter is the following. In Section 2.2, some useful properties of the functional Bregman divergence are briefly discussed. In Section 2.3, we propose two different approaches (rate comparison and direct comparison) for Bayesian model diagnostics using a functional Bregman Divergence. In this section, we further show that the rate comparison approach is equivalent to Peng and Dey (1995). In the direct comparison approach, we develop two different approximation methods, so that the results can be applied to more general and complex statistical models. Sensitivity analysis is discussed in Section 2.4 through Monte Carlo simulation study. In Section 2.5, two examples are studied, in which we consider a Bayesian generalized linear model for binary response data with logit, probit, and complementary log-log links and Bayesian circular data analysis. Section 2.6 concludes the chapter with some remarks.

2.2 Functional Bregman divergence

In this section, we discuss some useful properties of the functional Bregman divergence. The functional Bregman divergence includes varieties of distortion functions according to the choice of the convex function ψ. However, we mainly consider the following class of convex functions $\psi_\alpha(x)$ for $\alpha \in \mathbb{R}$ (Eguchi and Kano, 2001).

$$\psi_\alpha(x) = \begin{cases} x\log x - x + 1, & \alpha = 1 \\ -\log x + x - 1, & \alpha = 0 \\ \frac{x^\alpha - \alpha x + \alpha - 1}{\alpha(\alpha-1)}, & \text{otherwise} \end{cases} . \tag{2.3}$$

Note that the convex function $\psi_\alpha(x)$ is continuous with respect to α (Hennequin et al., 2011). Using (2.3), the functional Bregman divergence in (1.2) is given as

$$D_{\psi_\alpha}(f_1, f_2) = \begin{cases} \int \left\{ \frac{f_1}{f_2} - \log\left(\frac{f_1}{f_2}\right) - 1 \right\} d\nu, & \alpha = 0 \\ \int \left\{ f_1 \log\left(\frac{f_1}{f_2}\right) - (f_1 - f_2) \right\} d\nu, & \alpha = 1 \\ \int \frac{f_1^\alpha - \alpha f_1 f_2^{\alpha-1} + (\alpha-1) f_2^\alpha}{\alpha(\alpha-1)} d\nu, & \text{otherwise} \end{cases} . \tag{2.4}$$

According to an appropriate choice of α, the functional Bregman divergence in (2.4) becomes some well-known divergences. For instance, when $\alpha = 0$, the divergence is the IS divergence. If $\alpha = 1$, then it reduces to the KL divergence. For $\alpha = 2$, it becomes the $L^2/2$ Euclidean distance. The standard Bregman divergence for vectors has some well-known useful properties (Banerjee et al., 2005). The functional Bregman divergence also has similar properties. The proofs of the properties are shown in Frigyik et al. (2008).

1. (Nonnegativity) $D_\psi(f_1, f_2) \geq 0$ for any non-negative measurable functions and the equality holds if and only if $f_1 = f_2$ almost surely.

2. (Convexity) $D_\psi(f_1, f_2)$ is always convex with respect to f_1, but not necessarily convex with respect to f_2.

3. (Linearity) $D_{c_1\psi_1+c_2\psi_2}(f_1, f_2) = c_1 D_{\psi_1}(f_1, f_2)+c_2 D_{\psi_2}(f_1, f_2)$ for $c_1, c_2 \geq 0$, where ψ_1 and ψ_2 are two functions over $(0, \infty) \to \mathbb{R}$, strictly convex and differentiable.

4. (Equivalence Classes) If $\psi(f_1) = \psi_1(f_1)+bf_1+c$ where b,c $\in \mathbb{R}$, then $D_\psi(f_1, f_2) = D_{\psi_1}(f_1, f_2)$. Therefore, the set of strictly convex functions can be partitioned into equivalence classes such that $[\psi_1] = \{\psi | D_\psi(f_1, f_2) = D_{\psi_1}(f_1, f_2)\}$.

5. (Linear Separation) The locus of the non-negative measurable function f that has the same distance from two fixed functions f_1 and f_2 is a hyperplane.

6. (Dual Divergence) Let ψ be a Legendre function and ψ^* be its conjugate, then $D_\psi(f_1, f_2) = D_{\psi^*}(f_1^*, f_2^*)$, where f_1 and f_2 are respectively related to f_1^* and f_2^* by the Legendre transformation.

7. (Generalized Pythagorean Inequality) For any non-negative measurable function f_1, f_2, and f_3, the functional Bregman divergence satisfies the following equation : $D_\psi(f_1, f_3) = D_\psi(f_1, f_2) + D_\psi(f_2, f_3) + \int(\psi'(f_2) - \psi'(f_3))(f_1 - f_2)d\nu$.

2.3 Bayesian model diagnostics

Let $\{f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be a class of statistical models and $\mathbf{X}$ be a matrix of covariates. We define a general perturbation as follows:

$$\delta(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X}) = \frac{f_\delta(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})\pi_\delta(\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})\pi(\boldsymbol{\theta})}. \tag{2.5}$$

In the outlier detection context, the following two perturbations can be considered due to independence of observations:

$$\begin{aligned} \delta_1(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X}) &= \frac{f(\mathbf{y}_{(i)}|\boldsymbol{\theta}, \mathbf{X}_{(i)})\pi(\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})\pi(\boldsymbol{\theta})} \\ &= \frac{1}{f(y_i|\boldsymbol{\theta}, \mathbf{x}_i)} \end{aligned} \tag{2.6}$$

and

$$\begin{aligned} \delta_2(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}_{[i(j)]})\pi(\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})\pi(\boldsymbol{\theta})} \\ &= \frac{f(y_i|\boldsymbol{\theta}, \mathbf{x}_{i(j)})}{f(y_i|\boldsymbol{\theta}, \mathbf{x}_i)}, \end{aligned} \tag{2.7}$$

where y_i and $\mathbf{x}_i$, respectively, are the i^{th} observation and the i^{th} covariate vector, $\mathbf{y}_{(i)}$ denotes a random data vector with the i^{th} observation deleted, and $\mathbf{X}_{[i(j)]}$ indicates the matrix obtained by deleting the jth component of i^{th} covariate vector from $\mathbf{X}$ while $\mathbf{x}_{i(j)}$ is the i^{th} covariate vector with the j component deleted. Here, the above perturbations can be used to measure the effect of i^{th} observation and the effect of covariates on the model, respectively. For the Bayesian robustness or sensitivity

analysis, the following perturbation can be considered:

$$\begin{aligned} \delta_3(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X}) &= \frac{f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})\pi_\epsilon(\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})\pi(\boldsymbol{\theta})} \\ &= \frac{\pi_\epsilon(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}, \end{aligned} \tag{2.8}$$

where π_ϵ is a class of contaminated priors or their mixture. For example, Dey and Birmiwal (1994) considered the ϵ-contaminated class of priors such as

$$\mathbf{\Pi} = \{\pi_\epsilon(\boldsymbol{\theta}) \ : \ \pi_\epsilon(\boldsymbol{\theta}) = (1-\epsilon)\pi(\boldsymbol{\theta}) + \epsilon q(\boldsymbol{\theta}), \ q \in \mathbf{Q}, \ 0 \leq \epsilon \leq 1\},$$

where π is the elicited prior and $\mathbf{Q}$ is a class of distribution. Note that all the results obtained in this chapter can be applied to any types of perturbations as far as the perturbed posterior distribution exists.

2.3.1 Rate comparison approach

For Bayesian model diagnostics measures, f-divergence has played a major role in developing a general model diagnostics methods. In general, the Bregman divergence is different than the f-divergence due to the fact that the f-divergence is related to the alpha geometrical structure but the Bregman divergence is based on a dually flat geometrical structure, see Amari (2009) for more details. Nevertheless, the following theorem shows that we can develop a method for the model diagnostics based on the functional Bregman divergence that is equivalent to the f-divergence.

Theorem 2.1. *Let f_1 and f_2 be two probability densities with corresponding distribution functions F_1 and F_2 and $\Phi_\psi(f_1, f_2)$ be the f-divergence between f_1 and f_2 with*

a convex function $\psi(\cdot)$. If the Borel measure ν is F_2, then

$$\Phi_\psi\left(f_1, f_2\right) = D_\psi(f_1/f_2, 1) + \psi(1). \tag{2.9}$$

According to the above theorem, we know that the f-divergence between two posteriors $\pi(\boldsymbol{\theta}|\mathbf{y})$ and $\pi_\delta(\boldsymbol{\theta}|\mathbf{y})$ is equivalent to the functional Bregman divergence between the rate of two posteriors $\pi_\delta(\boldsymbol{\theta}|\mathbf{y})/\pi(\boldsymbol{\theta}|\mathbf{y})$ and the constant function $h(\boldsymbol{\theta}) = 1$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. As a result, we propose the following quantity for model diagnostics under Bayesian framework by using the functional Bregman divergence:

$$\begin{aligned} D_\psi^R &= D_\psi\left(\frac{\pi_\delta(\boldsymbol{\theta}|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})}, 1\right) \quad (\text{letting } d\nu = \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}) \\ &= \int \left\{\psi\left(\frac{\pi_\delta(\boldsymbol{\theta}|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})}\right) - \psi(1) - \left(\frac{\pi_\delta(\boldsymbol{\theta}|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})} - 1\right)\psi'(1)\right\}\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\ &= \int \psi\left(\frac{\pi_\delta(\boldsymbol{\theta}|\mathbf{y})}{\pi(\boldsymbol{\theta}|\mathbf{y})}\right)\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} - \psi(1). \end{aligned}$$

From (2.5), we obtain

$$\pi_\delta(\boldsymbol{\theta}|\mathbf{y}) = \frac{\delta(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X})\pi(\boldsymbol{\theta}|\mathbf{y})}{\int \delta(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}}. \tag{2.10}$$

By using (2.10), D_ψ^R can be written as

$$D_\psi^R = \int \psi\left(\frac{\delta(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X})}{\int \delta(\boldsymbol{\theta}, \mathbf{y}, \mathbf{X})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}}\right)\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} - \psi(1).$$

A Monte Carlo estimate of D_ψ^R is then given by

$$\hat{D}_\psi^R = \frac{1}{N}\sum_{s=1}^{N}\left[\psi\left(\frac{\delta(\boldsymbol{\theta}^s,\mathbf{y},\mathbf{X})}{\frac{1}{N}\sum_{s=1}^{N}\delta(\boldsymbol{\theta}^s,\mathbf{y},\mathbf{X})}\right)\right] - \psi(1), \tag{2.11}$$

where $\{\boldsymbol{\theta}^s\}_{s=1}^N$ are samples from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$. Note that our rate comparison approach based on the functional Bregman divergence is equivalent to Peng and Dey (1995).

2.3.2 Direct comparison approach

In the previous section, we showed that the rate of the posteriors could be used for the Bayesian model diagnostics based on a functional Bregman divergence. In this section, we propose a new method to measure the discrepancy between two posteriors via a direct comparison of two posteriors $\pi(\boldsymbol{\theta}|\mathbf{y})$ and $\pi_\delta(\boldsymbol{\theta}|\mathbf{y})$ using the functional Bregman divergence. Since the functional Bregman divergence measures the discrepancies between two probability densities, we can directly compare two posteriors using the functional Bregman divergence as given below. Observe that the following quantity can be used for an outlier detection or a sensitivity analysis corresponding to a choice of a certain perturbation:

$$D_\psi = \int \left\{\psi(\pi(\boldsymbol{\theta}|\mathbf{y})) - \psi(\pi_\delta(\boldsymbol{\theta}|\mathbf{y})) - (\pi(\boldsymbol{\theta}|\mathbf{y}) - \pi_\delta(\boldsymbol{\theta}|\mathbf{y}))\,\psi'\left(\pi_\delta(\boldsymbol{\theta}|\mathbf{y})\right)\right\} d\boldsymbol{\theta}. \tag{2.12}$$

The problem here is that we cannot obtain the above quantity directly due to the fact that in many cases the integral cannot be expressed in a closed form. In order to overcome this problem, we propose two different ways to approximate the functional Bregman divergence; Gaussian approximation and Importance-Weighted Marginal

Density Estimation. If our study focuses on Generalized Linear models (GLMs), the Gaussian approximation could be reasonable; see Ghosal et al. (1995); Ghosal (1997). In addition, the integral has a closed form with respect to the convex function in (2.3), so we can directly solve the integral problem. However, it could not be always expressed as a closed form for any choice of convex functions, in this case importance sampling method can be used for any choice of the convex function. Finally, we develop a flexible approximation technique using Importance-weighted marginal density estimation for complex models.

Divergence approximation using Gaussian approximation

From the Bernstein-von Mises theorem, we can derive the following approximation for the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \approx g\left(\boldsymbol{\theta}\ ; \hat{\boldsymbol{\theta}}_{\mathbf{y}}, V(\hat{\boldsymbol{\theta}}_{\mathbf{y}})\right) \quad (\equiv \tilde{\pi}_G(\boldsymbol{\theta}|\mathbf{y}))\,, \tag{2.13}$$

where $g(\cdot\ ; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicates the density of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\hat{\boldsymbol{\theta}}_{\mathbf{y}}$ is the mode of the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$, and $V(\hat{\boldsymbol{\theta}}_{\mathbf{y}})$ is the inverse of the negative Hessian of $\log(\pi(\boldsymbol{\theta}|\mathbf{y}))$ evaluated at the mode $\hat{\boldsymbol{\theta}}_{\mathbf{y}}$. The Gaussian approximation could be a short-cut to estimate the posterior density because it provides relatively quick result. To determine the mode of the posterior, many algorithms are available such as Newton's method, Fisher's scoring method, and Nelder-Mead algorithm, etc. Here, we suggest the use of Nelder-Mead algorithm since it is less sensitive to the choice of initial value than that of Newton's method.

Using (2.13), the functional Bregman divergence between the two posteriors in (2.12)

can be approximated by

$$\begin{aligned}\hat{D}_{\psi}^{G} &= \int \left\{\psi(\tilde{\pi}^{G}(\boldsymbol{\theta}|\mathbf{y})) - \psi(\tilde{\pi}_{\delta}^{G}(\boldsymbol{\theta}|\mathbf{y}))\right. \\ &\quad \left. - \left(\tilde{\pi}^{G}(\boldsymbol{\theta}|\mathbf{y}) - \tilde{\pi}_{\delta}^{G}(\boldsymbol{\theta}|\mathbf{y})\right)\psi'\left(\tilde{\pi}_{\delta}^{G}(\boldsymbol{\theta}|\mathbf{y})\right)\right\} d\boldsymbol{\theta}. \qquad (2.14)\end{aligned}$$

Note that under the Gaussian approximation technique, the functional Bregman divergence in (2.12) with the convex function in (2.3) becomes infinity when $\alpha < 1$, so we only consider the case when $\alpha \geq 1$. The following lemma shows how to simplify the integral in (2.14).

Lemma 2.2. *Let f_1 and f_2 be pdfs of $N_p(\boldsymbol{\mu}_1,\ \boldsymbol{\Sigma}_1)$ and $N_p\ (\boldsymbol{\mu}_2,\ \boldsymbol{\Sigma}_2)$, respectively. Suppose that $\alpha \geq 1$ in (2.3).*
Then $D_{\psi_\alpha}(f_1, f_2)$ as defined in (2.4) can be obtained as follows:

If $\alpha = 1$,

$$D_{\psi_\alpha}(f_1, f_2) = \frac{1}{2}\left[\mathbf{tr}\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\right) - \log\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) - p + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathbf{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right].$$

If $\alpha > 1$,

$$\begin{aligned}D_{\psi_\alpha}(f_1, f_2) &= \frac{|\boldsymbol{\Sigma}_1^{-1}|^{\frac{\alpha^*}{2}} + \alpha^*|\boldsymbol{\Sigma}_2^{*-1}|^{\frac{\alpha^*}{2}}}{\alpha^{\frac{p}{2}+1}\alpha^*(2\pi)^{\frac{p\alpha^*}{2}}} - \frac{|\boldsymbol{\Sigma}_1^{-1}|^{\frac{1}{2}}|\boldsymbol{\Sigma}_2^{*-1}|^{\frac{\alpha^*}{2}}}{(\alpha^*)^{\frac{p\alpha^*}{2}+1}(2\pi)^{\frac{p\alpha^*}{2}}|\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{*-1}|^{\frac{1}{2}}} \\ &\times \exp\left[-\frac{1}{2}\left\{\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^{\mathbf{T}}\boldsymbol{\Sigma}_2^{*-1}\boldsymbol{\mu}_2\right\}\right] \\ &\times \exp\left[\frac{1}{2}\left\{(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{*-1}\boldsymbol{\mu}_2)^{\mathbf{T}}(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{*-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_2^{*-1}\boldsymbol{\mu}_2)\right\}\right],\end{aligned}$$

where $\alpha^* = \alpha - 1$ *and* $\boldsymbol{\Sigma}_2^* = \boldsymbol{\Sigma}_2/(\alpha - 1)$.

Since the proof is straightforward, we omit it. In equation (2.14), suppose that $\tilde{\pi}^G(\boldsymbol{\theta}|\mathbf{y})$ and $\tilde{\pi}_\delta^G(\boldsymbol{\theta}|\mathbf{y})$ are pdfs of $\mathrm{N}_p(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$ and $\mathrm{N}_p(\boldsymbol{\mu}_\delta,\ \boldsymbol{\Sigma}_\delta)$, respectively. From above lemma, we obtain the results in the following examples.

Example 2.3 (KL Divergence)**.** Choose $\alpha = 1$, then the estimate of divergence is given by

$$\hat{D}_{KL}^G = \frac{1}{2}\left[\mathbf{tr}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\delta^{-1}\right) - \log\left(\frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}_\delta|}\right) - p + (\boldsymbol{\mu} - \boldsymbol{\mu}_\delta)^{\mathbf{T}}\boldsymbol{\Sigma}_\delta^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_\delta)\right].$$

Example 2.4 (Half Squared Euclidean Distance)**.** Choose $\alpha = 2$, then the estimate of the divergence is given by

$$\begin{aligned}\hat{D}_{L^2/2}^G &= \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2} + |\boldsymbol{\Sigma}_\delta^{-1}|^{1/2}}{2(4\pi)^{p/2}} - \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_\delta|^{1/2}} \\ &\quad \times \exp\left[-\frac{1}{2}\left\{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}_\delta^{\mathbf{T}}\boldsymbol{\Sigma}_\delta^{-1}\boldsymbol{\mu}_\delta\right\}\right] \\ &\quad \times \exp\left[\frac{1}{2}\left\{(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma}_\delta^{-1}\boldsymbol{\mu}_\delta)^{\mathbf{T}}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_\delta^{-1})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma}_\delta^{-1}\boldsymbol{\mu}_\delta)\right\}\right].\end{aligned}$$

In fact, our choice for the convex function ψ is very flexible as far as it is strictly convex and differentiable. Nevertheless, unfortunately, for some convex functions, the integral in (2.12) cannot be expressed in a closed form. In this situation, importance sampling could be one of the solution for solving such integration problem in (2.12). Suppose $q(\cdot)$ is a sampling density. Using this $q(\cdot)$, the equation (2.12) can

be expressed as

$$
\begin{aligned}
(2.12) \quad &= \int \left\{\psi(\pi(\boldsymbol{\theta}|\mathbf{y})) - \psi(\pi_\delta(\boldsymbol{\theta}|\mathbf{y})) - (\pi(\boldsymbol{\theta}|\mathbf{y}) - \pi_\delta(\boldsymbol{\theta}|\mathbf{y}))\,\psi'\left(\pi_\delta(\boldsymbol{\theta}|\mathbf{y})\right)\right\} \frac{q(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= E_{q(\boldsymbol{\theta})}\left\{ \frac{\psi(\pi(\boldsymbol{\theta}|\mathbf{y})) - \psi(\pi_\delta(\boldsymbol{\theta}|\mathbf{y})) - (\pi(\boldsymbol{\theta}|\mathbf{y}) - \pi_\delta(\boldsymbol{\theta}|\mathbf{y}))\,\psi'\left(\pi_\delta(\boldsymbol{\theta}|\mathbf{y})\right)}{q(\boldsymbol{\theta})} \right\},
\end{aligned}
$$

which can be estimated as

$$
\hat{D}^I_\psi = \frac{1}{N}\sum_{s=1}^{N} \left\{ \frac{\psi(\tilde{\pi}^G(\boldsymbol{\theta}^s|\mathbf{y})) - \psi(\tilde{\pi}^G_\delta(\boldsymbol{\theta}^s|\mathbf{y})) - \left(\tilde{\pi}^G(\boldsymbol{\theta}^s|\mathbf{y}) - \tilde{\pi}^G_\delta(\boldsymbol{\theta}^s|\mathbf{y})\right)\psi'\left(\tilde{\pi}^G_\delta(\boldsymbol{\theta}^s|\mathbf{y})\right)}{q(\boldsymbol{\theta}^s)} \right\},
$$

where $\{\boldsymbol{\theta}^s\}_{s=1}^N$ are samples from $q(\boldsymbol{\theta})$ and the estimated posterior $\tilde{\pi}^G(\cdot|\cdot)$ is determined by Gaussian approximation. We define the sampling density $q(\cdot)$ as follows using Gaussian approximations of posteriors in (2.13) so that it reduces the simulation error. Suppose that $q(\boldsymbol{\theta})$ is a density function of an approximated posterior via Gaussian approximation, i.e., $q(\boldsymbol{\theta}) = \tilde{\pi}^G(\boldsymbol{\theta}|\mathbf{y})$. Then the following example follows:

Example 2.5. Let $\psi(x) = \psi_\alpha(x)$ and $\{\boldsymbol{\theta}^s\}_{s=1}^N$ be samples from the approximated posterior distribution $\tilde{\pi}^G(\boldsymbol{\theta}|\mathbf{y})$, then a Monte Carlo estimate of the divergence is given as follows:

$$
\begin{aligned}
&\text{If} \quad \alpha = 1, \\
&\qquad \hat{D}^I_{\psi_\alpha} = \frac{1}{N}\sum_{s=1}^{N} \left\{ -\log\left(\frac{\tilde{\pi}^G_\delta(\boldsymbol{\theta}^s|\mathbf{y})}{\tilde{\pi}^G(\boldsymbol{\theta}^s|\mathbf{y})} \right) \right\}. \\
&\text{If} \quad \alpha > 1, \\
&\qquad \hat{D}^I_{\psi_\alpha} = \frac{1}{N}\sum_{s=1}^{N} \left[\frac{1 - \alpha\left\{\frac{\tilde{\pi}^G_\delta(\boldsymbol{\theta}^s|\mathbf{y})}{\tilde{\pi}^G(\boldsymbol{\theta}^s|\mathbf{y})}\right\}^{\alpha-1} + (\alpha-1)\left\{\frac{\tilde{\pi}^G_\delta(\boldsymbol{\theta}^s|\mathbf{y})}{\tilde{\pi}^G(\boldsymbol{\theta}^s|\mathbf{y})}\right\}^{\alpha}}{\alpha(\alpha-1)\left\{\tilde{\pi}^G(\boldsymbol{\theta}^s|\mathbf{y})\right\}^{1-\alpha}} \right].
\end{aligned}
$$

Note that the Gaussian approximation method can be also applied to measure influence on the marginal posterior distribution of single parameter θ_i, where $i = 1, \ldots, p$. From (2.13), we derive the following approximation for the marginal posterior density of θ_i:

$$\pi(\theta_i|\mathbf{y}) \approx g\left(\theta_i \,; (\hat{\boldsymbol{\theta}}_{\mathbf{y}})_i, V(\hat{\boldsymbol{\theta}}_{\mathbf{y}})_{ii}\right), \tag{2.15}$$

where $(\hat{\boldsymbol{\theta}}_{\mathbf{y}})_i$ indicates i^{th} element of $\hat{\boldsymbol{\theta}}_{\mathbf{y}}$ and $V(\hat{\boldsymbol{\theta}}_{\mathbf{y}})_{ii}$ is i^{th} diagonal element of the dispersion matrix $V(\hat{\boldsymbol{\theta}}_{\mathbf{y}})$. Then the problem becomes a special case of where the dimension of the parameter space is one, i.e., $p = 1$.

Divergence approximation using IWMDE

The Gaussian approximation method could provide relatively accurate and fast approximation under the GLM framework. Nonetheless, it cannot be implemented if the posterior has complex structures such as in a Bayesian hierarchical model. In this case, it is hard to avoid Markov chain Monte Carlo (MCMC) approach under such a complex model. However, implementation of MCMC for a posterior comparison method using a functional Bregman divergence is a challenge for the following reasons. First, in many cases an unperturbed posterior density and perturbed posterior densities are unknown, so that they should be estimated. Second, estimating the divergence in (2.12) for all perturbations could be a burden. Lastly, implementing MCMC more than once is not realistic because MCMC requires long time until the chain converges to the posterior distribution. In order to overcome these problems, we develop the estimation of the functional Bregman divergence using Importance-Weighted Marginal

Density Estimation (IWMDE) (Chen, 1994). Surprisingly IWMDE can be used to estimate both the divergence and the posterior densities simultaneously using a set of samples from the full posterior distribution.

The normalizing constant for $\pi(\boldsymbol{\theta}|\mathbf{y})$ is given by

$$m^{-1}(\mathbf{y}) = \int \frac{w(\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}, \tag{2.16}$$

where $m(\mathbf{y})$ is the marginal density and $w(\cdot)$ is a probability density function. Let $\{\boldsymbol{\theta}^s\}_{s=1}^N$ be samples from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$, where the samples could be generated by MCMC. Using (2.16), the IWMDE is given by

$$\tilde{m}^{\mathrm{IW}}(\mathbf{y}) = \left[\frac{1}{N}\sum_{s=1}^{N}\frac{w(\boldsymbol{\theta}^s)}{f(\mathbf{y}|\boldsymbol{\theta}^s)\pi(\boldsymbol{\theta}^s)}\right]^{-1}. \tag{2.17}$$

From (2.10), the estimate of perturbed posterior is given by

$$\tilde{\pi}_{\delta}^{\mathrm{IW}}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\tilde{\pi}^{\mathrm{IW}}(\boldsymbol{\theta}|\mathbf{y})\delta(\boldsymbol{\theta},\mathbf{y},\mathbf{X})}{\frac{1}{N}\sum_{s=1}^{N}\delta(\boldsymbol{\theta}^s,\mathbf{y},\mathbf{X})},$$

where $\tilde{\pi}^{\mathrm{IW}}(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/\tilde{m}^{\mathrm{IW}}(\mathbf{y})$. Therefore, a Monte Carlo estimate of the functional Bregman divergence in (2.12) is given as

$$\begin{aligned}\hat{D}_{\psi}^{\mathrm{IW}} &= \frac{1}{N}\sum_{s=1}^{N}\left\{\frac{\psi(\tilde{\pi}^{\mathrm{IW}}(\boldsymbol{\theta}^s|\mathbf{y})) - \psi(\tilde{\pi}_{\delta}^{\mathrm{IW}}(\boldsymbol{\theta}^s|\mathbf{y}))}{\tilde{\pi}^{\mathrm{IW}}(\boldsymbol{\theta}^s|\mathbf{y})}\right. \\ &\quad \left. - \frac{\left(\tilde{\pi}^{\mathrm{IW}}(\boldsymbol{\theta}^s|\mathbf{y}) - \tilde{\pi}_{\delta}^{\mathrm{IW}}(\boldsymbol{\theta}^s|\mathbf{y})\right)\psi'\left(\tilde{\pi}_{\delta}^{\mathrm{IW}}(\boldsymbol{\theta}^s|\mathbf{y})\right)}{\tilde{\pi}^{\mathrm{IW}}(\boldsymbol{\theta}^s|\mathbf{y})}\right\},\end{aligned}$$

where $\{\boldsymbol{\theta}^s\}_{s=1}^N$ are samples from a posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$. In many cases, this approach is relied on MCMC which is relatively slow. However, this approach is quite

efficient with respect to Bayesian modeling due to the fact that the procedure requires only one time sampling from posterior density and we would already have the samples of posterior for the inference procedure. In IWMDE, a choice of a good $w(\cdot)$ directly effects the simulation error. Hence the best choice of w is to choose the closest density to the target function $f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. We assume w as a multivariate Gaussian distribution for the sake of simplicity. Under the Gaussian assumption, we choose w that minimizes the functional Bregman divergence between $\pi(\boldsymbol{\theta}|\mathbf{y})$ and $w(\boldsymbol{\theta})$ due to the fact $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Then the following theorem provides the optimal choice of w.

Theorem 2.6. *Let $f(\mathbf{x})$ be a density function of a continuous random variable with mean $\boldsymbol{\eta}$ and variance $\mathbf{V}$ and $g(\mathbf{x})$ be a Gaussian density with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, where $\mathbf{x} \in \mathbb{R}^p$. If $\psi(x) = x \log x$, then*

$$\arg\min_{(\boldsymbol{\mu},\boldsymbol{\Sigma})} D_\psi \left(f(\mathbf{x}), g(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) = (\boldsymbol{\eta}, \mathbf{V}). \tag{2.18}$$

According to the above theorem, we propose the estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as follows:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{s=1}^{N} \boldsymbol{\theta}^s, \tag{2.19}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{s=1}^{N} (\boldsymbol{\theta}^s - \hat{\boldsymbol{\mu}})(\boldsymbol{\theta}^s - \hat{\boldsymbol{\mu}})^{\mathbf{T}}, \tag{2.20}$$

where $\{\boldsymbol{\theta}^s\}_{s=1}^N$ is a sample from a posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$. Note that the function $w(\cdot)$ is only related to the estimate of the normalizing constant in (2.16), thus the proposed $w(\cdot)$ based on Theorem 2 can be used to estimate the functional Bregman

divergence in (2.12) regardless of any convex function ψ. Using IWMDE method, the following example can be given:

Example 2.7. Let $\psi(x) = \psi_\alpha(x)$ and $\{\boldsymbol{\theta}^s\}_{s=1}^N$ be samples from a posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$, then a Monte Carlo estimate of the divergence is given as follows:

$$\begin{aligned}
\hat{D}_{\psi_\alpha}^{\mathrm{IW}} &= \frac{1}{N}\sum_{s=1}^{N}\left[\frac{\left\{\delta(\boldsymbol{\theta}^s,\mathbf{y},\mathbf{X})/\bar{\delta}\right\}^{-1}+\log\left\{\delta(\boldsymbol{\theta}^s,\mathbf{y},\mathbf{X})/\bar{\delta}\right\}-1}{\tilde{\pi}^{\mathrm{IW}}(\boldsymbol{\theta}^s|\mathbf{y})}\right] && \text{if } \alpha{=}0,\\
\hat{D}_{\psi_\alpha}^{\mathrm{IW}} &= \frac{1}{N}\sum_{s=1}^{N}\left[-\log\left\{\delta(\boldsymbol{\theta}^s,\mathbf{y},\mathbf{X})/\bar{\delta}\right\}\right] && \text{if } \alpha{=}1,\\
\hat{D}_{\psi_\alpha}^{\mathrm{IW}} &= \frac{1}{N}\sum_{s=1}^{N}\left[\frac{1-\alpha\left\{\delta(\boldsymbol{\theta}^s,\mathbf{y},\mathbf{X})/\bar{\delta}\right\}^{\alpha-1}+(\alpha-1)\left\{\delta(\boldsymbol{\theta}^s,\mathbf{y},\mathbf{X})/\bar{\delta}\right\}^{\alpha}}{\alpha(\alpha-1)\left\{\tilde{\pi}^{\mathrm{IW}}(\boldsymbol{\theta}^s|\mathbf{y})\right\}^{1-\alpha}}\right] && \text{o.w.},
\end{aligned}$$

where $\bar{\delta} = \frac{1}{N}\sum_{s=1}^{N}\delta(\boldsymbol{\theta}^s,\mathbf{y},\mathbf{X})$.

2.4 Sensitivity analysis of functional Bregman divergence

In this section, we implement sensitivity analysis of a functional Bregman divergence on internal and external changes through Monte Carlo simulation. Here, we define that the internal change represents the changes in functional Bregman divergence itself such as in different types of convex functions and the external change means the changes in non-negative measurable functions in (1.2).

2.4.1 Sensitivity to internal change

First we study the sensitivity of the functional Bregman divergence to the choice of a convex function (internal change). Consider a Gaussian distribution which is the

most well-known location-scale family. Let f_1 and f_2 be pdfs of $N(0,1)$ and $N(\mu,\sigma^2)$, respectively. Suppose that $\psi = \psi_\alpha$. Now, we consider two scenarios. In scenario 1, we fix $\sigma^2 = 1$ and change μ away from zero under different choices of α. In scenario 2, we fix $\mu = 0$ and then move σ^2 away from one. Figure 2.1 displays the plots of the measured divergences between f_1 and f_2 in scenario 1 and scenario 2, respectively. From the plots, we notice that when α is smaller the functional Bregman divergence tends to be more sensitive because the derivative of the function decreases as α increases. Note that this study provides us a guideline for the use of an appropriate convex function. For example, in social science field, the data naturally include many outliers, in this case we choose a convex function with large α, so that it is robust with respect to the natural variation. However, in medical research, unexpected outliers are fatal, thus we need to use the most sensitive measure ($\alpha = 1$).

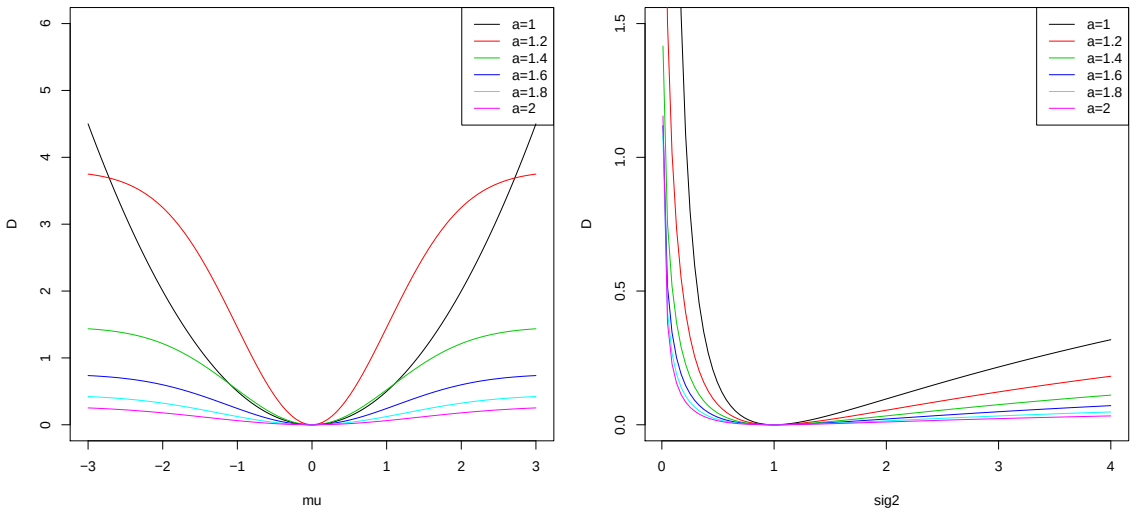

FIGURE 2.1: Plots of functional Bregman divergence in scenario 1 (left) and in scenario 2 (right).

2.4.2 Sensitivity to external change

In this section, we examine sensitivity to the changes in posteriors (external change) via Monte Carlo simulation. The sensitivity analysis of the posteriors can be interpreted as the Bayesian robustness or sensitivity analysis if the changes in the posteriors are caused by priors. Consider a Bayesian linear regression model as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{2.21}$$

where $\epsilon \sim N(\mathbf{0}, \sigma^2\mathrm{I})$. Suppose that σ^2 is known and the prior of $\boldsymbol{\beta}$ is $N(\mathbf{b_0}, \sigma_\mathbf{0}^\mathbf{2}\mathrm{I})$, then the posterior distribution is given as

$$\boldsymbol{\beta}|\mathbf{y} \sim N\left(\left(\frac{\mathbf{X^T X}}{\sigma^2} + \frac{\mathrm{I}}{\sigma_0^2}\right)^{-1}\left(\mathbf{X^T y} - \mathbf{b_0}\right), \left(\frac{\mathbf{X^T X}}{\sigma^2} + \frac{\mathrm{I}}{\sigma_0^2}\right)^{-1}\right). \tag{2.22}$$

Under the model, we first focus on sensitivity of sample size n in outlier detection context using perturbation based on case deletion in form (2.6). In simulation, for given n we assign $\boldsymbol{\beta} = 1$, $\mathbf{x}_i \overset{iid}{\sim} U(0, 100)$, and $\sigma^2 = 1$ for the model and $\mathbf{b_0} = 0$ and $\sigma_0^2 = 1000$ for the prior , where $\mathbf{x}_i$ is i^{th} row vector of $\mathbf{X}$. In each iteration, we intentionally insert one outlier $y_1 = \mathbf{x}_1\boldsymbol{\beta} + q$, where q is $1 - 10^{-8}$ quantile of a standard normal distribution. Using 1000 iterations, we calculate the average of divergences between the full posterior and the outlier deleted posterior for given n. As the sample size n increases from 10 to 1000, we repeat the simulation. Figure 2.2 (left) displays the trace of the calculated divergence against the increased sample sizes, where lower dotted line and upper dashed line respectively indicate cut-off points for mild outliers and severe outliers (discussed in Section 2.5). The plot shows that a larger sample size dramatically reduces the effect of the outlier on the changes in posterior density.

Note that if sample size n is larger than 800, then the outlier is not even an outlier in a Bayesian viewpoint.

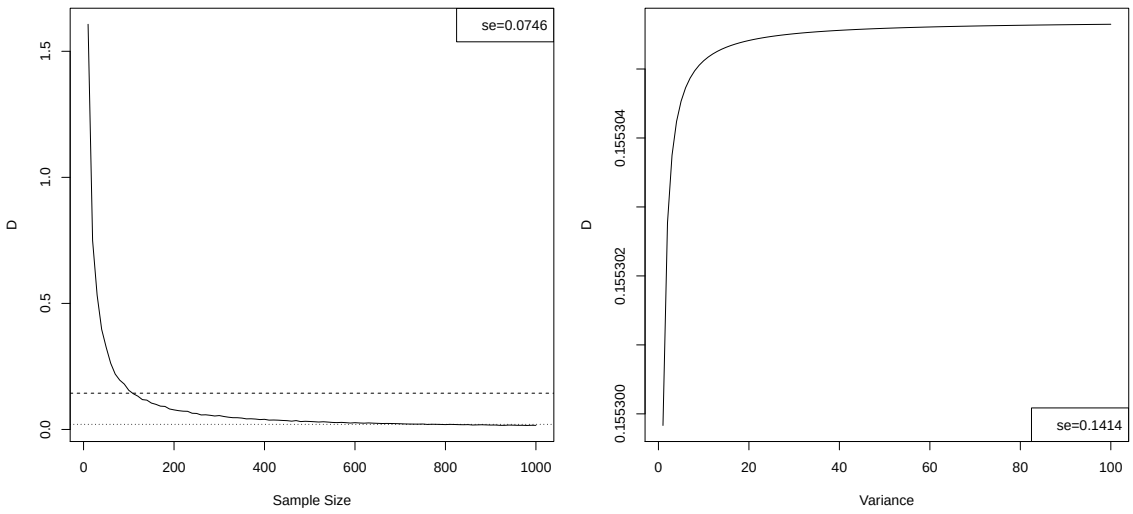

FIGURE 2.2: Trace plots of functional Bregman divergence between the full posterior and the outlier deleted posterior against sample size (left) and flatness of prior distribution (right).

Second, we study the sensitivity to flatness (variance) of prior distribution using perturbation based on case deletion. In simulation, we use the same simulation setting as above but we fix the sample size $n = 100$ and change the variance of prior σ_0^2 from 1 to 100. Figure 2.2 (right) presents the trace of the calculated divergence. From the plot, one might argue that the divergence is sensitive to the small variance of prior, while the divergence is relatively robust to the large variance. However, we notice that the difference between maximum (= 0.155305) and minimum (= 0.155299) divergences is only 6×10^{-6}. Therefore, we conclude that the functional Bregman divergence is robust to flatness or variance of priors. Note that one can also consider sensitivity analysis to variation of the likelihood instead of the prior distribution as

the following example.

Example 2.8 (Sensitivity analysis to likelihood functions)**.** Suppose $X|\theta \sim N(\theta, 1)$ and $\pi(\theta) \propto 1$, then the posterior density $p_1(\theta|X)$ is $N(X, 1)$. Consider $X|\theta \sim Cauchy(\theta, 1)$ with the same non-informative prior $\pi(\theta) \propto 1$ then the posterior density $p_2(\theta|X)$ is $Cauchy(X, 1)$. In this case, the functional Bregman divergence between p_1 and p_2, $D_\psi(p_1, p_2)$, can be applied to the sensitivity analysis to variation between the normal likelihood and the Cauchy likelihood. Hence if $D_\psi(p_1, p_2)$ provides a large value, then it implies that the posterior distribution is sensitive to the variation of the likelihood functions. The details are avoided here due to space constraint.

2.5 Applications

In this section, we consider two examples to show applicability of our methods to Bayesian outlier detection problem using case deleted perturbation in (2.6). For the calibration, we use similar method described in Peng and Dey (1995) and McCulloch (1989). Consider a biased coin with success probability p. Then the functional Bregman Divergence between an unbiased and the biased coin is given by

$$D_\psi\left(f_0, f_1\right) = \int \left\{\psi\left(f_0(x)\right) - \psi(f_1(x)) - \left(f_0(x) - f_1(x)\right)\psi'(f_1(x))\right\} d\nu(x),$$

where $f_0 = 1/2$, $f_1(x) = p^x(1-p)^{1-x}$, and ν is a counting measure for $x = 0, 1$. Define $c(p) = D_\psi(f_0, f_1)$, then $c(p)$ can be expressed as

$$c(p) = 2\psi(1/2) - \psi(p) - \psi(1-p) - (1/2-p)(\psi'(p) - \psi'(1-p)). \tag{2.23}$$

Since $c(p)$ measures the biasedness of a coin, it can be applied to determine the cut-off point of the divergence. Hence an observation is considered to be a mild outlier if the divergence with respect to the observation is in $[c(0.6), c(0.75)]$ and a strong outlier if the divergence with respect to the observation is larger than c(0.75).

2.5.1 Bayesian generalized linear model for binary response data

Bayesain GLMs have been widely used for the analysis of binary response data (Albert and Chib, 1993). In this section we apply our method to the GLMs under different links. The data set we used here is a subset of the Invasive Plant Atlas of New England (IPANE) data. There are 100 observations, which are randomly sampled from 1,789 original observations. For i^{th} observation, the response variable y_i is the binary outcome, indicating whether or not the percentage of coverage of the species is larger than zero and the covariate $\mathbf{x}_i$ is the maximum temperature of the warmest month; see Mehrhoff et al. (2003); Ibáñez et al. (2009); Wang and Dey (2010) for more details on the data. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)^{\mathbf{T}}$ denote an $n \times 1$ vector of n independent Bernoulli random variables with success probability $p_i =$ Probability$(y_i = 1)$. Suppose $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathbf{T}}$ is a $p \times 1$ vector of covariates for $i = 1, \ldots, n$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathbf{T}}$ is a $p \times 1$ vector of regression coefficients. In the GLMs, the success probability p_i is defined by

$$h(p_i) = \mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}, \tag{2.24}$$

where, $h(\cdot)$ is a link function. For a Bayesian model, a choice of a prior for $\boldsymbol{\beta}$ is required, we use $N(\mathbf{b}_0, \boldsymbol{\Sigma}_0)$, which is a commonly used prior (Gelfand and Ghosh, 2000). We consider $\mathbf{b}_0 = \mathbf{1}$ and $\boldsymbol{\Sigma}_0 = 1000\mathbf{I}$ with the following link functions for the IPANE data.

1. Logit link

   In GLMs, the most commonly used link function is the canonical link. For binary response data, the canonical link function is a symmetric link and given as $h(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$, which is the logit link. Under this link, we have

   $$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta} \quad \text{or} \quad p_i = \frac{\exp(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta})}.$$

   Using the prior $\boldsymbol{\beta} \sim N(\mathbf{b}_0, \boldsymbol{\Sigma}_0)$, the posterior distribution is obtained as

   $$\pi(\boldsymbol{\beta}|\text{data}) \propto \exp\left[\sum_{i=1}^{n}\left\{y_i\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta} - \log\left(1+e^{\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}}\right)\right\} - \frac{1}{2}(\boldsymbol{\beta}-\mathbf{b}_0)^{\mathbf{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\mathbf{b}_0)\right].$$

2. Probit link

   Another symmetric link is the probit link, which is obtained by setting $h(p_i) = \Phi^{-1}(p_i)$, where Φ^{-1} is the inverse function of the standard normal cdf. Under the probit link, we have

   $$\Phi^{-1}(p_i) = \mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta} \quad \text{or} \quad p_i = \Phi(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}).$$

Using the prior $\boldsymbol{\beta} \sim N(\mathbf{b}_0, \boldsymbol{\Sigma}_0)$, the posterior density is given as

$$\begin{aligned}\pi(\boldsymbol{\beta}|\text{data}) \quad \propto \quad & \exp\left[\sum_{i=1}^{n}\left\{y_i \log\left(\frac{\Phi(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta})}{1-\Phi(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta})}\right) + \log\left(1-\Phi(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta})\right)\right\}\right] \\ & \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{b}_0)^{\mathbf{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\mathbf{b}_0)\right\}.\end{aligned}$$

3. Complementary log-log link

   One of the most popular asymmetric link is a complementary log-log (clog-log) link, which is specified as $h(p_i) = \log\{-\log(1-p_i)\}$. Under the clog-log link, we have

$$\log\{-\log(1-p_i)\} = \mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta} \quad \text{or} \quad p_i = 1 - \exp\left\{-\exp(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta})\right\}.$$

   Using the prior $\boldsymbol{\beta} \sim N(\mathbf{b}_0, \boldsymbol{\Sigma}_0)$, the posterior distribution is obtained as

$$\begin{aligned}\pi(\boldsymbol{\beta}|\text{data}) \quad \propto \quad & \exp\left[\sum_{i=1}^{n}\left\{y_i \log\left(e^{\exp(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta})}-1\right) - \exp(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta})\right\}\right] \\ & \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{b}_0)^{\mathbf{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\mathbf{b}_0)\right\}.\end{aligned}$$

In order to approximate a functional Bregman divergence, we use the proposed approximation methods: Gaussian approximation and IWMDE. In the Gaussian approximation, Nelder-Mead algorithm is applied for finding the posterior mode. Importance sampling approximation within the Gaussian approximation is also implemented. For the implementation of IWMDE, we use $5,000$ samples (after $5,000$ burn-in iterations) from the posterior distribution using Metropolis-Hastings algorithm (Chib and Greenberg, 1995), where a multivariate Gaussian distribution is

used as a proposal distribution. For each link, the plot of estimated divergences with a convex function $\psi_{\alpha=1}$ (KL-divergence) against deleted observations is presented in Figure 2.3, where dotted line and dashed line indicate cut-off points for mild outliers and severe outliers respectively. According to Figure 2.3, we assure that our approximation methods provide the same result even though the estimated divergences are different in the case of severe outliers. For the clog-log link in Figure 2.3 (bottom), only 66^{th} observation is considered as the severe outlier while 66^{th} and 97^{th} observations are considered as severe outliers for logit link in Figure 2.3 (top left) as well as probit link in Figure 2.3 (top right).

2.5.2 Bayesian circular data analysis

In this section, we will apply our methods for circular data analysis under the Bayesian framework. One of the common methods for circular data is the intrinsic approach using von Mises distribution. Suppose that data $y_1, ..., y_n$ are observations from a random variable Y which follows von Mises distribution with the direction parameter $\mu \in [0, 2\pi)$ and the concentration parameter $\kappa > 0$. The density function of Y is

$$f(y|\mu, \kappa) = \frac{\exp\{\kappa \cos(y - \mu)\}}{2\pi I_0(\kappa)}, \tag{2.25}$$

where $I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} \exp\{\kappa \cos(y)\}dy$. Assume that μ is unknown but κ is known, then the likelihood function is given as

$$f(\mathbf{y}|\mu) \propto \exp\left\{\kappa \sum_{i=1}^{n} \cos(y_i - \mu)\right\}. \tag{2.26}$$

Using Guttorp and Lockhart (1988)'s conjugate prior $\pi(\mu) \propto \exp\{a_0 \cos(\mu - b_0)\}$,

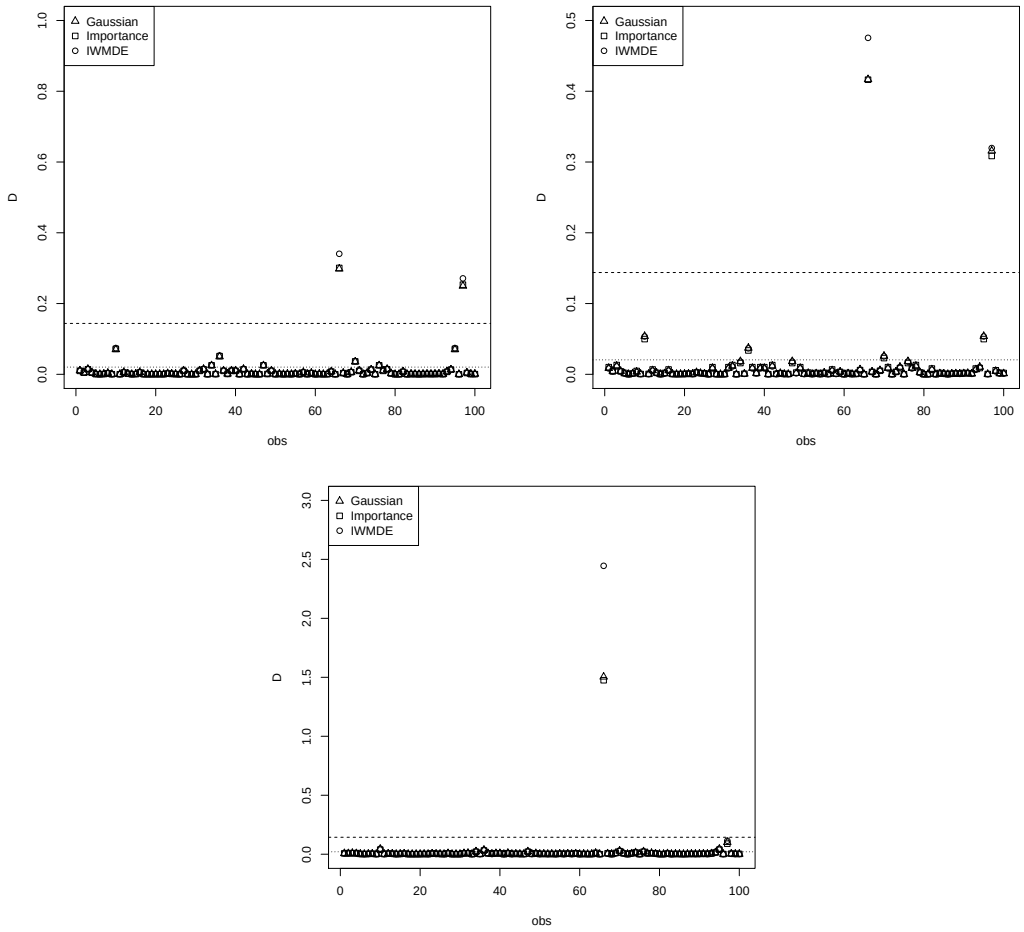

FIGURE 2.3: Plots of estimated divergence with a convex function $\psi_{\alpha=1}$ (KL-divergence) against deleted observations in logit (top left), probit (top right), and clog-log (bottom) links for IPANE data.

where $a_0 > 0$ and $0 \leq b_0 < 2\pi$, the posterior distribution is given as

$$\pi(\mu|\mathbf{y}) \propto \exp\{a\cos(\mu - b)\}, \tag{2.27}$$

where $a = \kappa\left(a_0\sin(b_0) + \sum_{i=1}^{n}\sin(y_i)\right)$ and $b = \arctan\left(\frac{a_0\sin(b_0)+\sum_{i=1}^{n}\sin(y_i)}{a_0\cos(b_0)+\sum_{i=1}^{n}\cos(y_i)}\right)$. Note the posterior density is von Mises distribution with the concentration a and the

direction b, so that we can directly sample from the posterior distribution. For circular data, the Gaussian approximation is inappropriate because the observation has a value (or direction) between 0 and 2π. Hence, we only consider IWMDE in this section. To illustrate our method, we use the movements of turtles data (Stephens, 1969). The data set consists of directions of movements taken by 76 turtles after treatment. The data originally collected by Dr. E. Gould of Johns Hopkins school of Hygiene. Later Stephens (1969) first cited the data in the study of a directional data modeling. They fitted von Mises distribution to find a preferred direction of the turtle's movement. The plot of turtles data in Figure 2.4 (left) provides that the turtles could have a preferred direction. Similarly, we fit Von Mises model in (2.25) but we fix the concentration parameter $\kappa = 1.1423$ (MLE of κ). After fitting the model, we implement the outlier detection technique using IWMDE method with $5,000$ samples from (2.27), where we set $a_0 = 0.01$ and $b_0 = 0$. Figure 2.4 (right) shows a plot of the estimated divergence for the convex function $\psi_{\alpha=2}$ ($L^2/2$ Euclidean distance) against deleted observations along with cut-off points for mild outliers (dotted line) and severe outliers (dashed line). Estimated divergences for the outliers are shown in Table 2.1. It shows that there is no severe outlier but 1^{st}, 57^{th}, 58^{th}, 59^{th}, 60^{th}, 74^{th}, 75^{th}, and 76^{th} observations are considered as mild outliers. We denote the outliers as triangle symbol ($\triangle$) in Figure 2.4 (left). Interestingly, turtles that moved toward left or right side of the main direction are outliers, but not turtles that went toward opposite direction. This is due to the fact that the turtles toward opposite direction are influential in the concentrate parameter and in this example we assumed that the parameter is fixed.

TABLE 2.1: The estimated divergences of mild outliers for turtle data.

Obs	1	57	58	59	60	74	75	76
D	0.0102	0.0134	0.0144	0.0144	0.0143	0.0133	0.0142	0.0135

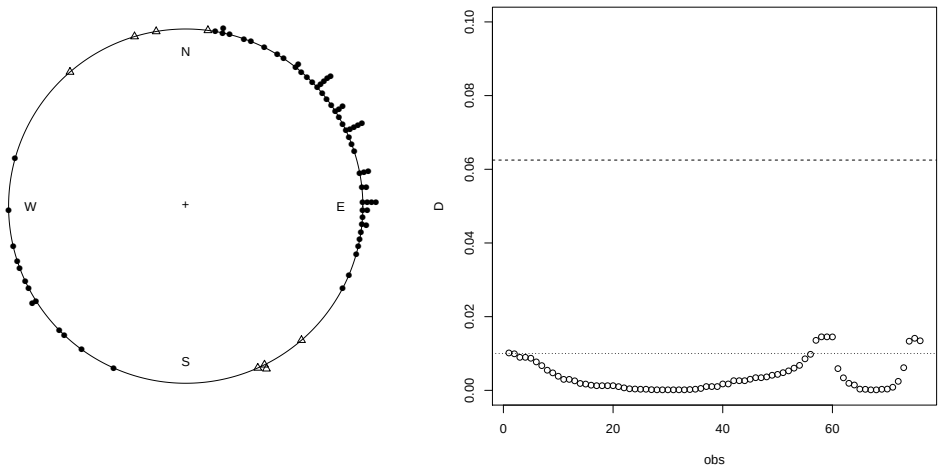

FIGURE 2.4: Circular plot of turtles data (left) and plot of estimated functional Bregman divergence ($L^2/2$ Euclidean distance) against deleted observations for turtles data (right).

2.6 Conclusion and remarks

This chapter is motivated by the fact that comparison of posterior distributions provides model diagnostics measures in a Bayesian viewpoint. The diagnostics measure can be specified by defining a perturbation on the prior distribution (sensitivity analysis) or the likelihood (outlier detection). Two different comparison approaches based on functional Bregman divergence have been introduced in this chapter. We demonstrate that comparison of posteriors based on f-divergence is equivalent to comparison of rate of posteriors in view of a functional Bregman divergence. We propose a direct comparison approach based on functional Bregman divergence. Two

approximation techniques; Gaussian approximation and IWMDE, are developed to estimate the functional Bregman divergence. If the posterior can be approximated by the Gaussian approximation and the sample size is relatively small, say less than 100, we recommend the Gaussian approximation method because it is much faster and accurate. If you have a large sample size or a model that has a complex structure, then you should use the IWMDE due to the fact it is efficient and flexible for such models. In advance, the idea of the posterior comparison using functional Bregman divergence can be extended in several directions depending on several other choices of perturbation schemes. Also, one can apply our outlier detection technique to model selection problems since the model with less outliers is more appropriate for the given data set.

Chapter 3

Bayesian Model Assessment and Selection using Bregman Divergence

3.1 Introduction

In a Bayesian framework, predictive distributions have been utilized as the essential ingredient for the development of the Bayesian model selection criteria. For instance, the Bayes Factor, used while comparing two candidate models for the data, is the ratio of the marginals (or prior predictive densities) of the two models and represents an intuitively appealing tool for selecting between the models. Stone (1974) and Geisser (1975) proposed the pseudo-Bayes Factor based on the conditional predictive densities (or posterior predictive densities) coupled with the idea of leave-one-out cross-validation. In similar spirit, Berger and Pericchi (1996b) introduced the Intrinsic Bayes Factor based on the posterior predictive distributions characterized by the

notion of minimal training samples to ensure the properness of the predictive densities.

Our primary goal here is to develop a general tool for Bayesian model selection and assessment using predictive distributions. To unify and extend many existing predictive model selection and assessment procedures, we consider various predictive distributions using a more flexible conditioning scheme and introduce fairly general summary measures of the evaluated predictive densities at the observed data. Developing the summary measures is directly related to developing a scoring rule which assigns a numerical score based on the predictive distribution and observed data. In general, there are many kinds of scoring rules available such as logarithmic score, quadratic score, and ranked probability score. From a Bayesian perspective, we adopt a scoring rule introduced by Grünwald and Dawid (2004) using Bregman divergence (Bregman, 1967). This rule offers a general approach to scoring in the sense that the Bregman divergence includes, as special cases, many well-known loss functions such as squared error loss, Kullback-Leibler (KL) divergence, and Mahalanobis distance.

In a frequentist framework, the predictive density approach for model selection provides an attractive tool to incorporate a scoring rule because the predictive density can be easily calculated by plug-in method. However, the calculation of the predictive density in the Bayesian framework is a challenge due to the fact that, in many cases, it is not available in closed form. Several attempts have been made to alleviate this problem. Gelfand and Dey (1994) proposed a Monte Carlo estimator for several types of conditional predictive densities with independent data. Chen (1994) developed a relatively accurate estimation method for the prior predictive (or marginal) density. As an extension of Gelfand and Dey (1994) and Chen (1994), we introduce a general Monte Carlo estimator to calculate various predictive densities.

Some remarks are due on the notation and the assumptions used in this chapter.

We use $f(\mathbf{y}|\boldsymbol{\theta})$, $m(\mathbf{y})$, $p(\mathbf{y}_1|\mathbf{y}_2)$, $\pi(\boldsymbol{\theta})$, and $\pi(\boldsymbol{\theta}|\mathbf{y})$ to denote, respectively, the likelihood function, the marginal density (or prior predictive density), conditional predictive density (or posterior predictive density), the prior, and the posterior, where $\mathbf{y} = (y_1, \ldots, y_n)$ is the observed data and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ represents the vector of parameters such that $\boldsymbol{\theta} \in \Theta$ a suitable parameter space. We remark here that there is no restriction on the nature of the posterior, i.e., the posterior can be a density with respect to the Lebesgue measure or with respect to a counting measure or can also be a mixture of densities. For notational convenience, the distributions, Normal, Inverse-Gamma, Inverse-Wishart, and Uniform are respectively denoted as N, IG, IW, and U.

The outline of the remainder of the chapter is as follows. In Section 3.2 we introduce our main construct: the Bregman Divergence Criterion (BDC) that is constructed from the predictive model selection perspective; also, importantly, the calculation of the conditional predictive density for a particular candidate model is detailed with some examples. In Section 3.3 we propose a calibration tool for the BDC based on the (generalized) probability integral transform under the Bayesian setup; furthermore, a sampling method from the conditional predictive distribution is also provided. Simulation studies are conducted with a linear regression model and a longitudinal data model in Section 3.4. Section 3.5 offers some concluding remarks.

3.2 Model selection using Bregman divergence

3.2.1 Predictive model selection and decision making

The predictive model selection problem can be considered as a decision making problem; this point of view was laid out in the works of Grünwald and Dawid (2004) and Gneiting and Raftery (2007). In other words, with respect to a loss function, the optimal model among proposed models is the one whose predictive density is "closest" to the true density. To formalize the idea, let M^{true} be the true model and $\mathcal{M} = \{M^1, \ldots, M^K\}$ be a finite set of proposed models, where K is assumed to be known and further there is no preferred model. Suppose that $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m$ are future observations corresponding to the observed $\mathbf{y}_1, \ldots, \mathbf{y}_m$, respectively. Let $\mathbf{p}^{\text{true}}$ be the vector of the true conditional predictive densities of $\tilde{\mathbf{y}}$ given $\mathbf{y}$ such that $\mathbf{p}^{\text{true}} = [p(\tilde{\mathbf{y}}_1|\mathbf{y}_1, M^{\text{true}}), \ldots, p(\tilde{\mathbf{y}}_m|\mathbf{y}_m, M^{\text{true}})]$. For the model M^k, define the vector of its conditional predictive densities by $\mathbf{p}^k = \left[p(\tilde{\mathbf{y}}_1|\mathbf{y}_1, M^k), \ldots, p(\tilde{\mathbf{y}}_m|\mathbf{y}_m, M^k)\right]$, for $k = 1, \ldots, K$. Then the predictive model selection problem can be formulated in the following manner: find the model that has a minimum "dissimilarity" between $\mathbf{p}^{\text{true}}$ and $\mathbf{p}^k$; i.e., determine M^* such that

$$M^* = \arg \min_{M^k \in \mathcal{M}} D\left(\mathbf{p}^k, \mathbf{p}^{\text{true}}\right), \tag{3.1}$$

where $D(\cdot, \cdot)$ is an appropriate divergence measure. In view of Bayesian decision theory, this strategy looks very simple and reasonable, but two serious problems occur in practice: first, since the true model M^{true} is unknown, consequently $\mathbf{p}^{\text{true}}$ is unknown; and second, the future observations $\tilde{\mathbf{y}}_i$'s are not available in many cases.

Since the observations, $\tilde{\mathbf{y}}_i$'s and $\mathbf{y}_i$'s, are generated from the true model, we expect

that the true model provides the best predictive performance: the maximum predictive density will be attained under the true model, i.e., $p(\tilde{\mathbf{y}}_i|\mathbf{y}_i, M^{\text{true}}) \geq p(\tilde{\mathbf{y}}_i|\mathbf{y}_i, M^k)$ for any i and k. Motivated by the aforementioned aspect, as an alternative of (3.1), we define the optimal model M^* by

$$M^* = \arg \max_{M^k \in \mathcal{M}} D\left(\mathbf{p}^{\text{worst}}, \mathbf{p}^k\right),$$

where $\mathbf{p}^{\text{worst}} = [p(\tilde{\mathbf{y}}_1|\mathbf{y}, M^{\text{worst}}), \ldots, p(\tilde{\mathbf{y}}_m|\mathbf{y}, M^{\text{worst}})]$ satisfies

$$p(\tilde{\mathbf{y}}_i|\mathbf{y}_i, M^{\text{worst}}) < p(\tilde{\mathbf{y}}_i|\mathbf{y}_i, M^k) \tag{3.2}$$

for any $i = 1, \ldots, m$ and $k = 1, \ldots, K$. The model M^{worst} in (3.2) can be interpreted as the worst model due to the fact that its conditional predictive density is always smaller than any proposed models. Consequently, the farthest model M^* from the worst model M^{worst} will be the closest model to the true model M^{true}. Note that probability density (or mass) functions are bounded below by zero. From this fact, the worst model M^{worst} can be simply defined by assigning zero probability density (or mass) for all observations, i.e., $\mathbf{p}^{\text{worst}} = (0, \ldots, 0)$. Therefore, the use of $\mathbf{p}^{\text{worst}}$ allows us to avoid the determination of $\mathbf{p}^{\text{true}}$.

For the second issue, cross-validation methods for independent data have been proposed by several authors; see for example, Stone (1974), Geisser (1975), and Geisser and Eddy (1979). For dependent data, Dawid (1984) proposed prequential approach in time series context. In this chapter, we consider a similar strategy as in Gelfand and Dey (1994), that provides more flexible grouping and conditioning schemes than partitioning approaches in the cross-validation methods. Suppose that $y_1, \ldots, y_n$ are

observed. Let $\{\mathbf{y}_{s_1}\}, \ldots, \{\mathbf{y}_{s_m}\}$ be $m(\leq n)$ subsets of $\{\mathbf{y}\} = \{y_1, \ldots, y_n\}$ such that $\{\mathbf{y}\} = \cup_{i=1}^{m}\{\mathbf{y}_{s_i}\}$. Define $\{\mathbf{y}_{(s_i)}\}$ for $i = 1, \ldots, m$, such that $\{\mathbf{y}_{(s_i)}\} \subseteq \{\mathbf{y}_{s_i}\}^c$, i.e., $\{\mathbf{y}_{(s_i)}\}$ is a subset of the complement of $\{\mathbf{y}_{s_i}\}$ for each i. Now, we treat $\{\mathbf{y}_{s_i}\}$'s and $\{\mathbf{y}_{(s_i)}\}$'s as the future observations and observed data sets, respectively.

Obviously, if the true conditional predictive densities are known and the future observations are available then the equation (3.1) can be directly applied, but it is unrealistic situation in many cases. In this chapter, we assume throughout that the true model is unknown and no future observation exists, so that our methodology under this premise is more representative of the nature of the problems faced when employing Bayesian techniques in model selection problems. Under the assumption, our methodology stands on finding M^* such that

$$M^* = \arg \max_{M^k \in \mathcal{M}} D\left(\mathbf{p}^{\text{worst}}, \mathbf{p}^k\right), \tag{3.3}$$

where $\mathbf{p}^k = (p_1^k, \ldots, p_m^k)$ with $p_i^k = p(\mathbf{y}_{s_i}|\mathbf{y}_{(s_i)}, M^k)$ and $\mathbf{p}^{\text{worst}} = (0, \ldots, 0)$.

3.2.2 Bregman divergence criterion

To employ the maximum distance approach in (3.3), the appropriate divergence measure should be determined. We consider a general class of divergence measures, called Bregman divergence. It is worthwhile to note that the Bregman divergence reduces to various divergence measures according to the choice of the convex function ψ; few illuminating examples are enumerated below.

Example 3.1. Suppose $\psi(\mathbf{x}) = \sum_{i=1}^{m}\{x_i \log x_i\}$, then the Bregman divergence is

given as

$$\begin{aligned} BD_{\psi}(\mathbf{x},\mathbf{y}) &= \sum_{i=1}^{m}\{x_i \log x_i\} - \sum_{i=1}^{m}\{y_i \log y_i\} - \sum_{i=1}^{m}\{(x_i - y_i)(\log y_i + 1)\} \\ &= \sum_{i=1}^{m}\left\{x_i \log \frac{x_i}{y_i} - x_i + y_i\right\}, \end{aligned}$$

which is the generalized Kullback-Leibler divergence between $\mathbf{x}$ and $\mathbf{y}$.

Example 3.2. Suppose $\psi(\mathbf{x}) = \sum_{i=1}^{m}\{-\log x_i\}$, then the Bregman divergence is written as

$$\begin{aligned} BD_{\psi}(\mathbf{x},\mathbf{y}) &= \sum_{i=1}^{m}\{-\log x_i\} - \sum_{i=1}^{m}\{-\log y_i\} - \sum_{i=1}^{m}\{(x_i - y_i)(-1/y_i)\} \\ &= \sum_{i=1}^{m}\left\{\frac{x_i}{y_i} - \log \frac{x_i}{y_i} - 1\right\}, \end{aligned}$$

which is called Itakura-Saito distance (Itakura and Saito, 1970) between $\mathbf{x}$ and $\mathbf{y}$.

Example 3.3. Suppose $\psi(\mathbf{x}) = \mathbf{x^T A x}$, where $\mathbf{A}$ is a positive definite matrix, then Bregman divergence is given as

$$\begin{aligned} BD_{\psi}(\mathbf{x},\mathbf{y}) &= \mathbf{x^T A x} - \mathbf{y^T A y} - \langle \mathbf{x} - \mathbf{y}, 2\mathbf{A y}\rangle \\ &= (\mathbf{x} - \mathbf{y})^{\mathbf{T}}\mathbf{A}(\mathbf{x} - \mathbf{y}). \end{aligned}$$

If $\mathbf{A}$ is assumed to be the inverse of the covariance matrix, then the Bregman divergence is the Mahalanobis distance between $\mathbf{x}$ and $\mathbf{y}$. If we assume that $\mathbf{A}$ is the identity matrix, then the Bregman divergence reduces to the squared Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ such that $BD_{\psi}(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$, where $\|\mathbf{x}\| = \sqrt{\mathbf{x^T x}}$.

Consequently, the dissimilarity can be measured by various measuring devices (or

loss functions). Therefore, using the Bregman divergence in (1.1), we reduce (3.3) to

$$\begin{aligned} M^* &= \arg\max_{M^k \in \mathcal{M}} \left\{ \psi(\mathbf{p}^{\text{worst}}) - \psi(\mathbf{p}^k) - \left\langle \mathbf{p}^{\text{worst}} - \mathbf{p}^k, \nabla\psi(\mathbf{p}^k) \right\rangle \right\} \\ &= \arg\max_{M^k \in \mathcal{M}} \left\{ \left\langle \mathbf{p}^k, \nabla\psi(\mathbf{p}^k) \right\rangle - \psi(\mathbf{p}^k) \right\}, \end{aligned} \tag{3.4}$$

for a suitable convex function ψ. From (3.4), we now define a new Bayesian model selection criterion, called Bregman divergence criterion (BDC).

Definition 3.4 (Bregman Divergence Criterion)**.**

$$BDC_\psi(\mathbf{p}^k) = \left\langle \mathbf{p}^k, \nabla\psi(\mathbf{p}^k) \right\rangle - \psi(\mathbf{p}^k), \tag{3.5}$$

where $\mathbf{p}^k = (p_1^k, \ldots, p_m^k)$ with $p_i^k = p(\mathbf{y}_{s_i} | \mathbf{y}_{(s_i)}, M^k)$.

The defined BDC in (3.5) measures the dissimilarity (with respect to a certain loss function in Bregman divergence) between the vector of predictive densities of the proposed model M^k and the vector of predictive densities of the worst model which has zero densities at the hypothetical future data given the observed data. Hence the *optimal model* has the largest BDC among the proposed models. It is worth noting that BDC includes many well-known scoring rules that evaluate the quality of probabilistic forecasts (Gneiting and Raftery, 2007). For example, let $\psi_{\log}(\mathbf{x}) = -\sum_{i=1}^m \log x_i - m$, where m indicates the dimension of $\mathbf{x}$, then BDC reduces to *logarithm score* as follows:

$$BDC_{\psi_{\log}}(\mathbf{p}^k) = \sum_{i=1}^m \log p_i^k. \tag{3.6}$$

Note that, depending on the definition of $\mathbf{p}^k$, the BDC in (3.6) is equivalent to the

several existing Bayesian model selection methods. For notational convenience, we denote $\{y_1, \ldots, y_{i-1}\} = \{\mathbf{y}_{1:i-1}\}$, where $\{\mathbf{y}_{1:0}\} = \emptyset$ and $\{\mathbf{y}_s\}^c = \{\mathbf{y}_{-s}\}$).

Example 3.5. Let $\{\mathbf{y}_{s_i}\} = \{y_i\}$ and $\{\mathbf{y}_{(s_i)}\} = \{\mathbf{y}_{1:i-1}\}$ for $i = 1, \ldots, n$. In this set-up, the i^{th} component of $\mathbf{p}^k$ is defined as $p_i^k = p(y_i|\mathbf{y}_{1:i-1}, M^k)$. Then the BDC with $\psi_{\log}$ is equivalent to the prequential approach or the Bayes Factor as follows:

$$BDC_{\psi_{\log}}(\mathbf{p}^k) = \sum_{i=1}^{n} \log p(y_i|\mathbf{y}_{1:i-1}, M^k) = \log\left\{m(\mathbf{y}|M^k)\right\}.$$

Example 3.6. Define $\{\mathbf{y}_{s_i}\} = \{y_i\}$ and $\{\mathbf{y}_{(s_i)}\} = \{\mathbf{y}_{-i}\}$. In this set-up, the BDC with $\psi_{\log}$ is equivalent to the pseudo-Bayes factor (PSBF) such that

$$BDC_{\psi_{\log}}(\mathbf{p}^k) = \sum_{i=1}^{n} \log p(y_i|\mathbf{y}_{-i}, M^k),$$

where $p(y_i|\mathbf{y}_{-i}, M^k)$ is called the *conditional predictive ordinate* (CPO).

Example 3.7. Suppose that $\{\mathbf{y}_{(s_i)}\}$ is a minimal subset (Berger and Pericchi, 1996b) and $\{\mathbf{y}_{s_i}\} = \{\mathbf{y}_{-(s_i)}\}$. Then, the BDC with $\psi_{\log}$ is equivalent to the Intrinsic Bayes factor (IBF) such that:

$$BDC_{\psi_{\log}}(\mathbf{p}^k) = \sum_{i=1}^{n} \log p(\mathbf{y}_{-(s_i)}|\mathbf{y}_{(s_i)}).$$

In addition, the BDC forms various scoring rules according to a choice of the convex function, see Table 3.1. Therefore, our proposed method is not only unifying existing methods, but also providing a general extension for Bayesian model selection.

TABLE 3.1: Examples of BDC induced by some convex functions ψ's.

$\psi(\mathbf{x})$	BDC	Scoring Rule
$\|\|\mathbf{x}\|\|^2$	$\|\|\mathbf{p}\|\|^2$	quadratic score
$\mathbf{x^T W x}$	$\mathbf{p^T W p}$	weighted quadratic score
$-\sum_{i=1}^m \log x_i - m$	$\sum_{i=1}^m \log p_i$	logarithm score
$\frac{1}{m}\sum_{i=1}^m x_i \log x_i$	$\frac{1}{m}\sum_{i=1}^m p_i$	mean score
$\frac{1}{m}\sum_{i=1}^m w_i x_i \log x_i$	$\frac{1}{m}\sum_{i=1}^m w_i p_i$	weighted mean score
$\sum_{i=1}^m \{e^{-cx_i} - 1\}$	$\sum_{i=1}^m \{e^{-cp_i}(e^{cp_i} - cp_i - 1)\}$	weighted Linex score

3.2.3 Calculation of Bregman divergence criterion

For the model M^k, calculation of the BDC in (3.5) is straightforward once the conditional predictive densities $p_i^k = p(\mathbf{y}_{s_i}|\mathbf{y}_{(s_i)}, M^k)$ for $i = 1, \ldots, m$ are obtained, because the BDC is calculated by substituting the obtained p_i^k's in (3.5). Therefore, in this section, we discuss mainly the calculation of the conditional predictive densities p_i based on a Monte Carlo (MC) method. For notational simplicity, we omit the index k and the model M^k in this section. For example, $p_i^k = p(\mathbf{y}_{s_i}|\mathbf{y}_{(s_i)}, M^k)$ will be denoted by $p_i = p(\mathbf{y}_{s_i}|\mathbf{y}_{(s_i)})$.

In some cases, if one is fortunate enough to obtain marginal densities, $m(\mathbf{y}_{s_i}, \mathbf{y}_{(s_i)})$ and $m(\mathbf{y}_{(s_i)})$, in closed-form, straightforward algebra informs us that p_i can be calculated as

$$p_i = p(\mathbf{y}_{s_i}|\mathbf{y}_{(s_i)}) = \begin{cases} m(\mathbf{y}_{s_i}, \mathbf{y}_{(s_i)})/m(\mathbf{y}_{(s_i)}) & \text{if} \quad m(\mathbf{y}_{(s_i)}) > 0 \\ m(\mathbf{y}_{s_i}) & \text{if} \quad m(\mathbf{y}_{(s_i)}) = 0 \end{cases}. \tag{3.7}$$

To fix ideas, the following example illustrates the above mentioned scenario in the case of the linear regression model with independent errors. We remark here that, for

purposes of notational simplicity, $\{\mathbf{y}_{s_i}, \mathbf{y}_{(s_i)}\}$ is denoted by $\{\mathbf{y}_{s_i \cup (s_i)}\}$.

Example 3.8 (Linear Regression Model)**.** Consider the linear model given in Berger and Pericchi (1996a). Suppose that model M is given as

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

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^{\mathbf{T}}$ and σ^2 are unknown, $\mathbf{y} = (y_1, \ldots, y_n)^{\mathbf{T}}$ is the data vector, $\mathbf{X}$ is the $(n \times q)$ design matrix with rank $q(< n)$, and $\mathbf{x}_i$ is the i^{th} row of $\mathbf{X}$. Consider the non-informative prior $\pi(\boldsymbol{\beta}, \sigma^2) = 1/\sigma^2$. Let n_i and $\tilde{n}_i$ be the number of observations in $\mathbf{y}_{s_i}$ and $\mathbf{y}_{(s_i)}$, respectively. With a straightforward calculation, the conditional predictive density p_i is obtained in a closed form as follows: if $n_i \geq q$ and $\tilde{n}_i = 0$, then

$$p_i = m(\mathbf{y}_{s_i}) \ = \ \frac{(\pi)^{-\frac{n_i - q}{2}} \Gamma\left(\frac{n_i - q}{2}\right)}{\left|\mathbf{X}_{s_i}^{\mathbf{T}} \mathbf{X}_{s_i}\right|^{1/2} \left[(\mathbf{y}_{s_i} - \hat{\mathbf{y}}_{s_i})^{\mathbf{T}} (\mathbf{y}_{s_i} - \hat{\mathbf{y}}_{s_i})\right]^{\frac{n_i - q}{2}}},$$

and if $\tilde{n}_i \geq q$, then

$$p_i = \frac{\Gamma\left(\frac{n_i + \tilde{n}_i - q}{2}\right)}{\Gamma\left(\frac{\tilde{n}_i - q}{2}\right)} \left[\frac{\left|\mathbf{X}_{(s_i)}^{\mathbf{T}} \mathbf{X}_{(s_i)}\right|}{\pi^{n_i} \left|\mathbf{X}_{s_i \cup (s_i)}^{\mathbf{T}} \mathbf{X}_{s_i \cup (s_i)}\right|}\right]^{1/2} \times \frac{\left[(\mathbf{y}_{(s_i)} - \hat{\mathbf{y}}_{(s_i)})^{\mathbf{T}} (\mathbf{y}_{(s_i)} - \hat{\mathbf{y}}_{(s_i)})\right]^{\frac{\tilde{n}_i - q}{2}}}{\left[(\mathbf{y}_{s_i \cup (s_i)} - \hat{\mathbf{y}}_{s_i \cup (s_i)})^{\mathbf{T}} (\mathbf{y}_{s_i \cup (s_i)} - \hat{\mathbf{y}}_{s_i \cup (s_i)})\right]^{\frac{n_i + \tilde{n}_i - q}{2}}},$$

where $\hat{\mathbf{y}}_s = \mathbf{X}_s (\mathbf{X}_s^{\mathbf{T}} \mathbf{X}_s)^{-1} \mathbf{X}_s^{\mathbf{T}} \mathbf{y}_s$ for $s = (s_i)$ or $s = s_i \cup (s_i)$. Note that p_i is undefined (improper) when $0 < \tilde{n}_i < q$ or $\tilde{n}_i = 0$ and $n_i < q$.

The example above represents a special case where the marginal densities are available in closed form. However, in general, calculation of the conditional predictive

density p_i in (3.7) is not an easy task within a Bayesian framework; pertinently, even if they are amenable to calculation via numerical methods, the huge computational burden associated with the task creates a significant obstacle; this is so since the conditional predictive density needs to be calculated for each sub-sample $\{\mathbf{y}_{s_i}\}$, $i = 1, \ldots, m$. In view of that, we propose a MC estimator of conditional predictive density p_i for each i, based on a single set of Markov chain Monte Carlo (MCMC) samples from the stationary full posterior distribution; in other words, all the required conditional predictive densities are simultaneously computed by the set of MCMC samples. This approach considerably mitigates the computational burden, in the general Bayesian framework, since the samples from the posterior should be already generated for the inference procedure in the first place. We formalize the preceding discussion with the following theorem.

Theorem 3.9. *Suppose $g(\cdot)$ is a probability density function satisfying $supp(g) \subset supp(\pi)$. Let $\mathbf{1}(A)$ be the indicator function for the set A and $f(\mathbf{y}|\boldsymbol{\theta})$ be the likelihood function. If $\{\boldsymbol{\theta}^j\}_{j=1}^N$ is a set of MCMC samples from a full posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ and $p(\mathbf{y}_s|\mathbf{y}_{-s}) < \infty$, where $\{\mathbf{y}_{-s}\} = \{\mathbf{y}_s\}^c$, then*

$$\lim_{N\to\infty} \left[\frac{1}{N} \sum_{j=1}^{N} \left\{ \frac{1}{f(\mathbf{y}_s|\mathbf{y}_{-s}, \boldsymbol{\theta}^j)} \left(\frac{g(\boldsymbol{\theta}^j)}{\pi(\boldsymbol{\theta}^j)} \right)^{\mathbf{1}(\{\mathbf{y}_{-s}\}=\emptyset)} \right\} \right]^{-1} \overset{a.s.}{=} p(\mathbf{y}_s|\mathbf{y}_{-s}). \tag{3.8}$$

Note that the conditional predictive density in (3.7) is re-written as

$$p_i = p(\mathbf{y}_{-(s_i)}|\mathbf{y}_{(s_i)}) \frac{m(\mathbf{y}_{s_i \cup (s_i)})}{m(\mathbf{y})}. \tag{3.9}$$

From Theorem 3.9 and (3.9), the MC estimator of the conditional predictive density

is given by

$$\hat{p}_i = \left\{ \frac{1}{N} \sum_{j=1}^{N} \frac{f(\mathbf{y}_{s_i \cup (s_i)} | \boldsymbol{\theta}^j)}{f(\mathbf{y} | \boldsymbol{\theta}^j)} \right\} \left\{ \frac{1}{N} \sum_{j=1}^{N} \frac{\{g(\boldsymbol{\theta}^j)/\pi(\boldsymbol{\theta}^j)\}^{\mathbf{1}(\{\mathbf{y}_{(s_i)}\}=\emptyset)}}{f(\mathbf{y}_{-(s_i)} | \mathbf{y}_{(s_i)}, \boldsymbol{\theta}^j)} \right\}^{-1}, \tag{3.10}$$

for $i = 1, \ldots, m$, where $g(\cdot)$ is a probability density function, $\pi(\cdot)$ is a prior density function, and $\{\boldsymbol{\theta}^j\}_{j=1}^N$ is a set of MCMC samples from the full posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$. Note that using Theorem 3.9 and (3.9), the following argument can be easily shown:

$$\hat{p}_i \overset{\text{a.s.}}{=} p(\mathbf{y}_{s_i} | \mathbf{y}_{(s_i)}).$$

If $\{\mathbf{y}_{(s_i)}\} = \emptyset$, then the determination of the density function $g(\cdot)$ in (3.10) is required. Otherwise, $g(\cdot)$ is eliminated from (3.10) due to the indicator function. In order to choose a good $g(\cdot)$, the function $g(\cdot)$ should be close to the target function $f(\mathbf{y}|\cdot)\pi(\cdot)$. Chen (1994) proposed to use a density function such that its mean and variance match the posterior mean and variance. Later, Goh and Dey (2014) showed that if $g(\cdot)$ is a normal density function then the mean and variance matching density minimizes the KL divergence between $g(\cdot)$ and $f(\mathbf{y}|\cdot)\pi(\cdot)$. Let $\phi(\cdot; \boldsymbol{\mu}, \Sigma)$ be the density function of $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$. Here, we define $g(\cdot) = \phi(\cdot; \boldsymbol{\mu}_\theta, \Sigma_\theta)$ with

$$\boldsymbol{\mu}_\theta = \frac{1}{N} \sum_{j=1}^{N} \boldsymbol{\theta}^j \quad \text{and} \quad \Sigma_\theta = \frac{1}{N-1} \sum_{j=1}^{N} (\boldsymbol{\theta}^j - \boldsymbol{\mu}_\theta)(\boldsymbol{\theta}^j - \boldsymbol{\mu}_\theta)^{\mathbf{T}},$$

where $\{\boldsymbol{\theta}^j\}_{j=1}^N$ is a set of MCMC samples from the full posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$. The following example describes the utility of our MC estimator in a generalized linear model.

Example 3.10 (Generalized Linear Model)**.** Suppose $\mathbf{y} = (y_1, \ldots, y_n)^{\mathbf{T}}$ is a vector of

independent observations such that the density of the observation y_j belongs to the natural exponential family, that is

$$f(y_j|\theta_j) = \exp\left(\theta_j y_j - a(\theta_j) + b(y_j)\right), \tag{3.11}$$

where $a(\cdot)$ and $b(\cdot)$ are known functions. Let θ_j's be related to the regression coefficients such that

$$\theta_j = h(\mathbf{x}_j^{\mathbf{T}}\boldsymbol{\beta}). \tag{3.12}$$

Then the model determined by (3.11) and (3.12) is called the generalized linear model (GLM). Let $\pi(\boldsymbol{\beta})$ be a prior density. From (3.11) and (3.12), then the posterior density of $\boldsymbol{\beta}$ is

$$\pi(\boldsymbol{\beta}|\mathbf{y}) \quad \propto \quad \exp\left[\sum_{j=1}^{n}\left\{y_j h(\mathbf{x}_j^{\mathbf{T}}\boldsymbol{\beta}) - a\left(h(\mathbf{x}_j^{\mathbf{T}}\boldsymbol{\beta})\right)\right\} + \log\left\{\pi(\boldsymbol{\beta})\right\}\right]. \tag{3.13}$$

Let $\left\{\boldsymbol{\beta}^l\right\}_{l=1}^{N}$ be a set of MCMC samples from the full posterior density in (3.13). The MC estimate in (3.10) is given as

$$\begin{aligned}
\hat{p}_i \quad = \quad & \left[\frac{1}{N}\sum_{l=1}^{N}\exp\left\{\sum_{j=1}^{n} d_{ij}\left\{a\left(h(\mathbf{x}_j^{\mathbf{T}}\boldsymbol{\beta}^l)\right) - y_j h(\mathbf{x}_j^{\mathbf{T}}\boldsymbol{\beta}^l) - b(y_j)\right\}\right\}\right] \\
& \times\left[\frac{1}{N}\sum_{l=1}^{N}\exp\left\{\sum_{j=1}^{n}\tilde{d}_{ij}\left\{a\left(h(\mathbf{x}_j^{\mathbf{T}}\boldsymbol{\beta}^l)\right) - y_j h(\mathbf{x}_j^{\mathbf{T}}\boldsymbol{\beta}^l) - b(y_j)\right\}\right.\right. \\
& \left.\left. + \mathbf{1}\left(\{\mathbf{y}_{(s_i)}\} = \emptyset\right)\log\left(\frac{\phi(\boldsymbol{\beta}^l;\boldsymbol{\mu}_\beta,\Sigma_\beta)}{\pi(\boldsymbol{\beta}^l)}\right)\right\}\right]^{-1},
\end{aligned}$$

where $d_{ij} = \mathbf{1}\left(y_j \notin \{\mathbf{y}_{s_i \cup (s_i)}\}\right)$, $\tilde{d}_{ij} = \mathbf{1}\left(y_j \notin \{\mathbf{y}_{(s_i)}\}\right)$,

$$\boldsymbol{\mu}_\beta = \frac{1}{N}\sum_{l=1}^{N} \boldsymbol{\beta}^l \quad \text{and} \quad \Sigma_\beta = \frac{1}{N-1}\sum_{l=1}^{N} (\boldsymbol{\beta}^l - \boldsymbol{\mu}_\beta)(\boldsymbol{\beta}^l - \boldsymbol{\mu}_\beta)^{\mathbf{T}}.$$

In practice, it might be hard to implement MCMC sampling for all the proposed models. In this situation, we propose to approximate the conditional predictive density p_i via the Laplace approximation as in Gelfand and Dey (1994). Assume $\{\mathbf{y}_{(s_i)}\} \neq \emptyset$. Then the conditional predictive density p_i can be approximated as

$$\begin{aligned} p_i &= \frac{\int_\Theta f(\mathbf{y}_{s_i \cup (s_i)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\left\{\int_\Theta f(\mathbf{y}_{(s_i)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\right\}^{\mathbf{1}\left(\{\mathbf{y}_{(s_i)}\}\neq\emptyset\right)}} \\ &\approx \frac{(2\pi)^{\frac{q}{2}} f(\mathbf{y}_{s_i \cup (s_i)}|\hat{\boldsymbol{\theta}}_1)\pi(\hat{\boldsymbol{\theta}}_1)\left|V_1(\hat{\boldsymbol{\theta}}_1)\right|^{\frac{1}{2}}}{\left\{(2\pi)^{\frac{q}{2}} f(\mathbf{y}_{(s_i)}|\hat{\boldsymbol{\theta}}_2)\pi(\hat{\boldsymbol{\theta}}_2)\left|V_2(\hat{\boldsymbol{\theta}}_2)\right|^{\frac{1}{2}}\right\}^{\mathbf{1}\left(\{\mathbf{y}_{(s_i)}\}\neq\emptyset\right)}} \quad (\equiv \tilde{p}_i)\,, \end{aligned} \tag{3.14}$$

where

$$\begin{aligned} \hat{\boldsymbol{\theta}}_1 &= \arg\max_{\boldsymbol{\theta}}\{f(\mathbf{y}_{s_i \cup (s_i)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\}, \\ \hat{\boldsymbol{\theta}}_2 &= \arg\max_{\boldsymbol{\theta}}\{f(\mathbf{y}_{(s_i)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\}, \\ V_1(\hat{\boldsymbol{\theta}}_1) &= \left[-\frac{\partial^2 \log\{f(\mathbf{y}_{s_i \cup (s_i)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathbf{T}}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_1}\right]^{-1}, \\ V_2(\hat{\boldsymbol{\theta}}_2) &= \left[-\frac{\partial^2 \log\{f(\mathbf{y}_{(s_i)}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathbf{T}}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_2}\right]^{-1}. \end{aligned}$$

To sum up, using (3.10) and (3.14), $BDC_\psi(\mathbf{p})$ in (3.5) can be calculated by $BDC_\psi(\hat{\mathbf{p}})$ or approximated by $BDC_\psi(\tilde{\mathbf{p}})$.

3.3 Calibration

An important ingredient of a consistent methodology involving any form of divergence measures is calibration. In order to validate the quality of predictions obtained from our framework, we develop a calibration method based on the *probability integral transform* (PIT) proposed by Dawid (1984). This aspect of model-building is crucial since a model chosen as the best with regards to a certain criterion may not be well-calibrated. As a consequence, a good model selection procedure is to choose the model that provides the best predictive performance, which attains the largest BDC, subject to calibration (Gneiting et al., 2007).

3.3.1 Probability integral transform

Let $F_i^{\text{true}}(y) = P(Y_i \leq y|y_{1:i-1}, M^{\text{true}})$ and $F_i^k(y) = P(Y_i \leq y|y_{1:i-1}, M^k)$ be cdfs of Y_i given $y_{1:i-1}$ under the true model M^{true} and under the k^{th} candidate M^k, respectively, for $i = 1, \ldots, n$. Without loss of generality, we suppose that outcomes, $y_1, y_2, \ldots, y_n$, are sequentially generated from the true model M^{true}. If $M^{\text{true}} = M^k$ and Y_i's are a continuous random variable, then $F_i^k(y_i)$, called PIT, follows independently the standard uniform distribution for $i = 1, \ldots, n$. This suggests that checking for the uniformity of PIT values is a reasonable way to perform a calibration check for continuous data. However, if data is not continuous, the traditional PIT technique fails since one does not have the uniformity. In this case, a *generalized probability integral transform* (GPIT) can be considered (Smith, 1985; Brockwell, 2007). Formally, let $(u_1, \ldots, u_n)$ be a random sample from $\text{U}(0, 1)$. Then for given model M^k the GPIT

is given as

$$G_i^k = (1 - u_i)F_i^k(y_i-) + u_i\{F_i^k(y_i)\}, \tag{3.15}$$

where $F(y-) = \lim_{t\uparrow y} F(t)$; indeed, this limit exists since the distribution function possesses left limits. Note that if $F_i^k(\cdot)$ is a continuous function, (3.15) reduces to the usual PIT. Similarly to PIT, GPIT follows independently the standard uniform distribution (Brockwell, 2007). As a consequence, in order for our framework to accommodate both discrete and continuous data, we propose to use GPIT for the purposes of calibration.

A problem persists however: in a Bayesian framework there is no guarantee for the existence of the GPITs in (3.15). To circumvent this issue, we use the idea of the minimal subset (Berger and Pericchi, 1996b) which would then ensure the existence of the GPITs. For given model M^k, let $\{y_1, \ldots, y_{m_k}\}$ be a minimal subset. Then, the corresponding set of GPIT's $\{G_{m_k+1}, \ldots, G_n\}$, can be used for calibration. For graphical inspection of uniformity, one could plot the histogram of the GPIT (Gneiting et al., 2007). However, checking uniformity only based on the histogram would hide the dependent nature of GPIT; note that the integral transform, in principle, guarantees independent uniform random variates. In order to check both uniformity and independence, we instead plot G_i versus i for $i = m_k + 1, \ldots, n$ for given model M^k; furthermore, we perform a Kolmogorov-Smirnov (KS) test verifying uniformity. In order to verify the independence condition, we consider the following transformation: for a given model M^k and $i = m_k + 1, \ldots, n$, consider

$$Z_i^k = \Phi^{-1}(G_i^k), \tag{3.16}$$

where $\Phi(\cdot)$ is a cdf of the standard normal distribution. If the GPIT follows independently a uniform distribution, then (3.16) follows independently the standard normal distribution. At this juncture, we can then employ formal tests for independence for data arising from a normal population. In this chapter, we choose the Ljung-Box (LB) (Ljung and Box, 1978) test to verify the independence condition.

3.3.2 Calculation of prequential distribution function

As in Section 3.2.3, for notational simplicity, we omit the index k and the model M^k in this section. Note that PIT or GPIT are determined by the prequential distribution function, $P(Y_i \leq y|y_{1:i-1})$, at the observed value. Therefore, we need to accurately calculate the prequential distribution function to obtain PIT or GPIT. However, calculating the prequential distribution function is not an easy task under the Bayesian framework. To address this issue, we propose a MC method to calculate the prequential distribution function using the MCMC sample from the full posterior distribution.

Theorem 3.11. *Let $\{\boldsymbol{\theta}^j\}_{j=1}^N$ be a set of MCMC samples from a full posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$ and $\{t^j\}_{j=1}^N$ be a random sample from $f(y_i|\mathbf{y}_{1:i-1}, \boldsymbol{\theta}^j)$ for given $j = 1, \ldots, N$. If the cdf $P(Y_i \leq y|\mathbf{y}_{1:i-1})$ exists for $y \in \mathbb{R}$, then*

$$\lim_{N\to\infty} \sum_{j=1}^{N} \frac{\mathbf{1}(t^j \leq y)}{N} \overset{a.s.}{=} P(Y_i \leq y|\mathbf{y}_{1:i-1}), \tag{3.17}$$

for $i = 2, \ldots, n$.

Theorem 3.11 informs us that the MC estimator in (3.17) converges to the prequential distribution function with probability 1. Based on the theorem, we propose

a MC estimator of PIT under model M^k in the following manner:

$$\hat{F}_i^k(y) = \sum_{j=1}^{N} \frac{\mathbf{1}(t^j \leq y)}{N}, \tag{3.18}$$

where $\{t^j\}_{j=1}^N$ is generated from pdf $f_{Y_i}(t|\mathbf{y}_{1:i-1}, \boldsymbol{\theta}_k^j, M^k)$ given the set of the full posterior samples $\left\{\boldsymbol{\theta}_k^j\right\}_{j=1}^N$. According to Theorem 3.11, our MC estimator in (3.18) converges to $F_i^k(y)$ with probability 1. Furthermore, using (3.18), the GPIT in (3.15) is easily obtained.

3.4 Illustrative examples

3.4.1 Linear regression model

In this section, we conduct some simulation studies to compare the model selection performance according to a choice of the convex function for a linear regression model, which is the most popular statistical models in practical application. To define the predictive density, we use the leave-one-out (LOO) cross validation approach (or CPO based approach) as the conditioning scheme, i.e., $p_i = p(y_i|\mathbf{y}_{-i})$ for $i = 1, \ldots, n$. We consider the following three types of BDC:

$$\begin{aligned} BDC_{\psi_1}(\mathbf{p}) &= \sum_{i=1}^{n} \log p_i, \\ BDC_{\psi_2}(\mathbf{p}) &= \frac{1}{n}\sum_{i=1}^{n} p_i, \\ BDC_{\psi_3}(\mathbf{p}) &= \sum_{i=1}^{n} p_i^2, \end{aligned}$$

which are induced by $\psi_1(\mathbf{p}) = -\sum_{i=1}^{n} \log p_i - n$, $\psi_2(\mathbf{p}) = \sum_{i=1}^{n} (p_i/n) \log p_i$, and $\psi_3(\mathbf{p}) = \sum_{i=1}^{n} p_i^2$, respectively. Consider three models M^1, M^2, and M^3 such that

$$M^k : \ \mathbf{y} = \mathbf{X}\boldsymbol{\beta}_k + \boldsymbol{\epsilon}, \quad k = 1, 2, 3$$

where $\boldsymbol{\beta}_1 = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathbf{T}}$, $\boldsymbol{\beta}_2 = (\beta_0, \beta_1, \beta_2, 0)^{\mathbf{T}}$, $\boldsymbol{\beta}_3 = (\beta_0, \beta_1, 0, 0)^{\mathbf{T}}$, $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}_n)$ (σ_k^2 is unknown), $\mathbf{y} = (y_1, \ldots, y_n)^{\mathbf{T}}$, and $\mathbf{X} = [\mathbf{1}_n, X_1, X_2, X_3]$ is an $(n \times 4)$ design matrix with full column rank. Define $\mathbf{X}_1 = [\mathbf{1}_n, X_1, X_2, X_3]$, $\mathbf{X}_2 = [\mathbf{1}_n, X_1, X_2]$, and $\mathbf{X}_3 = [\mathbf{1}_n, X_1]$. Consider the non-informative prior $\pi_k(\boldsymbol{\beta}_k, \sigma_k) = 1/\sigma_k^2$ for $k = 1, 2, 3$, then for given k, the posterior distribution is

$$\pi(\boldsymbol{\beta}_k, \sigma_k^2 | \mathbf{y}) \propto (\sigma_k^2)^{-\left(\frac{n}{2}+1\right)} \exp\left\{ -\frac{1}{2\sigma_k^2} (\mathbf{y} - \mathbf{X}_k\boldsymbol{\beta}_k)^{\mathbf{T}} (\mathbf{y} - \mathbf{X}_k\boldsymbol{\beta}_k) \right\}. \tag{3.19}$$

From (3.19), we can easily determine that

$$\sigma_k^2 | \mathbf{y} \quad \sim \mathrm{IG}\left(\frac{n - p_k}{2}, \frac{(\mathbf{y} - \hat{\mathbf{y}}_k)^{\mathbf{T}} (\mathbf{y} - \hat{\mathbf{y}}_k)}{2} \right), \tag{3.20}$$

$$\boldsymbol{\beta}_k, | \sigma_k^2, \mathbf{y} \ \sim \mathrm{N}\left(\hat{\boldsymbol{\beta}}_k, \sigma_k^2 (\mathbf{X}_k^{\mathbf{T}} \mathbf{X}_k)^{-1} \right), \tag{3.21}$$

where $\hat{\mathbf{y}}_k = \mathbf{X}_k (\mathbf{X}_k^{\mathbf{T}} \mathbf{X}_k)^{-1} \mathbf{X}_k^{\mathbf{T}} \mathbf{y}$, $\hat{\boldsymbol{\beta}}_k = (\mathbf{X}_k^{\mathbf{T}} \mathbf{X}_k)^{-1} \mathbf{X}_k^{\mathbf{T}} \mathbf{y}$, and q_k is the dimension of $\boldsymbol{\beta}_k$ for $k = 1, 2, 3$. Therefore, the full posterior sample, $\left\{ (\boldsymbol{\beta}_k^l, \sigma_k^{2l}) \right\}_{l=1}^{N}$, can be directly generated from (3.20) and (3.21).

We simulate 1,000 data sets under the following setting: for given $n = 50$, we set $\boldsymbol{\beta} = (2, -2, 2.5, 0)^{\mathbf{T}}$, $\mathbf{X} = [\mathbf{1}_n, X_1, X_2, X_3]$ with $x_{ij} \overset{iid}{\sim} \mathrm{N}(0, 1)$, where x_{ij} indicates i^{th} element of a vector X_j, and $\sigma^2 = 1$. In this setting, the true model is M^2. The model

TABLE 3.2: Summary of selection rates.

Model	BDC_{ψ_1}	BDC_{ψ_2}	BDC_{ψ_3}	DIC
M^1	0.172	0.198	0.194	0.184
M^2	0.828	0.802	0.806	0.816
M^3	0.000	0.000	0.000	0.000

TABLE 3.3: Mean and MSE of BDCs

	method	M^1	M^2	M^3
BDC_{ψ_1}	exact	-73.6954	-73.1707	-122.0472
	estimate	-73.6961	-73.1656	-122.0471
	MSE	0.0992	0.0801	0.0712
BDC_{ψ_2}	exact	0.2690	0.2715	0.1021
	estimate	0.2690	0.2715	0.1021
	MSE	0.0004	0.0004	0.0001
BDC_{ψ_3}	exact	4.2371	4.3130	0.6086
	estimate	4.2369	4.3137	0.6087
	MSE	0.0159	0.0160	0.0020

M^1 and M^3 are over-fitting and under-fitting, respectively. The size of full posterior samples is $N = 2,000$ in each simulation.

Table 3.2 summarizes the model selection performance of the BDCs along with Deviance Information Criterion (DIC), which is a standard Bayesian model selection criterion. As a result, BDC_{ψ_1} provides the highest selection rate for the true model ($= M^2$), and this shows that BDC_{ψ_1} (or logarithm scoring) performs better than the others in linear regression models.

To check the accuracy of our MC estimator for BDC, using Example 3.8, we compare the mean of estimated BDCs with the mean of the exact BDCs, see Table 3.3. All estimated mean values are close to true mean values and MSEs are small enough. Therefore, we assure that the proposed MC estimator is relatively accurate for linear regression models.

3.4.2 Bayesian longitudinal data model

Many biological experiments are conducted under longitudinal study setup, where the measurements on subjects are repeatedly measured over time. Due to the experimental scheme, a correlation between measurements on the given subject naturally occurs, so the determination of the correlation structure requires a special care in the statistical modelling for longitudinal data models. In this section, we assume that every individual has the same number of observations, but unequal size of observations can be occurred due to missing data in practice. Suppose that y_{ij} denotes the j^{th} measurement on the i^{th} subject and $\mathbf{X}_i$ is an $(a \times q)$ design matrix of covariates for i^{th} subject in the longitudinal study for $i = 1, \ldots, n$ and $j = 1, \ldots, a$. Then the following linear model can be used to fit the longitudinal data:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \tag{3.22}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{ia})^{\mathbf{T}}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{q-1})^{\mathbf{T}}$ is a $q-$dimensional parameter vector, and $\boldsymbol{\epsilon}_i$ is the random error vector with $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$ and $Cov(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j) = \mathbf{V}_i$ if $i = j$ and $Cov(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j) = \mathbf{0}$ if $i \neq j$ for $i, j = 1, \ldots, n$. In general the normality is assumed for the error vector $\boldsymbol{\epsilon}_i$ with the homogeneity, i.e., $\mathbf{V}_i = \mathbf{V}$ for all i, and then we can write (3.22) as

$$\mathbf{y} \sim \mathrm{N}(\mathbf{Z}\boldsymbol{\beta}, \mathbf{I}_n \otimes \mathbf{V}), \tag{3.23}$$

where $\mathbf{y} = (\mathbf{y}_1^{\mathbf{T}}, \ldots, \mathbf{y}_n^{\mathbf{T}})^{\mathbf{T}}$, $\mathbf{Z} = \left[\mathbf{X}_1^{\mathbf{T}}, \ldots, \mathbf{X}_n^{\mathbf{T}}\right]^{\mathbf{T}}$, $\mathbf{I}_n$ is the $(n \times n)$ identity matrix, and $\mathbf{I}_n \otimes \mathbf{V}$ denotes the Kronecker product of $\mathbf{I}_n$ and $\mathbf{V}$. In model (3.23), the correlation between measurements is determined by the covariance matrix $\mathbf{V}$.

In this section, we exemplify the determination of the covariance matrix using our method with rat population growth data (Gelfand et al., 1990). The data consist of the weights of 30 young rats (n=30) that were measured weekly for five weeks (a=5). Define that Y_{ij} is the weight of the i^{th} rat at time point j and the j^{th} row of $\mathbf{X}_i$ is $\mathbf{x}_{ij}^{\mathbf{T}} = (1, x_{ij})$, where x_{ij} is the age in days at the point j for $i = 1, \ldots, 30$ and $j = 1, \ldots, 5$. In many cases, a uniform correlation between measurements (called compound symmetry) on the individual subject is assumed on the covariance matrix $\mathbf{V}$ such that

$$\mathbf{V} = \sigma^2 \mathbf{V}^1(\rho) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix},$$

where $0 < \rho < 1$. If the correlation between any two measurements on the same subject dramatically decreases towards zero as the time distance between the measurements increases, then the exponential correlation function can be considered as follows:

$$\rho(l, k) = \exp\left(-b|t_l - t_k|\right), \tag{3.24}$$

where b is an unknown constant and t_l and t_k indicate the time points of l^{th} and k^{th} measurements, respectively. Since the weights were measured once every week, i.e.,

$|t_l - t_k| \propto |l - t|$, then (3.24) can be re-written as

$$\rho(l,k) = \rho^{|l-k|}, \tag{3.25}$$

where $\rho = \exp(-b)$ and a common constant in $|t_l - t_k|$ is absorbed into the unknown constant b for $l, k = 1, \ldots, 5$. According to (3.25) the covariance matrix $\mathbf{V}$ is given as

$$\mathbf{V} = \sigma^2 \mathbf{V}^2(\rho) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix},$$

where $0 < \rho < 1$. Now we compare the following four models: 1) model M^1 (uniform correlation) with $\mathbf{V} = \sigma^2 \mathbf{V}^1(\rho)$, 2) model M^2 (exponential correlation) with $\mathbf{V} = \sigma^2 \mathbf{V}^2(\rho)$, 3) model M^3 (no correlation) with $\mathbf{V} = \sigma^2 \mathbf{I}_5$, and 4) model M^4 (unspecified dependence structure) with $\mathbf{V} = \mathbf{\Sigma}$. Let us consider the following priors:

$$\pi(\boldsymbol{\beta}, \sigma^2, \rho) \propto \sigma^{-2} \qquad \text{for model } M^1,\ M^2,\ M^3, \tag{3.26}$$

$$(\boldsymbol{\beta}, \mathbf{\Sigma}) \sim \text{IW}\,(v_0, \mathbf{\Sigma}_0) \qquad \text{for model } M^4, \tag{3.27}$$

where $v_0 = 5$ and $\mathbf{\Sigma}_0 = 0 \cdot 001 \mathbf{I}_5$. After some calculations with (3.23), (3.26), and (3.27), we obtain the following full conditionals for each model:

For model M^1 and M^2,

$$\begin{aligned}
\boldsymbol{\beta}|\sigma^2,\rho,\mathbf{y},M^k &\sim \mathrm{N}\left(\tilde{\boldsymbol{\beta}}^k,\sigma^2\left[\mathbf{Z}^{\mathbf{T}}\tilde{\mathbf{V}}^k(\rho)^{-1}\mathbf{Z}\right]^{-1}\right),\\
\sigma^2|\rho,\boldsymbol{\beta},\mathbf{y},M^k &\sim \mathrm{IG}\left(\frac{na}{2},\frac{(\mathbf{y}-\mathbf{Z}\boldsymbol{\beta})^{\mathbf{T}}\tilde{\mathbf{V}}^k(\rho)^{-1}(\mathbf{y}-\mathbf{Z}\boldsymbol{\beta})}{2}\right),\\
\pi(\rho|\boldsymbol{\beta},\sigma^2,\mathbf{y},M^k) &\propto \left|\mathbf{V}^k(\rho)\right|^{-\frac{n}{2}}\exp\left\{-\frac{(\mathbf{y}-\mathbf{Z}\boldsymbol{\beta})^{\mathbf{T}}\tilde{\mathbf{V}}^k(\rho)^{-1}(\mathbf{y}-\mathbf{Z}\boldsymbol{\beta})}{2\sigma^2}\right\},
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}^k=\left(\mathbf{Z}^{\mathbf{T}}\tilde{\mathbf{V}}^k(\rho)^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathbf{T}}\tilde{\mathbf{V}}^k(\rho)^{-1}\mathbf{y}$ and $\tilde{\mathbf{V}}^k(\rho)^{-1}=\mathbf{I}_n\otimes\mathbf{V}^k(\rho)^{-1}$ for $k=1,2$. For model M^3,

$$\begin{aligned}
\boldsymbol{\beta}|\sigma^2,\mathbf{y},M^3 &\sim \mathrm{N}\left(\hat{\boldsymbol{\beta}},\sigma^2\left(\mathbf{Z}^{\mathbf{T}}\mathbf{Z}\right)^{-1}\right),\\
\sigma^2|\boldsymbol{\beta},\mathbf{y},M^3 &\sim \mathrm{IG}\left(\frac{na}{2},\frac{(\mathbf{y}-\mathbf{Z}\boldsymbol{\beta})^{\mathbf{T}}(\mathbf{y}-\mathbf{Z}\boldsymbol{\beta})}{2}\right),
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}=\left(\mathbf{Z}^{\mathbf{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathbf{T}}\mathbf{y}$. For model M^4,

$$\begin{aligned}
\boldsymbol{\beta}|\boldsymbol{\Sigma},\mathbf{y},M^4 &\sim \mathrm{N}\left(\tilde{\boldsymbol{\beta}}^4,\left[\mathbf{Z}^{\mathbf{T}}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}\right]^{-1}\right),\\
\boldsymbol{\Sigma}|\boldsymbol{\beta},\mathbf{y},M^4 &\sim \mathrm{IW}\left(v_0+n,\boldsymbol{\Sigma}_0+\sum_{i=1}^{n}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^{\mathbf{T}}\right),
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}^4=\left(\mathbf{Z}^{\mathbf{T}}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathbf{T}}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$, $\tilde{\boldsymbol{\Sigma}}^{-1}=\mathbf{I}_n\otimes\boldsymbol{\Sigma}^{-1}$. Using the full conditionals, we generate 10,000 samples from the full posterior (after 5,000 burn-in iterations) using Gibbs sampler in each model. In model M^1 and M^2, we use Metropolis-Hastings algorithm within the Gibbs chain in order to generate samples from ρ. For the proposal distribution, we approximate the conditional distribution of ρ given posterior

mean of $(\boldsymbol{\beta}, \sigma^2)$ (obtained from M^3) using Gaussian approximation and then truncate the distribution with lower tail 0 and upper tail 1. Since the data have the dependency, we use the prequential predictive densities as follows:

$$p_j^k = \begin{cases} m^k(\mathbf{y}_{\cdot 1}) & \text{if } j = 1 \\ p(\mathbf{y}_{\cdot j}|\mathbf{y}_{\cdot 1:j-1}, M^k) & \text{if } j = 2, \ldots, 5 \end{cases}, k = 1, \ldots, 4, \tag{3.28}$$

where $\mathbf{y}_{\cdot j} = \{y_{1j}, \ldots, y_{30j}\}$ for $j = 1, \ldots, 5$. The MC estimator of (3.28) is given as

$$\hat{p}_1^k = \left\{ \frac{1}{N} \sum_{l=1}^{N} \frac{f(\mathbf{y}_{\cdot 1}|\boldsymbol{\theta}_k^l, M^k)}{f(\mathbf{y}|\boldsymbol{\theta}_k^l, M^k)} \right\} \left\{ \frac{1}{N} \sum_{l=1}^{N} \frac{g^k(\boldsymbol{\theta}_k^l)/\pi(\boldsymbol{\theta}_k^l)}{f(\mathbf{y}|\boldsymbol{\theta}_k^l, M^k)} \right\}^{-1}, \tag{3.29}$$

and for $j = 2, \ldots, 5$

$$\hat{p}_j^k = \left\{ \frac{1}{N} \sum_{l=1}^{N} \frac{f(\mathbf{y}_{\cdot 1:j}|\boldsymbol{\theta}_k^l, M^k)}{f(\mathbf{y}|\boldsymbol{\theta}_k^l, M^k)} \right\} \left\{ \frac{1}{N} \sum_{l=1}^{N} \frac{f(\mathbf{y}_{\cdot 1:j-1}|\boldsymbol{\theta}_k^l, M^k)}{f(\mathbf{y}|\boldsymbol{\theta}_k^l, M^k)} \right\}^{-1}, \tag{3.30}$$

where $\left\{\boldsymbol{\theta}_1^l = (\boldsymbol{\beta}_1^l, \sigma_1^{2l}, \rho_1^l)\right\}_{l=1}^N$, $\left\{\boldsymbol{\theta}_2^l = (\boldsymbol{\beta}_2^l, \sigma_2^{2l}, \rho_2^l)\right\}_{l=1}^N$, $\left\{\boldsymbol{\theta}_3^l = (\boldsymbol{\beta}_3^l, \sigma_3^{2l})\right\}_{l=1}^N$,

and $\left\{\boldsymbol{\theta}_4^l = (\boldsymbol{\beta}_4^l, \boldsymbol{\Sigma}_4^l)\right\}_{l=1}^N$ are respectively sets of MCMC samples of full posterior distributions under model M^1, M^2, M^3, and M^4. Note that the likelihood functions in (3.29) and (3.30) can be easily obtained from (3.23). For the choice of the convex function, we utilize ψ_1, which provide the best model selection performance in the previous section. For model M^1, M^2, M^3, and M^4, the obtained BDCs are -643.2493, -551.3066, -514.8821, and -514.8509, respectively. This result shows that the unspecified covariance structure model (M^4) provides the best performance for the rat population growth data. The plots of PITs are shown in Figure 3.1 along with p-values of formal tests for calibration check. In Figure 3.1, model M^4 provides

randomly scattered PITs and satisfies the uniformity and the independence, while other models present some patterns and fail the independence test. Hence, model M^4 is the best model for the given data.

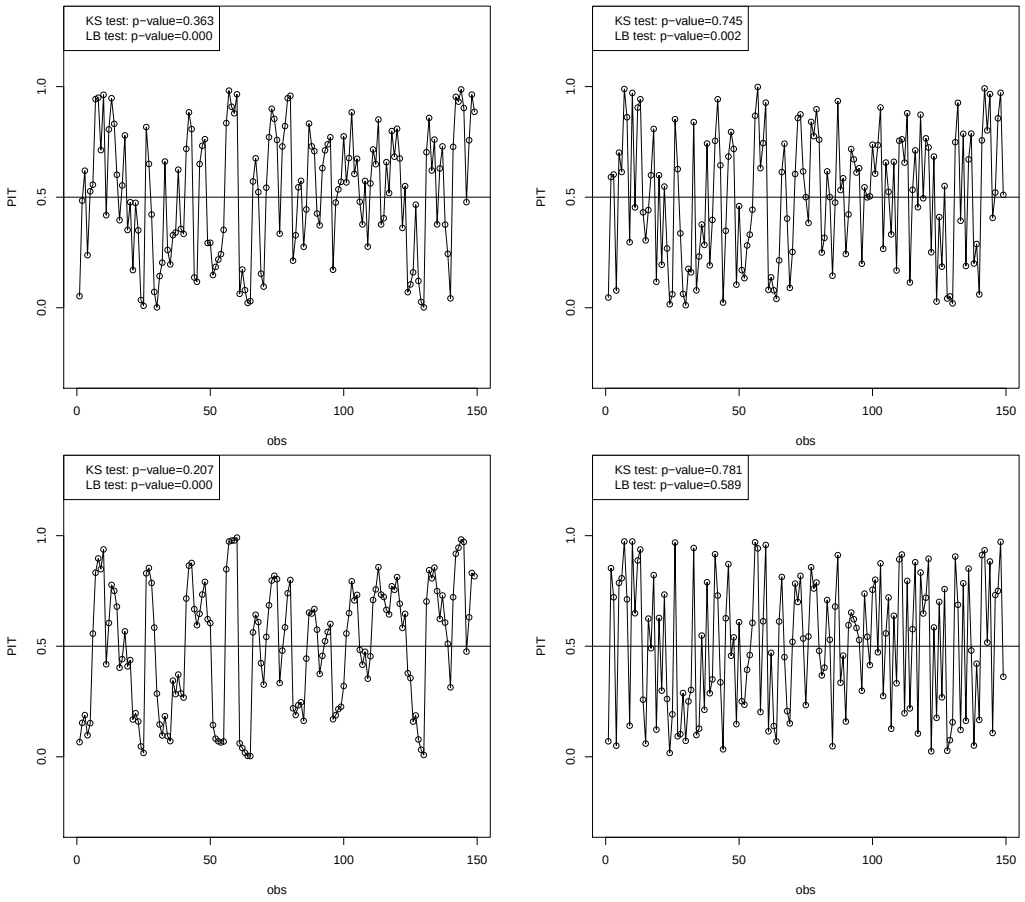

FIGURE 3.1: Plots of PITs under model M^1, M^2, M^3, and M^4 with p-values of KS test and LB test.

3.5 Concluding remarks

We have introduced the generalization of Bayesian predictive model selection measure, named BDC, along with the calibration method based on GPIT. In order to

calculate BDC and GPIT, we propose MC estimators based on MCMC samples from the full posterior distribution. Using our estimators, various predictive densities and prequential distribution functions can be obtained in the Bayesian framework. In addition, our MC estimators can be directly applied to calculations of many kinds of Bayes factors such as BF, PSBF, and IBF.

In this chapter, our computation of the BDC relies on an importance sampling approach, but this may fail if the proposal distribution does not sufficiently well approximate the target. To improve the accuracy, using *Truncated Importance Sampling* (Ionides, 2008), (3.10) can be replaced by

$$\hat{p}_i = \left[\frac{1}{N}\sum_{j=1}^{N}\min\{W_1(\boldsymbol{\theta}^j), \sqrt{N}\bar{W}_1\}\right]\left[\frac{1}{N}\sum_{j=1}^{N}\min\{W_2(\boldsymbol{\theta}^j), \sqrt{N}\bar{W}_2\}\right]^{-1},$$

where

$$W_1(\boldsymbol{\theta}^j) = \frac{f(\mathbf{y}_{s_i\cup(s_i)}|\boldsymbol{\theta}^j)}{f(\mathbf{y}|\boldsymbol{\theta}^j)}; \quad W_2(\boldsymbol{\theta}^j) = \frac{\{g(\boldsymbol{\theta}^j)/\pi(\boldsymbol{\theta}^j)\}^{\mathbf{1}(\{\mathbf{y}_{(s_i)}\}=\emptyset)}}{f(\mathbf{y}_{-(s_i)}|\mathbf{y}_{(s_i)}, \boldsymbol{\theta}^j)};$$

$$\bar{W}_1 = \sum_{j=1}^{N} W_1(\boldsymbol{\theta}^j)/N; \quad \bar{W}_2 = \sum_{j=1}^{N} W_2(\boldsymbol{\theta}^j)/N.$$

In practice, penalized likelihood criteria such as Akaike information criterion (AIC), Bayesian information criterion (BIC), and Deviance information criterion (DIC) are commonly utilized in model selection problems. Even though the penalized likelihood approach is built from a different perspective with the predictive distribution approach, many studies revealed that these approaches are asymptotically equivalent when the sample size is large enough, see Gelfand and Dey (1994); Stone (1974); Kass and Vaidyanathan (1992) for more details. Therefore, BDC generalizes asymptotically the penalized likelihood methods as well.

Chapter 4

Bayesian Modeling of Sparse High-Dimensional Data using Bregman Divergence

4.1 Introduction

Suppose we have n independent observations of the response value y_i and the predictor vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$. To reveal the relationship between the response and the predictors, the following parametric model has been widely used: for $i = 1, \ldots, n$,

$$E(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \mathbf{h}(\mathbf{X}\boldsymbol{\beta}), \tag{4.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{h}(\mathbf{X}\boldsymbol{\beta}) = \big(h(\mathbf{x}_1^{\mathbf{T}}\boldsymbol{\beta}), \ldots, h(\mathbf{x}_n^{\mathbf{T}}\boldsymbol{\beta})\big)^{\mathbf{T}}$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathbf{T}}$, and $h(\cdot)$ is a known non-deceasing link function. Here, the unknown coefficient vector $\boldsymbol{\beta}$ is of our primary interest.

In many practical situations, especially in genetic research, we encounter frequently that the number of considered predictors (=p) is possibly of high-dimensionality, but only a few predictors may be significantly related to the response, known as "*sparse high-dimensional problems*". For instance, a genome-wide association study looks at millions of single nucleotide polymorphisms (SNPs) to identify several relevant genes to a certain phenotype.

A very convenient way of eliminating the irrelevant predictors from the model in (4.1) is to define the corresponding coefficient values as *zero*; this is where the "*sparse*" comes from. Motivated by this idea, to achieve the estimation and the variable elimination simultaneously, many studies (Tibshirani, 1996; Zou, 2006; Fan and Li, 2001) have proposed the Penalized Loss function (PL) estimator,

$$\hat{\boldsymbol{\beta}}_{\mathrm{PL}} = \arg\min_{\boldsymbol{\beta}} \left[L\{\mathbf{y}, \mathbf{h}(\mathbf{X}\boldsymbol{\beta})\} + \mathrm{Pe}(\boldsymbol{\beta}|\lambda) \right], \tag{4.2}$$

where $L(\cdot, \cdot)$ denotes a loss function and $\mathrm{Pe}(\cdot|\lambda)$ is a penalty function with a regularization (or tuning) parameter λ controlling the degree of penalization.

There are numerous choices of the sparsity-inducing penalty function in (4.2). For example, Akaike (1974) and Schwarz (1978) utilized the ℓ_0-norm penalty,

$$\ell_0(\boldsymbol{\beta}|\lambda) = \lambda \sum_{j=1}^{p} \mathbf{1}\{\beta_j \neq 0\}, \tag{4.3}$$

where $\lambda \geq 0$ and $\mathbf{1}\{\cdot\}$ denotes an indicator function. Owing to the indicator function in (4.3), we can directly restrict the number of non-zero coefficients, and successfully induce the sparsity. However, the indicator function leads a severe burden on the computation of (4.2) due to its discontinuity. To overcome the computational

difficulty, Tibshirani (1996) proposed to utilize the ℓ_1-norm penalty, called LASSO,

$$\text{LASSO}(\boldsymbol{\beta}|\lambda) = \lambda \sum_{j=1}^{p} |\beta_j| \,, \tag{4.4}$$

where $\lambda \geq 0$. Since LASSO is a continuous and convex function for $\boldsymbol{\beta}$, it defeats the computational challenge. However, Zou (2006) pointed out that the use of the common tuning parameter λ in (4.4) could produce undesirable bias estimators, because it imposes the same amount of penalty on both relevant and irrelevant predictors. As an alternative, Zou (2006) proposed a weighted ℓ_1-norm penalty, called the adaptive LASSO (a-LASSO),

$$\text{a-LASSO}(\boldsymbol{\beta}|\lambda, \mathbf{w}) = \lambda \sum_{j=1}^{p} w_j \, |\beta_j| \,,$$

where $\lambda \geq 0$ and $\mathbf{w} = (w_1, \ldots, w_p)$ is the data–driven weight such that $w_j = 1/|\hat{\beta}_j|^{\gamma}$ and $\hat{\beta}_j$ is a $\sqrt{n}$–consistent estimator for β_j for $j = 1, \ldots, p$ and $\gamma > 0$. A well-defined weight releases the possible non-zero coefficients from the penalization, so this remedies the bias problem. However, in the high dimensional setup, it is somewhat unrealistic to find a good weight, due to the fact that we should not know which predictors will be irrelevant. Recently, to avoid the determination of the weight function, Dicker et al. (2013) introduced a continuous approximation to the ℓ_0-norm penalty, called SELO,

$$\text{SELO}(\boldsymbol{\beta}|\lambda, \tau) = \lambda \sum_{j=1}^{p} \frac{1}{\log(2)} \log\left(\frac{|\beta_j|}{|\beta_j| + \tau} + 1 \right),$$

where $\lambda \geq 0$ and $\tau > 0$. Interestingly, SELO coverges to ℓ_0-norm penalty as τ

approaches 0. In other words, $\ell_0(\boldsymbol{\beta}|\lambda) \approx \text{SELO}(\boldsymbol{\beta}|\lambda, \tau)$ for a sufficiently small τ; e.g., $\tau = 0.01$ (Dicker et al., 2013). Consequently, SELO enables us to overcome the computational drawback of ℓ_0-norm penalty as well as the theoretical and practical limitations of ℓ_1-norm penalties.

As a choice of the loss function in (4.2), a general class of divergence measures, called *Bregman divergence*, has been in the spotlight (Banerjee et al., 2005; Zhang et al., 2010). Since many well-known loss functions, such as squared error loss, Kullback-Leibler (KL) divergence, Itakura-Saito distance (Itakura and Saito, 1970), and Mahalanobis distance, belong to Bregman divergence, the use of Bregman divergence generalizes and unifies many existing loss functions.

While the PL technique produces an attractive point estimator of $\boldsymbol{\beta}$ in sparse high-dimensional problems, it is a challenge to explain the uncertainty of the obtained estimator (Kyung et al., 2010). However, from a Bayesian perspective, the unknown parameter $\boldsymbol{\beta}$ is considered as a random variable and thus the uncertainty of its estimator can be explained by the induced probability distribution of $\boldsymbol{\beta}$, called posterior distribution, for given data $\mathcal{D} = \{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$. This aspect motivates us to utilize a Bayesian approach to the sparse high-dimensional problems. In fact, the Bayesian methodology has been used by many researchers. Park and Casella (2008), for example, introduced the Bayesian lasso using the Laplace prior, which corresponds to the LASSO penalty. Kyung et al. (2010) discussed general relationship between the Bayesian approach and the ℓ_1-norm penalized loss-function methods. Recently, using the exponential power priors, Polson et al. (2014) proposed the Bayesian bridge, which is a counterpart of ℓ_α-norm penalized loss-function method for $\alpha \in (0, 1]$.

To the best of our knowledge, neither the idea of approximating ℓ_0-norm penalty

nor the Bregman divergence loss function has been considered in sparse high-dimensional problems from a Bayesian perspective until now. In this chapter, our aim is to develop a new Bayesian approach to the high-dimensional challenge using both Bregman divergence and ℓ_0-norm approximation. Our approximation, however, differs from SELO in Dicker et al. (2013). Unlike SELO, our approximation to ℓ_0 norm leads to a continuous and differentiable penalty function, which is more beneficial to computation. Furthermore, in a hierarchical Bayesian framework, our approximation can be represented by Gaussian and gamma mixture model, so the estimation procedure can be easily conducted by standard Bayesian estimation algorithms such as Markov Chain Monte Carlo (MCMC) and Iterated Conditional Modes (ICM) algorithm (Besag, 1986).

The outline of the remainder of the chapter is as follows. In Section 4.2, using Bregman divergence and Gaussian and Diffused-gamma prior, we introduce a new Bayesian modeling for high-dimensional regression problems. In Section 4.3, we introduce a coordinate-wise ICM algorithm to find the posterior mode in the presence of high-dimensionality. In addition, a practical MCMC method is discussed. In Section 4.4, illustrative examples are described through simulation and real data studies. Section 4.5 offers several concluding remarks.

4.2 Bayesian modeling

From a Bayesian viewpoint, the PL estimator in (4.2) can be viewed as the *Maximum A Posteriori* (MAP) estimator of the following posterior distribution (Tibshirani,

1996):

$$\pi(\boldsymbol{\beta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}) \propto \exp\left[-L\{\mathbf{y}, \mathbf{g}(\mathbf{X}\boldsymbol{\beta})\}\right] \exp\left\{-\mathrm{Pe}(\boldsymbol{\beta}|\lambda)\right\},$$

where $f(\mathbf{y}|\boldsymbol{\beta})$ and $\pi(\boldsymbol{\beta})$ are respectively the likelihood and the prior density function. Let $\hat{\boldsymbol{\beta}}_{\mathrm{MAP}}$ be the MAP estimator, i.e., $\hat{\boldsymbol{\beta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\beta}}\{\pi(\boldsymbol{\beta}|\mathbf{y})\}$. Then, the uncertainty of $\hat{\boldsymbol{\beta}}_{\mathrm{MAP}}$ can be easily explained by the posterior $\pi(\boldsymbol{\beta}|\mathbf{y})$. From the aforementioned perspective, in this section, we will discuss the development of the likelihood, the prior, and the posterior.

4.2.1 Likelihood specification using Bregman divergence

The relationship between the loss function and the likelihood, also known as the *duality property*, was first discussed by (Bernardo and Smith, 1994). The duality property states that the loss function can be viewed as the negative of the log likelihood function. For example, if we define $L(\mathbf{z}_1, \mathbf{z}_2) = \frac{1}{2}||\mathbf{z}_1 - \mathbf{z}_2||^2$, then the corresponding likelihood function is given by $f(\mathbf{y}|\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}||\mathbf{y} - \mathbf{h}(\mathbf{X}\boldsymbol{\beta})||^2\right\}$, i.e., the Gaussian density function.

In order to develop a general class of likelihood functions rather than one specific likelihood, we propose to use Bregman divergence and then define the likelihood function as

$$f_\psi(\mathbf{y}|\boldsymbol{\beta}) \propto \exp\left[-\mathrm{BD}_\psi\left\{\mathbf{y}, \mathbf{h}\left(\mathbf{X}\boldsymbol{\beta}\right)\right\}\right]. \tag{4.5}$$

One may wonder, "what is the corresponding distribution family to the Bregman divergence?" To answer this question, Banerjee et al. (2005) showed that any member

of the natural exponential family corresponds to a unique and distinct member of Bregman divergence. This implies that the developed class of likelihood functions by Bregman divergence contains the natural exponential family distribution as a subset. For example, if we define $\psi(\mathbf{x}) = \sum_{i=1}^{n} \{x_i \log x_i\}$, then (4.5) reduces to the Poisson likelihood,

$$\begin{aligned} f_\psi(\mathbf{y}|\boldsymbol{\beta}) &\propto \exp\left[-\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{h(\mathbf{x}_i\boldsymbol{\beta})}\right) - (y_i - h(\mathbf{x}_i\boldsymbol{\beta}))\right\}\right] \\ &\propto \prod_{i=1}^{n}\left[e^{-h(\mathbf{x}_i\boldsymbol{\beta})}\left\{h(\mathbf{x}_i\boldsymbol{\beta})\right\}^{y_i}\right]. \end{aligned}$$

See Table 4.1 for more examples. It is worth noting that our likelihood is more general than the natural exponential family. For instance, Zhang et al. (2009) verified that the quasilikelihood function (Wedderburn, 1974) belongs to Bregman divergence. Consequently, owing to the generality of Bregman divergence, the proposed method allows us to handle various types of data such as count, binary, continuous, etc.

TABLE 4.1: Examples of the Bregman divergences generated by some convex functions and related distributions in the natural exponential family.

$\psi(z)$	$\mathrm{BD}_\psi(z_1, z_2)$	Distribution
$\frac{1}{2\sigma^2}z^2$	$\frac{1}{2\sigma^2}(z_1 - z_2)^2$	Gaussian
$z \log z$	$z_1 \log\left(\frac{z_1}{z_2}\right) - (z_1 - z_2)$	Poisson
$-\log z$	$\frac{z_1}{z_2} - \log\left(\frac{z_1}{z_2}\right) - 1$	Exponential
$z \log z + (1-z)\log(1-z)$	$z_1 \log\left(\frac{z_1}{z_2}\right) + (1-z_1)\log\left(\frac{1-z_1}{1-z_2}\right)$	Bernoulli

4.2.2 Prior specification using ℓ_0-norm approximation

Recall that our goal is to develop a prior that mimics very closely the ℓ_0-norm penalty function in (4.3). Let $\tilde{\ell}_0(\cdot|\cdot)$ be a good approximation of the ℓ_0-norm penalty. To specify $\tilde{\ell}_0(\cdot|\cdot)$, we introduce a new penalty function,

$$\tilde{\ell}_0(\boldsymbol{\beta}|\lambda,\tau) = \lambda \sum_{j=1}^{p} \frac{\{(\mathbf{X}_{[,j]})^{\mathbf{T}}\mathbf{X}_{[,j]}\}\beta_j^2}{\tau^2 + \{(\mathbf{X}_{[,j]})^{\mathbf{T}}\mathbf{X}_{[,j]}\}\beta_j^2}, \tag{4.6}$$

where τ is a deterministic constant and $\mathbf{X}_{[,j]}$ denotes the j^{th} column of the design matrix $\mathbf{X}$. To argue a nice property of our penalty function, we introduce the following lemma.

Lemma 4.1. *Define* $f(\tau|x) = x^2/(\tau^2 + x^2)$. *Then,*

$$\lim_{\tau\to 0} f(\tau|x=0) \to 0 \quad \textit{and} \quad \lim_{\tau\to 0} f(\tau|x\neq 0) \to 1.$$

The proof is straightforward, and we thus omit it. According to Lemma 4.1, we can easily show that $\tilde{\ell}_0(\boldsymbol{\beta}|\lambda,\tau) \to \lambda \sum_{j=1}^{p} \mathbf{1}\{\beta_j \neq 0\}$ as τ goes to zero for given λ and $\boldsymbol{\beta}$. Hence, if τ is chosen to be sufficiently small, then our penalty function well approximates the ℓ_0-norm penalty, i.e., $\tilde{\ell}_0(\boldsymbol{\beta}|\lambda,\tau \approx 0) \approx \ell_0(\boldsymbol{\beta}|\lambda)$. To illustrate, in Figure 4.1(left), we display graphs of $f(x) = x^2/(\tau^2 + x^2)$ for varying $\tau = 10^{-k}$, $k = 2,3,4,5$. As τ gets closer to zero, the function indeed approaches $\mathbf{1}\{x \neq 0\}$. Figure 4.1(right) compares our function, say GD, with SELO($\boldsymbol{\beta} = x|\lambda = 1,\tau) = \frac{1}{\log(2)} \log\left(\frac{|x|}{|x|+\tau} + 1\right)$ for a given $\tau = 10^{-4}$. Unlike the SELO, our GD is smooth (or differentiable) at the peak ($x = 0$), and this feature releases us from a lot of computational burden. Now, we define our new prior, called Gaussian and Diffused-

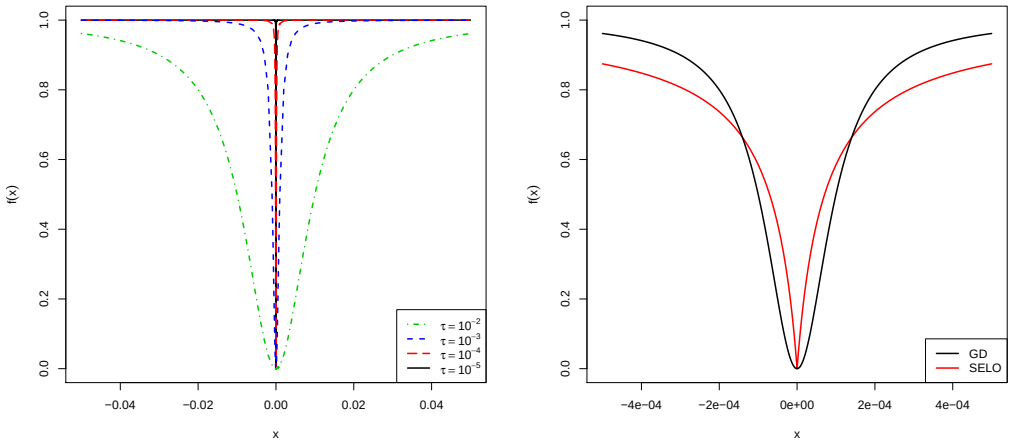

FIGURE 4.1: (left) graphs of $f(x) = x^2/(\tau^2 + x^2)$ for $\tau = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$; (right) GD versus SELO.

gamma (GD) prior by

$$\pi_{\mathrm{GD}}(\boldsymbol{\beta}, \mathbf{d}) \quad \propto \quad \pi_{\mathrm{G}}(\boldsymbol{\beta}|\mathbf{d})\pi_{\mathrm{D}}(\mathbf{d}), \tag{4.7}$$

such that

$$\begin{aligned} \pi_{\mathrm{G}}(\boldsymbol{\beta}|\mathbf{d}) &\propto \prod_{j=1}^{p} \left\{ d_j^{1/2} \exp\left(-\frac{d_j}{2}\beta_j^2\right)\right\}, \\ \pi_{\mathrm{D}}(\mathbf{d}) &\propto \prod_{j=1}^{p} \left\{ d_j^{(\lambda-1)/2} \exp\left(-\frac{\tau_0^2}{2\{(\mathbf{X}_{[,j]})^{\mathbf{T}}\mathbf{X}_{[,j]}\}}d_j\right)\right\}, \end{aligned}$$

where $\lambda \geq 0$ and τ_0 is determined to be sufficiently small. The following lemma reveals the relationship between the GD prior and the newly defined penalty in (4.6).

Lemma 4.2. *Let $\pi_{GD}(\boldsymbol{\beta}, \mathbf{d})$ be the defined GD prior in (4.7). Then, for any $\lambda \geq 0$*

and $\tau_0 > 0$*, we have that*

$$\max_{\mathbf{d}} \{\pi_{GD}(\boldsymbol{\beta}, \mathbf{d})\} \propto \exp\left(-\lambda \sum_{j=1}^{p} \frac{\{(\mathbf{X}_{[,j]})^{\mathbf{T}}\mathbf{X}_{[,j]}\}\beta_j^2}{\tau_0^2 + \{(\mathbf{X}_{[,j]})^{\mathbf{T}}\mathbf{X}_{[,j]}\}\beta_j^2}\right).$$

The proof of Lemma 4.2 can be easily shown by differentiating $\pi_{\mathrm{GD}}(\boldsymbol{\beta}, \mathbf{d})$ with respect to d_j 's, letting them to be zero, and finding the solutions of the equations. From Lemma 4.2, we can easily show that

$$\arg\min_{\boldsymbol{\beta}} \left[\min_{\mathbf{d}} \{f_\psi(\mathbf{y}|\boldsymbol{\beta})\pi_{\mathrm{GD}}(\boldsymbol{\beta}, \mathbf{d})\}\right] = \arg\min_{\boldsymbol{\beta}} \left[\mathrm{BD}_\psi \{\mathbf{y}, \mathbf{h}(\mathbf{X}\boldsymbol{\beta})\} + \tilde{\ell}_0(\boldsymbol{\beta}|\lambda, \tau_0)\right],$$

for $\lambda \geq 0$ and $\tau_0 > 0$. Hence, owing to Lemma 4.1, for a sufficiently small value τ_0, our MAP estimator of $\boldsymbol{\beta}$ well approximates the penalized Bregman divergence estimator with the ℓ_0-norm penalty. Furthermore, owing to our Bayesian framework, its uncertainty can be explained by the induced posterior distribution. Based on Figure 4.1, we throughout this chapter define $\tau_0 = 10^{-20}$ and it works very well in our numerical studies.

4.2.3 The posterior

The key idea of our GD prior-based approach is to replace the auxiliary $\mathbf{d}$ by its MAP estimate rather than integrating out. Let $\hat{\mathbf{d}}_{\mathrm{MAP}}$ denote the MAP estimate of $\mathbf{d}$. Then, the posterior is given by

$$\begin{aligned} \pi(\boldsymbol{\beta}|\mathbf{y}) &\propto f_\psi(\mathbf{y}|\boldsymbol{\beta})\pi_{\mathrm{GD}}(\boldsymbol{\beta}, \hat{\mathbf{d}}_{\mathrm{MAP}}) \\ &\propto f_\psi(\mathbf{y}|\boldsymbol{\beta})\pi_{\mathrm{G}}(\boldsymbol{\beta}|\hat{\mathbf{d}}_{\mathrm{MAP}}). \end{aligned} \tag{4.8}$$

Before implementing the posterior inference, it is important to check the propriety of the posterior distribution. The following lemma asserts that our approach guarantees the proper posterior.

Lemma 4.3. *Suppose that the prior* $\pi(\boldsymbol{\beta})$ *is proper and the likelihood* $f(\mathbf{y}|\boldsymbol{\beta})$ *satisfies* $\sup_{\boldsymbol{\beta}\in\mathbb{R}^p} f(\mathbf{y}|\boldsymbol{\beta}) < \infty$*, then the posterior* $\pi(\boldsymbol{\beta}|\mathbf{y})$ *is also proper.*

Since $\pi_{\mathrm{G}}(\boldsymbol{\beta}|\hat{\mathbf{d}}_{\mathrm{MAP}})$ in (4.8) is the multivariate Gaussian density function for $\boldsymbol{\beta}$, it is proper. Note that $\mathrm{BD}_{\psi}(\mathbf{z}_1, \mathbf{z}_2) = \psi(\mathbf{z}_1) - \psi(\mathbf{z}_2) - (\mathbf{z}_1 - \mathbf{z}_2)^{\mathbf{T}} \nabla\psi(\mathbf{z}_2) \geq 0$ for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^p$, due to the convexity of ψ. Hence, $f_{\psi}(\mathbf{y}|\boldsymbol{\beta}) \propto \exp\left[-\mathrm{BD}_{\psi}\left\{\mathbf{y}, \mathbf{h}\left(\mathbf{X}\boldsymbol{\beta}\right)\right\}\right] \leq 1$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$. From Lemma 4.3, hence, our posterior is proper.

4.3 Posterior computation

Due to the high-dimensionality of parameter space, our Bayesian approach requires special care in posterior computation. In this section, we introduce new ICM algorithm and MCMC sampling to make a Bayesian inference for high-dimensional problems. In addition, we discuss the optimal determination of the hyperparameter from a Bayesian perspective. For notational simplicity, in this section we omit the subscript "MAP" which indicates a MAP estimator; throughout this section, $\hat{\boldsymbol{\beta}}_{\mathrm{MAP}}$ and $\hat{\mathbf{d}}_{\mathrm{MAP}}$ are denoted by $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{d}}$, respectively.

4.3.1 Maximum A Posteriori estimation

It is straightforward to check that our target estimator $\hat{\boldsymbol{\beta}}(= \arg\max_{\boldsymbol{\beta}}\{\pi(\boldsymbol{\beta}|\mathbf{y})\})$ can be obtained by

$$(\hat{\boldsymbol{\beta}}, \hat{\mathbf{d}}) = \arg\max_{\boldsymbol{\beta},\mathbf{d}} \{f_{\psi}(\mathbf{y}|\boldsymbol{\beta})\pi_{\mathrm{G}}(\boldsymbol{\beta}|\mathbf{d})\pi_{\mathrm{D}}(\mathbf{d})\}.$$

From this aspect, we propose to obtain $\hat{\boldsymbol{\beta}}$ using the full posterior,

$$\pi(\boldsymbol{\beta}, \mathbf{d}|\mathbf{y}) \propto f_{\psi}(\mathbf{y}|\boldsymbol{\beta})\pi_{\mathrm{G}}(\boldsymbol{\beta}|\mathbf{d})\pi_{\mathrm{D}}(\mathbf{d}).$$

Note that the full conditionals are given by

$$\begin{aligned} \pi(\mathbf{d}|\text{others}) &\propto \pi_{\mathrm{G}}(\boldsymbol{\beta}|\mathbf{d})\pi_{\mathrm{D}}(\mathbf{d}), & (4.9)\\ \pi(\boldsymbol{\beta}|\text{others}) &\propto f_{\psi}(\mathbf{y}|\boldsymbol{\beta})\pi_{\mathrm{G}}(\boldsymbol{\beta}|\mathbf{d}). & (4.10) \end{aligned}$$

Then, using the ICM algorithm, $\hat{\boldsymbol{\beta}}$ can be obtained by iteratively updating the current $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)}$ as follows:

$$\begin{aligned} \mathbf{d}^{(t+1)} &\leftarrow \arg\max_{\mathbf{d}} \left\{\pi_{\mathrm{G}}(\boldsymbol{\beta}^{(t)}|\mathbf{d})\pi_{\mathrm{D}}(\mathbf{d})\right\},\\ \boldsymbol{\beta}^{(t+1)} &\leftarrow \arg\max_{\boldsymbol{\beta}} \left\{f_{\psi}(\mathbf{y}|\boldsymbol{\beta})\pi_{\mathrm{G}}(\boldsymbol{\beta}|\mathbf{d}^{(t+1)})\right\},\\ \hat{\boldsymbol{\beta}} &= \boldsymbol{\beta}^{(t+1)}, \end{aligned}$$

until convergence. However, in the presence of high-dimensionality, this algorithm is infeasible. To overcome this difficulty, we develop a new component-wise updating ICM algorithm.

Lemma 4.4. *Let* $q(\boldsymbol{\beta}) = BD_{\psi}\left\{\mathbf{y}, \mathbf{h}(\mathbf{X}\boldsymbol{\beta})\right\}$. *Then, the gradient vector and the Hessian matrix of* $q(\boldsymbol{\beta})$ *at* $\boldsymbol{\beta}_0$ *are respectively given by*

$$\left.\frac{\partial q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \mathbf{X}^{\mathbf{T}}\mathbf{z}(\boldsymbol{\beta}_0) \quad \textit{and} \quad \left.\frac{\partial^2 q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathbf{T}}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \mathbf{X}^{\mathbf{T}}\mathbf{S}(\boldsymbol{\beta}_0)\mathbf{X},$$

where $\mathbf{z}(\boldsymbol{\beta}_0) = [z_1(\boldsymbol{\beta}_0), \ldots, z_n(\boldsymbol{\beta}_0)]^{\mathbf{T}}$ *and* $\mathbf{S}(\boldsymbol{\beta}_0) = \mathbf{Diag}\{s_1(\boldsymbol{\beta}_0), \ldots, s_n(\boldsymbol{\beta}_0)\}$ *with*

$$\begin{aligned}
z_i(\boldsymbol{\beta}_0) &= -\{y_i - h(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}_0)\}\psi''\left(h(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}_0)\right) h'(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}_0), \\
s_i(\boldsymbol{\beta}_0) &= \{h'(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}_0)\}^2\psi''\left(h(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}_0)\right) \\
&\quad -\{y_i - h(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}_0)\}\psi'''\left(h(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}_0)\right)\{h'(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}_0)\}^2 \\
&\quad -\{y_i - h(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}_0)\}\psi''\left(h(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}_0)\right) h''(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}_0).
\end{aligned}$$

Using Lemma 4.4, it is straightforward to show that the Laplace approximation to $f_{\psi}(\mathbf{y}|\boldsymbol{\beta})$ at the current state $\boldsymbol{\beta}^{(t)}$ is given by $\tilde{f}_{\psi}(\mathbf{y}|\boldsymbol{\beta}) = \phi\left(\boldsymbol{\beta}|\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}\right)$ with

$$\begin{aligned}
\boldsymbol{\mu}^{(t)} &= \left[\mathbf{X}^T\mathbf{S}(\boldsymbol{\beta}^{(t)})\mathbf{X}\right]^{-1}\left[\mathbf{X}^T\mathbf{S}(\boldsymbol{\beta}^{(t)})\mathbf{X}^T\boldsymbol{\beta}_0 - \mathbf{X}^T\mathbf{Z}(\boldsymbol{\beta}^{(t)})\right] \\
\boldsymbol{\Sigma}^{(t)} &= \left[\mathbf{X}^T\mathbf{S}(\boldsymbol{\beta}^{(t)})\mathbf{X}\right]^{-1},
\end{aligned}$$

where $\phi\left(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ indicates a multivariate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. By replacing $f_{\psi}(\mathbf{y}|\boldsymbol{\beta})$ with $\tilde{f}_{\psi}(\mathbf{y}|\boldsymbol{\beta})$ in (4.10), the full conditionals can be viewed as

$$[d_j|\text{others}] \sim \text{Gamma}\left(\frac{\lambda}{2} + 1, \frac{\tau_0^2/\{(\mathbf{X}_{[,j]})^{\mathbf{T}}\mathbf{X}_{[,j]}\} + \beta_j^2}{2}\right), \tag{4.11}$$

$$[\beta_j|\text{others}] \approx \text{N}\left(\mu_j, \sigma_j\right), \tag{4.12}$$

for $j = 1, \ldots, p$, where

$$\begin{aligned} \mu_j &= \frac{\mathbf{X}_{[,j]}^{\mathbf{T}}\mathbf{S}(\boldsymbol{\beta}^{(t)})\mathbf{X}\boldsymbol{\beta}^{(t)} - \mathbf{X}_{[,j]}^{\mathbf{T}}\mathbf{z}(\boldsymbol{\beta}^{(t)}) - \mathbf{X}_{[,j]}^{\mathbf{T}}\mathbf{S}(\boldsymbol{\beta}^{(t)})\mathbf{X}_{[,-j]}\boldsymbol{\beta}_{-j}}{d_j + \mathbf{X}_{[,j]}^{\mathbf{T}}\mathbf{S}(\boldsymbol{\beta}^{(t)})\mathbf{X}_{[,j]}}, \\ \sigma_j &= \frac{1}{d_j + \mathbf{X}_{[,j]}^{\mathbf{T}}\mathbf{S}(\boldsymbol{\beta}^{(t)})\mathbf{X}_{[,j]}}, \end{aligned}$$

where $\mathbf{X}_{[,-j]}$ denotes the sub-matrix of $\mathbf{X}$ without the j^{th} column and $\boldsymbol{\beta}_{-j}$ denotes the sub-vector of $\boldsymbol{\beta}$ without its j^{th} component. Using (4.11) and (4.12), we derive the following component-wise updating ICM algorithm.

Algorithm 4.5. *(ICM algorithm)*

Set *an initial value* $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(0)}$ *and a threshold value* $\xi(> 0)$.

For $j = 1, \ldots, p$.

If $|\hat{\beta}_j| > 0$,

Update $\hat{\beta}_j$ *by*

$$\begin{aligned} \boldsymbol{\beta}^{(t)} &\leftarrow \hat{\boldsymbol{\beta}} \\ d_j^{(t+1)} &\leftarrow \frac{\lambda}{\tau_0^2/\{(\mathbf{X}_{[,j]})^{\mathbf{T}}\mathbf{X}_{[,j]}\} + (\beta_j^{(t)})^2}, \\ \beta_j^{(t+1)} &\leftarrow \frac{\mathbf{X}_{[,j]}^{\mathbf{T}}\mathbf{S}(\boldsymbol{\beta}^{(t)})\mathbf{X}_{[,j]}\beta_j^{(t)} - \mathbf{X}_{[,j]}^{\mathbf{T}}\mathbf{z}(\boldsymbol{\beta}^{(t)})}{d_j^{(t+1)} + \mathbf{X}_{[,j]}^{\mathbf{T}}\mathbf{S}(\boldsymbol{\beta}^{(t)})\mathbf{X}_{[,j]}}, \\ \hat{\beta}_j &= sign(\beta_j^{(t+1)})(|\beta_j^{(t+1)}| - \xi)\mathbf{1}\{|\beta_j^{(t+1)}| > \xi\}, \end{aligned}$$

Repeat *until convergence.*

Return $\hat{\boldsymbol{\beta}}$.

Note that, owing to the soft thresholding method in the proposed ICM algorithm,

$\hat{\boldsymbol{\beta}}$ contains zero values. By excluding the zero-valued coordinates from the update, our algorithm overcomes the high-dimensionality. In all our numerical studies, we set the tolerance level $\xi = 10^{-10}$ to determine zero estimates.

4.3.2 Posterior sampling

Keep in mind that the posterior $\pi(\boldsymbol{\beta}|\mathbf{y}) \propto f_\psi(\mathbf{y}|\boldsymbol{\beta})\pi_{\mathrm{G}}(\boldsymbol{\beta}|\hat{\mathbf{d}})$ is of our interest. To obtain the MCMC sample from our target posterior, the Gibbs sampler can be used by iteratively generating β_j's from the full conditionals. However, in general, the conditional distribution of β_j may not have an explicit form. To overcome this difficulty, we propose to use the Metropolis-Hastings within the Gibbs sampler with a proposal density $p(\beta_j^{(t+1)}|\beta_j^{(t)})$ generating the move from the current state state $\beta_j^{(t)}$ to a new state $\beta_j^{(t+1)}$. Hence, the proposed moves are accepted with probabilities

$$\alpha = \min\left\{1, \frac{f_\psi(\mathbf{y}|\beta_j^{(t+1)}, \boldsymbol{\beta}_{-j}^{(t)})\pi_{\mathrm{GD}}(\beta_j^{(t+1)}, \boldsymbol{\beta}_{-j}^{(t)})p(\beta_j^{(t)}|\beta_j^{(t+1)})}{f_\psi(\mathbf{y}|\beta_j^{(t)}, \boldsymbol{\beta}_{-j}^{(t)})\pi_{\mathrm{GD}}(\beta_j^{(t)}, \boldsymbol{\beta}_{-j}^{(t)})p(\beta_j^{(t+1)}|\beta_j^{(t)})}\right\}.$$

From (4.12), we propose to define the proposal distribution $p(\cdot|\beta_j^{(t)})$ as a Gaussian distribution with the mean $\beta_j^{(t)}$ and the variance $\left[\mathbf{X}_{[,j]}^{\mathbf{T}}\mathbf{S}(\hat{\boldsymbol{\beta}})\mathbf{X}_{[,j]} + \hat{d}_j\right]^{-1}$, where $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{d}})$ indicates the obtained MAP estimate.

Remark 4.6. Let $\mathcal{B} = \{\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(m)}\}$ be the obtained MCMC sample of size m from the posterior. Then, we can easily construct the exact $100 \times (1-\alpha)\%$ credible intervals or Highest Posterior Density (HPD) intervals (Chen and Shao, 1999).

Remark 4.7. If $\hat{\beta}_j$ is determined to be zero by the ICM algorithm, then we recommend to update $\beta_j^{(t+1)} = 0$ with probability one in the Gibbs chain so that we can reduce the computational time under the curse of high-dimensionality.

4.3.3 Prior specification

In our Bayesian approach, the determination of the hyperparameter λ is extremely crucial because it controls the degree of penalization (or the sparsity on the MAP estimator $\hat{\boldsymbol{\beta}}_{\text{MAP}}$).

Let $\mathbf{d}_\lambda$ be the obtained MAP estimate of $\mathbf{d}$ for a given λ. To select the optimal λ, from Bayesian perspective, we propose to utilize the prior predictive density $m(\mathbf{y}|\lambda) = \int f_\psi(\mathbf{y}|\boldsymbol{\beta})\pi_{\text{G}}(\boldsymbol{\beta}|\mathbf{d}_\lambda)d\boldsymbol{\beta}$. Using $m(\mathbf{y}|\lambda)$, we define the optimal λ^* such that $\lambda^* = \arg\max_\lambda\{m(\mathbf{y}|\lambda\}$. In many cases, however, $m(\mathbf{y}|\lambda)$ would not be expressed in a closed form. To calculate $m(\mathbf{y}|\lambda)$ in such cases, we propose to use the Importance-Weighted Marginal Density Estimation (IWMDE) (Chen, 1994),

$$\tilde{m}(\mathbf{y}) = \left[\frac{1}{m}\sum_{s=1}^{m}\frac{\phi(\boldsymbol{\beta}^{(s)}|\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})}{f(\mathbf{y}|\boldsymbol{\beta}^{(s)})\pi_{\text{G}}(\boldsymbol{\beta}^{(s)}|\mathbf{d}_\lambda)}\right]^{-1}, \tag{4.13}$$

where $\{\beta^{(s)}, s = 1, \ldots, m\}$ is a set of MCMC samples from $\pi(\boldsymbol{\beta}|\mathbf{y})$, $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{\beta}^{(s)}$, and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \frac{1}{m-1}\sum_{j=1}^{m}(\boldsymbol{\beta}^{(s)} - \boldsymbol{\mu}_{\boldsymbol{\beta}})(\boldsymbol{\beta}^{(s)} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^{\mathbf{T}}$.

4.4 Numerical studies

4.4.1 Simulation studies

To compare the estimation performance of our Bayesian method with widely used PL methods (Ridge, Elastic-net, LASSO, and adaptive LASSO), Monte Carlo experiments are conducted in this section. We generate 500 data sets from each of the following three models: for $i = 1, \ldots, n$,

1. $y_i \overset{iid}{\sim} N(\mu_i, 1)$ with $\mu_i = h_1(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta})$, where $\boldsymbol{\beta} = (2, 2.5, -2, -2.5, \text{rep}(0, p-4))$, $h_1(x) = x$, and $\mathbf{x}_i \overset{iid}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\Sigma_{ij})_{p\times p} = (1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p$ and $\rho = 0.2$.

2. $y_i \overset{iid}{\sim} Bernoulli(p_i)$ with $p_i = h_2(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta})$, where $\boldsymbol{\beta} = (1.5, 2, -2, \text{rep}(0, p-3))$, $h_2(x) = \frac{1}{1+\exp(-x)}$, and $\mathbf{x}_i \overset{iid}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\Sigma_{ij})_{p\times p} = (1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p$ and $\rho = 0.2$.

3. $y_i \overset{iid}{\sim} Poisson(\mu_i)$ with $\mu_i = h_3(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta})$, where $\boldsymbol{\beta} = (2, 2.2, \text{rep}(0, p-2))$, $h_3(x) = \exp(x)$, and $\mathbf{x}_i = \boldsymbol{\Phi}(\mathbf{z}_i) - 0.5\mathbf{1}_p$ such that $\mathbf{z}_i \overset{iid}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = (\Sigma_{ij})_{p\times p} = (1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p$, $\rho = 0.2$, $\boldsymbol{\Phi}(\mathbf{z}_i) = (\Phi(z_{i1}), \ldots, \Phi(z_{ip}))^{\mathbf{T}}$, and $\Phi(\cdot)$ is the CDF of standard normal distribution.

To specify the likelihood, based on Table 4.1, we define

$$\begin{aligned}
\psi_1(\mathbf{x}) &= \sum_{i=1}^{n} \left\{ \frac{x_i^2}{2} \right\}, \\
\psi_2(\mathbf{x}) &= \sum_{i=1}^{n} \left\{ x_i \log x_i + (1-x_i)\log(x_i - 1) \right\}, \\
\psi_3(\mathbf{x}) &= \sum_{i=1}^{n} \left\{ x_i \log x_i \right\},
\end{aligned}$$

for Model 1, 2, and 3, respectively. The hyperparameter λ is determined by (4.13). All the PL methods are implemented by **R** package **glmnet**, where the tuning parameter is determined by the 10–fold cross validation. We measure the estimation accuracy using the following two types of mean squared error (MSE):

$$\text{MSE}_{\text{est}} = \frac{1}{p}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2; \quad \text{MSE}_{\text{pred}} = \frac{1}{n}\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

To assess the variable selection performance, we calculate False Positive Rate (FPR) and False Negative Rate (FNR) as follows:

$$\text{FPR}\% = 100 \times \frac{\text{FP}}{\text{TN} + \text{FP}}; \quad \text{FNR}\% = 100 \times \frac{\text{FN}}{\text{TP} + \text{FN}},$$

where TP, FP, TN and FN denote the numbers of true non-zeros, false non-zeros, true zeros and false zeros, respectively. Table 4.2 summarizes our Monte Carlo simulation result. It clearly shows that our GD method always performs better than all the PL methods. Furthermore, our GD method is comparable to an ideal estimation method, say *oracle*, in which the true zero coefficients are forced to be zero and the remaining non-zero coefficients are estimated by the Ordinary Least Squares (OLS).

4.4.2 Predictive binary classification: Leukemia data

In practice, especially in genetics study, a researcher conducts a pre-screening procedure such as Sure Independence Screening (SIS) (Fan and Lv, 2008; Fan and Song, 2010) to reduce the ultra-high dimensionality ($n \ll p$) prior to the estimation. In this section, we study collaborative performance of our proposed method and SIS for classification problem using *Leukemia data* (Fan and Lv, 2008); the data set is available at **R** package **SIS**.

This data set consists of 72 samples with 7,129 genes, where 38 samples and the remaining 34 are defined as *training set* and *test set*, respectively. For the i^{th} observation, the response variable y_i is a binary outcome, indicating the types of acute leukemia (Acute Lymphoblastic Leukemia=0 and Acute Myeloid Leukemia=1) and

TABLE 4.2: The MC estimates of MSE_{est}, MSE_{pred}, FPR%, and FNR%, where Alasso_1, Alasso_2, and Alasso_3 indicate adative lasso with ridge, elastic-net, and lasso weights, respectively.

Model	(n,p)		Oracle	GD	Alasso_1	Alasso_2	Alasso_3	Ridge	E-net	Lasso
1	(200,200)	MSE_{est}	0.0001	0.0001	0.0008	0.0025	0.0018	0.0233	0.0013	0.0008
		MSE_{pred}	0.0195	0.0207	0.1191	0.3721	0.2840	0.7980	0.1787	0.1207
		FPR%	0.0000	0.0092	8.8878	14.8939	8.8551	100.0000	15.0388	8.8837
		FNR%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	(200,500)	MSE_{est}	0.0000	0.0000	0.0004	0.0010	0.0007	0.0339	0.0007	0.0004
		MSE_{pred}	0.0196	0.0214	0.1528	0.4131	0.3216	7.5333	0.2324	0.1508
		FPR%	0.0000	0.0056	4.7524	8.0060	4.6766	100.0000	8.1569	4.7056
		FNR%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	(200,200)	MSE_{est}	0.0019	0.0022	0.0112	0.0403	0.0408	0.0371	0.0175	0.0112
		MSE_{pred}	0.0023	0.0039	0.0177	0.0541	0.0458	0.0530	0.0262	0.0175
		FPR%	0.0000	0.1117	10.7025	17.8437	10.4487	100.0000	19.3218	10.7005
		FNR%	0.0000	0.0000	0.0000	0.0000	0.2000	0.0000	0.0000	0.0000
	(200,500)	MSE_{est}	0.0009	0.0019	0.0076	0.0169	0.0185	0.0272	0.0123	0.0076
		MSE_{pred}	0.0023	0.0041	0.0233	0.0701	0.0616	0.0931	0.0330	0.0233
		FPR%	0.0000	0.0020	6.2052	9.6314	6.0165	100.0000	11.1421	6.2213
		FNR%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	(200,200)	MSE_{est}	0.0004	0.0010	0.0032	0.0099	0.0075	0.0310	0.0052	0.0033
		MSE_{pred}	0.0300	0.0580	0.1905	0.5411	0.4361	0.7366	0.2653	0.1940
		FPR%	0.0000	0.1737	5.4192	9.2333	5.6747	100.0000	9.6717	5.7677
		FNR%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	(200,500)	MSE_{est}	0.0002	0.0004	0.0015	0.0046	0.0034	0.0157	0.0026	0.0016
		MSE_{pred}	0.0292	0.0595	0.2275	0.6280	0.5112	0.9518	0.3159	0.2335
		FPR%	0.0000	0.0594	2.7289	4.5538	2.7795	100.0000	4.9382	2.8675
		FNR%	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

the predictor vector $\mathbf{x}_i$ gives the expression levels of 7,129 genes. Define the probability of being Acute Myeloid Leukemia (AML) for the i^{th} sample as p_i =Probability($y_i = 1$). The link function h is defined as $p_i = h(\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}) = \left[1 + \exp\left(-\mathbf{x}_i^{\mathbf{T}}\boldsymbol{\beta}\right)\right]^{-1}$, i.e., logit link. To specify the convex function, from Table 4.1, we define

$$\psi(\mathbf{x}) = \sum_{i=1}^{n} \{x_i \log x_i + (1 - x_i) \log(x_i - 1)\},$$

which induces the Bernoulli likelihood. First, we conduct a pre–screening procedure to reduce the ultra–high–dimensionality using SIS. According to Fan and Lv (2008),

we select the top $2n/\log(n) \approx 21$ genes and then implement all methods in previous section to determine the classification model. In the GD method, we set $\lambda = 1$ based on the marginal density computed by (4.13). In PL methods, the regularization parameter is determined by the 10–fold cross validation as in the previous simulation study. To predict the outcome, $\hat{p}_i = 0.5$ is used as a cut-off point. As a result, Table 4.3 shows that GD method and Elastic-net have the lowest classification error rate for the test set. The GD method identifies two genes, while the other methods select more than 10 genes. Note that the selected two genes by the GD method are commonly identified by all methods. Hence, we conclude that our GD method provides the best classification performance.

TABLE 4.3: Contingency table for test set

Prediction		True ALL	True AML	Seleceted Genes
GD	ALL	20	1	6,20
	AML	0	13	(2 out of 21)
Alasso_1	ALL	20	3	2,6,7,9,10,12,13,15,18,19,20
	AML	0	11	(11 out of 21)
Alasso_2	ALL	20	3	6,7,8,9,10,12,13,15,18,19,20
	AML	0	11	(11 out of 21)
Alasso_3	ALL	20	3	6,7,9,10,12,13,15,18,19,20
	AML	0	11	(10 out of 21)
Ridge	ALL	20	2	1 – 21
	AML	0	12	(21 out of 21)
E-net	ALL	20	1	1,2,4,5,6,7,8,9,10,11,12,13,14,15,17,18,19,20,21
	AML	0	13	(19 out of 21)
Lasso	ALL	20	3	6,7,9,10,12,13,15,18,19,20
	AML	0	11	(10 out of 21)

4.5 Concluding remarks

From a Bayesian perspective, we develop a new approach to sparse high–dimensional problems using Bregman divergence and a valid ℓ_0-norm approximation. To determine the optimal hyperparameter, we suggest to utilize the prior predictive density, which is easy to calculate with MCMC sample from the posterior. To reduce the computational burden of MCMC computation, alternatively, we can use the Bayesian Information Criterion (BIC), which is the second order approximation to the prior predictive density (Schwarz, 1978).

In the PL approach, cross-validation techniques have been widely used to choose the optimal tuning parameter from a viewpoint of predictive model selection. Similarly, Conditional Predictive Ordinate (CPO) can be utilized in our framework from a perspective of predictive Bayesian model selection. The CPO calculation can be easily accomplished by a single set of MCMC sample (Gelfand and Dey, 1994). Hence, this CPO-based approach would be computationally more efficient than other cross-validation methods.

One of the advantages of our divergence-based approach is that many extensions can be easily developed by assigning a new divergence measure in the likelihood function. For example, using Bregman matrix divergence (Kulis et al., 2009), our model can be adapted to multivariate regression models, which is a work in progress by the authors.

Chapter 5

Bayesian Sparse and Reduced-rank Regression

5.1 Introduction

In various fields of scientific research such as genomics, economics, image processing, astronomy, etc., massive amount of data are routinely collected, and many associated statistical problems can be cast in the framework of multivariate regression, where both the number of response variables and the number of predictors are possibly of high dimensionality. For example, in genomics study, it is critical to explore the relationship between genetic markers and gene expression profiles in order to understand the gene regulatory network; in a study of human lung disease mechanism, the detailed CT-scanned lung imaging data enable us to examine the systematic variations in airway tree measurements across various lung disease status and pulmonary function test results. To formulate, suppose we have n independent observations of the

response vector $\mathbf{y}_i \in \mathbb{R}^q$ and the predictor vector $\mathbf{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$. Consider the multivariate linear regression model

$$\mathbf{Y} = \mathbf{X}\mathbf{C} + \mathbf{E}, \tag{5.1}$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^{\mathbf{T}} \in \mathbb{R}^{n \times q}$ is the response matrix, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathbf{T}} \in \mathbb{R}^{n \times p}$ is the predictor matrix, $\mathbf{C} \in \mathbb{R}^{p \times q}$ is the unknown regression coefficient matrix, and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)^{\mathbf{T}} \in \mathbb{R}^{n \times q}$ is the error matrix with $\mathbf{e}_i$'s being independently and identically distributed (i.i.d.) with mean zero. We assume the response variables and the predictors are all centered, and there is no intercept term. (In what follows, we use $\mathbf{a}_j^{\mathbf{T}}$ to denote the j^{th} row of a generic matrix $\mathbf{A}$ and $\tilde{\mathbf{a}}_l$ the l^{th} column of $\mathbf{A}$, e.g., $\mathbf{C} = [\mathbf{c}_1\ \mathbf{c}_2\ \cdots\ \mathbf{c}_p]^{\mathbf{T}} = [\tilde{\mathbf{c}}_1\ \tilde{\mathbf{c}}_2\ \cdots\ \tilde{\mathbf{c}}_q]$.) A fundamental goal of multivariate regression is thus to estimate and make inference about the coefficient matrix $\mathbf{C}$ so that meaningful dependence structure between the responses and predictors can be revealed.

When the predictor dimension p and the response dimension q are large relative to the sample size n, classical estimation methods such as ordinary least squares (OLS) may fail miserably. The curse of dimensionality can be mitigated by assuming that $\mathbf{C}$ admits certain low-dimensional structures, and regularization/penalization approaches are then commonly deployed to conduct dimension reduction and model estimation. The celebrated reduced rank regression (RRR) (Anderson, 1951; Izenman, 1975; Reinsel and Velu, 1998) achieved dimension reduction through constraining the coefficient matrix $\mathbf{C}$ to be rank deficient, building upon the belief that the response variables are related to the predictors through only a few latent directions, i.e., some linear combinations of the original predictors. As such, low-rank structure induces and

models dependency among responses, which is the essence of conducting multivariate analysis. Bunea et al. (2011) generalized the classical RRR to high dimensional settings, casting reduced-rank estimation as a penalized least squares problem with the penalty being proportional to the rank of $\mathbf{C}$. Yuan et al. (2007) utilized the nuclear norm penalty, defined as the l_1 norm of the singular values. See also, Negahban and Wainwright (2011), Rohde and Tsybakov (2011), Mukherjee and Zhu (2011) and Chen et al. (2013).

It is clear that low-rankness in $\mathbf{C}$ is of intrinsic multivariate nature, which, when further combined with other structures such as sparsity and/or smoothness, can further lift dimension reduction and facilitate model interpretation. For example, in the aforementioned genomics study, it is plausible that the gene expression profiles (responses) and the genetic markers (predictors) are associated through only a few latent pathways (linear combinations of possibly highly-correlated genetic markers), and moreover, very likely such linear associations only involve a small subset of genetic markers and/or gene profiles. Therefore, recovering a low-rank and also sparse coefficient matrix $\mathbf{C}$ in model (5.1) hold the key to reveal such interesting connections between the responses and predictors. Chen et al. (2012) proposed a regularized sparse singular value decomposition (SVD) approach with known rank, in which each latent variable is constructed from only a subset of the predictors and is associated with only a subset of the responses. Chen and Huang (2012) proposed a rank-constrained adaptive group Lasso approach to recover a low-rank coefficient matrix $\mathbf{C}$ with sparse rows; for each zero row in $\mathbf{C}$, the corresponding predictor is then completely eliminated from the model. Bunea et al. (2012) also proposed a joint sparse and low-rank estimation approach and derived its nonasymptotic oracle error bounds. Both methods required to solve the nonconvex rank-constrained problem by fitting models of

various ranks. Recently, Ma and Sun (2014) proposed a subspace assisted regression with row sparsity method which was shown to achieve near optimal nonasymptotic minimax rates in estimation.

While all the aforementioned regularized regression techniques produce attractive point estimators of the coefficient matrix $\mathbf{C}$, it remains a difficult problem to assess the uncertainty of the obtained estimators. To overcome this limitation, there has already been a rich literature on Bayesian approaches of the reduced rank regression. From a Bayesian perspective, the unknown parameter is considered as a random variable, and thus the statistical inference can be made by the posterior distribution. The first attempt to develop the Bayesian reduced rank regression was made by Geweke (1996). The coefficient matrix is assumed to be $\mathbf{C} = \mathbf{AB^T}$ with $\mathbf{A} \in \mathbb{R}^{p\times r}$ and $\mathbf{B} \in \mathbb{R}^{q\times r}$, where $r < \min\{p, q\}$ is assumed to be known. Then, by assigning Gaussian prior on $(\mathbf{A}, \mathbf{B})$, the induced posterior achieves the low-rank structure of the prespecified rank. As an alternative, Lim and Teh (2007) proposed to start from the largest possible rank $r = \min\{p, q\}$, assign a column-wise shrinkage Gaussian prior on each columns of $\mathbf{A}$ and $\mathbf{B}$. The posterior for redundant columns of $\mathbf{A}$ and $\mathbf{B}$ is forced to be concentrated around zero, so the (approximate) rank reduction can be accomplished. The main challenge of this Bayesian approach is the choice of the hyperparameters of the Gaussian priors in order to control the amount of shrinkage. There have been several attempts to overcome this challenge by assigning priors on the hyperparameters, so that they can be determined in the estimation procedure. For instance, Salakhutdinov and Mnih (2008) proposed to utilize the Wishart distribution as the hyperprior. Similar hierarchical Bayesian methods were also proposed in the context of matrix completion, matrix completion deals with missing values, but we do not (Zhou et al., 2010; Babacan et al., 2011). However, none of the aforementioned

studies dealt with the sparsity of the coefficient matrix $\mathbf{C}$. Recently, Zhu et al. (2014) introduced a Bayesian low-rank regression model with high-dimensional responses and covariates. To enable sparse estimation under low rank constraint with a prefixed rank, they utilized a sparse singular value decomposition (SVD) structure (Chen et al., 2012) with Gaussian-mixtures of gamma priors on all the elements of the decomposed matrices. Then, the sparsity of $\mathbf{C}$ was achieved using Bayesian thresholding method. For a survey on Bayesian reduce rank models, see Alquier (2013) and the references therein.

We develop in this chapter a novel Bayesian simultaneous dimension reduction and variable selection approach. Our method aims to tackle several challenges regarding both the estimation and inference in the sparse and low-rank regression problems. First, the proposed method enables us to simultaneously estimate the unknown rank and remove irrelevant predictors, in contrast to several existing methods in which rank selection has to be resolved by comparing fitted models of various ranks or by some ad hoc approach such as scree plot. In addition, we also seek potential column sparsity of the coefficient matrix, so that it is applicable to problems with high-dimensional responses where response selection is highly desirable (to be elaborated below). Second, by careful construction of the prior distribution, our method alleviates the many difficulties brought by the use of nonsmooth and nonconvex penalty functions and by the tuning parameter selection procedure in penalized regression analysis. From a Bayesian perspective, the penalty function can be viewed as a negative logarithm of the prior density function (Tibshirani, 1996; Park and Casella, 2008; Kyung et al., 2010). We develop a general prior for $\mathbf{C}$ mimicking the rank penalty and the group l_0 row/column penalty, to achieve simultaneous rank reduction and variable selection through the induced posterior distribution, yet the computation is kept tractable

and efficient. Since the tuning parameters are considered as random variables in our Bayesian formulation, the optimal ones are selected to achieve the highest posterior probability given the data. Furthermore, using our Bayesian approach, the credibility intervals for the regression coefficients and their functions can be easily constructed using the Markov Chain Monte Carlo (MCMC) technique. In contrast, there has been little work on quantifying the estimation uncertainty in regularized regression approaches.

We now formally state our assumptions or prior beliefs about the coefficient matrix $\mathbf{C}$ in model (5.1).

A1. (Reduced rank) $r^* \leq r$, where $r^* = \text{rank}(\mathbf{C})$ indicates the rank of $\mathbf{C}$ and $r = \min(p, q)$.

A2. (Row-wise sparsity) $p^* \leq p$, where $p^* = \mathbf{card}\{j : \mathbf{c}_j^\mathbf{T}\mathbf{c}_j \neq 0\}$ and $\mathbf{c}_j^\mathbf{T}$ denotes the j^{th} row of $\mathbf{C}$, where $\mathbf{card}\{\cdot\}$ denotes the cardinality of a set.

A3. (Column-wise sparsity) $q^* \leq q$, where $q^* = \mathbf{card}\{l : \tilde{\mathbf{c}}_l^\mathbf{T}\tilde{\mathbf{c}}_l \neq 0\}$ and $\tilde{\mathbf{c}}_l$ denotes the l^{th} column of $\mathbf{C}$.

A1 states that $\mathbf{C}$ is possibly of low rank. In A2, excluding the j^{th} predictor from model (5.1) is equivalent to setting all entries of the j^{th} row of $\mathbf{C}$ as zero. Therefore, the first two assumptions concern rank reduction and predictor selection. The third assumption is about "response selection", i.e., if the l^{th} column of $\mathbf{C}$ is zero, the l^{th} response is modeled as a noise variable. While such structural assumption can be treated as optional depending on the specific application, we stress that there are many circumstances where response selection is highly desirable. For example, in many applications the dimension of the responses can be very high, and there may exist noise variables that are not related to any predictors in the model. It is

also possible that some responses only relate to some predictors nonlinearly, hence it would not be sensible to force them appear in a linear model setup. Nonetheless, allowing possible response selection provides more flexibility in the multivariate linear regression framework.

The remainder of the chapter is organized as follows. In Section 5.2, we briefly introduce a general penalized regression approach for conducting sparse and low-rank estimation. In Section 5.3, we develop our new Bayesian approach, and explore the connections between the PLS and our Bayes estimators. The full conditionals are obtained in Section 5.4, and we describe the posterior optimization algorithm and posterior sampling technique. In Section 5.5, we study the posterior consistency of the proposed method. Simulation studies and a real application on yeast cycle data are presented in Section 5.6 and Section 5.7. Some concluding remarks are given in Section 5.8. Lastly, in Section 5.9, using Bregman divergence we introduce a general extension of our proposed method.

5.2 Penalized regression approach

In the regularized estimation framework, the unknown coefficient matrix $\mathbf{C}$ in model (5.1) can be estimated by the following penalized least squares (PLS) method,

$$\hat{\mathbf{C}}_{\text{pls}} = \arg\min_{\mathbf{C}} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_F^2 + \mathcal{P}_{\boldsymbol{\lambda}}(\mathbf{C}) \right\}, \tag{5.2}$$

where $\|\mathbf{C}\|_F = \sqrt{\mathbf{tr}(\mathbf{C}^\mathbf{T}\mathbf{C})} = \sqrt{\mathbf{tr}(\mathbf{C}\mathbf{C}^\mathbf{T})}$ denotes the Frobenius norm, and $\mathcal{P}_{\boldsymbol{\lambda}}(\mathbf{C})$ a penalty function with non-negative tuning parameter $\boldsymbol{\lambda}$ controlling the amount of

regularization. It is natural to construct a penalty function of an additive form,

$$\mathcal{P}_{\boldsymbol{\lambda}}(\mathbf{C}) = \mathcal{P}_{\lambda_1}^{\mathrm{RR}}(\mathbf{C}) + \mathcal{P}_{\lambda_2}^{\mathrm{RS}}(\mathbf{C}) + \mathcal{P}_{\lambda_3}^{\mathrm{CS}}(\mathbf{C}),$$

where $\mathcal{P}_{\lambda_1}^{\mathrm{RR}}(\mathbf{C})$, $\mathcal{P}_{\lambda_2}^{\mathrm{RS}}(\mathbf{C})$ and $\mathcal{P}_{\lambda_3}^{\mathrm{CS}}(\mathbf{C})$ induce the low-rankness, row-wise sparsity and column-wise sparsity in $\mathbf{C}$, with tuning parameters λ_1, λ_2 and λ_3, respectively. There are numerous choices of the penalty functions. Note that the rank of matrix $\mathbf{C}$ is same as the number of non-zero singular values, i.e., $\mathrm{rank}(\mathbf{C}) = \mathbf{card}\{k : s_k(\mathbf{C}) > 0\} = r^*$, where $s_k(\mathbf{C})$ denotes the rank and the k^{th} singular value of $\mathbf{C}$. Hence, rank reduction can be achieved by penalizing the singular values of $\mathbf{C}$, i.e.,

$$\mathcal{P}_{\lambda_1}^{\mathrm{RR}}(\mathbf{C}) = \lambda_1 \sum_{k=1}^{r} \rho_1(s_k(\mathbf{C})), \tag{5.3}$$

where ρ_1 is a sparsity-inducing penalty function. In particular, choosing $\rho_1(|a|) = \mathbf{1}\{|a| \neq 0\}$ corresponds to directly penalizing/restraining the rank of $\mathbf{C}$, and that $\rho_1(|a|) = |a|^{\beta_1}$ gives the Schatten-β quasi-norm penalty when $0 < \beta_1 < 1$ and the convex nuclear norm penalty $\lambda_1\|\mathbf{C}\|_*$ when $\beta_1 = 1$. For promoting row-wise/column-wise sparsity, selecting or eliminating parameters by groups is needed, which can be achieved by penalizing the row/column ℓ_2 norms of $\mathbf{C}$,

$$\mathcal{P}_{\lambda_2}^{\mathrm{RS}}(\mathbf{C}) = \frac{1}{2}\lambda_2 \sum_{j=1}^{p} \rho_2(\|\mathbf{c}_j\|_2), \tag{5.4}$$

$$\mathcal{P}_{\lambda_3}^{\mathrm{CS}}(\mathbf{C}) = \frac{1}{2}\lambda_3 \sum_{l=1}^{q} \rho_3(\|\tilde{\mathbf{c}}_l\|_2), \tag{5.5}$$

where $\|\mathbf{c}\|_2 = \sqrt{\mathbf{c}^{\mathbf{T}}\mathbf{c}}$ denotes the ℓ_2 norm. Choosing $\rho_2(|a|) = \mathbf{1}\{|a| \neq 0\}$ corresponds to directly counting and penalizing the number of nonzero rows, and $\rho_2(|a|) = |a|$

corresponds to the convex group Lasso penalty (Yuan and Lin, 2006). Other methods include group SCAD (Fan and Li, 2001) and group MCP (Breheny and Huang, 2009; Zhang, 2010); see Fan and Song (2010) and Huang et al. (2012) for comprehensive reviews. In principal, rank reduction and variable selection can be accomplished by solving the PLS problem (5.2) with any sparsity-inducing penalties ρ_1, ρ_2 and ρ_3.

The pros and cons of using convex penalties in model selection are well understood. In low-rank estimation, Bunea et al. (2011) showed that while the convex nuclear norm penalized estimator has similar estimation properties to those of the nonconvex rank penalized estimator, the former requires stronger conditions and is in general not as parsimonious as the latter in rank selection. For sparse group selection, it is known that the convex group Lasso criterion often leads to over-selection and substantial estimation bias, and adopting nonconvex penalties may lead to superior properties in both model estimation and variable selection under milder conditions (Huang et al., 2008; Ma and Sun, 2014). Unfortunately, the nonconvexity of a penalized regression criterion also imposes great challenges in both understanding its theoretical properties and solving the optimization problem in computation. Therefore, trading off computation efficiency and statistical properties is critical in formulating penalized estimation criterion, and it is particularly relevant when dealing with large data applications. The problem of tuning parameter selection can also be troublesome, especially so for the problem of interest here as it requires multiple tuning parameters. Furthermore, it is still a largely unsolved problem on how to make statistical inference and attach error measures to any penalized estimator. All these concerns motivate us to tackle the sparse and low-rank estimation problem in a Bayesian fashion, to achieve a computationally efficient implementation and be able to make valid inference about the composite low-dimensional structure of $\mathbf{C}$.

5.3 Bayesian sparse and low-rank regression

From a Bayesian perspective, the PLS estimate in (5.2) can be viewed as the maximum a posteriori (MAP) estimate from the following posterior density function,

$$\begin{aligned}\pi(\mathbf{C} \mid \mathbf{Y}, \boldsymbol{\lambda}) &\propto f(\mathbf{Y} \mid \mathbf{C})\pi(\mathbf{C} \mid \boldsymbol{\lambda}) \\ &\propto \exp\left(-\frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_F^2\right) \exp\left\{-\mathcal{P}_{\boldsymbol{\lambda}}(\mathbf{C})\right\},\end{aligned}$$

where $f(\mathbf{Y} \mid \mathbf{C})$ denotes the likelihood function and $\pi(\mathbf{C} \mid \boldsymbol{\lambda})$ denotes the prior density function of $\mathbf{C}$ given the tuning parameter $\boldsymbol{\lambda}$. Motivated by the connections between PLS and MAP and by the penalty function defined in (5.3)–(5.5), it is natural to consider the following prior,

$$\pi(\mathbf{C} \mid \boldsymbol{\lambda}) \propto \exp\left[-\frac{1}{2}\left\{\lambda_1 \sum_{k=1}^{r} \ell_0(|s_k(\mathbf{C})|) + \lambda_2 \sum_{j=1}^{p} \ell_0(\|\mathbf{c}_j\|_2) + \lambda_3 \sum_{l=1}^{q} \ell_0(\|\tilde{\mathbf{c}}_l\|_2)\right\}\right] \quad (5.6)$$

where $\ell_0(a) = \mathbf{1}\{a \neq 0\}$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$. In penalized regression, this ℓ_0 penalty corresponds to directly penalize the rank, the number of nonzero rows, and the number of nonzero columns of $\mathbf{C}$, which leads to an intractable combinatory problem. Similarly, in Bayesian framework, even though this setup directly targets on the desired structure of $\mathbf{C}$, there are several difficulties in using such a prior distribution. Since the prior density function in (5.6) involves the singular values of $\mathbf{C}$, it induces an improper posterior distribution, as it is generally difficult to be considered as a probability density function of $\mathbf{C}$. Moreover, the non-differentiability of the ℓ_0 functions at zero induces a discontinuous posterior density function.

To overcome the first difficulty, i.e., avoiding direct use of the singular values,

we propose an indirect modeling method through decomposing the matrix $\mathbf{C}$. We write $\mathbf{C} = \mathbf{A}\mathbf{B}^{\mathbf{T}}$, where $\mathbf{A}$ is a $p \times r$ matrix, $\mathbf{B}$ is a $q \times r$ matrix, and r is an upper bound of the true rank r^* of $\mathbf{C}$, e.g., a trivial one is $r = \min(p, q)$. Apparently such a decomposition is not unique, as with any nonsingular $r \times r$ matrix $\mathbf{Q}$, $\mathbf{C} = \mathbf{A}\mathbf{B}^{\mathbf{T}} = \mathbf{A}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{B}^{\mathbf{T}} = \tilde{\mathbf{A}}\tilde{\mathbf{B}}^{\mathbf{T}}$ where $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{Q}$ and $\tilde{\mathbf{B}} = \mathbf{B}\left(\mathbf{Q}^{-1}\right)^{\mathbf{T}}$. Interestingly, the following lemma reveals that the low-rankness and the row/column sparsity of $\mathbf{C}$ can all be represented as certain row/column sparsity of $\mathbf{A}$ and $\mathbf{B}$, and more importantly, the representations are invariant to any nonsingular transformation.

Lemma 5.1. *Let $\mathbf{C} \in \mathbb{R}^{p\times q}$, and suppose $\mathbf{C} = \mathbf{A}\mathbf{B}^{\mathbf{T}}$ for some $\mathbf{A} \in \mathbb{R}^{p\times r}$, $\mathbf{B} \in \mathbb{R}^{q\times r}$ with $r = \min(p, q)$. Let $\tilde{\mathbf{a}}_k$ and $\tilde{\mathbf{b}}_k$ denote the k^{th} column of $\mathbf{A}$ and $\mathbf{B}$, respectively. Let $\mathbf{a}_j^{\mathbf{T}}$ and $\mathbf{b}_l^{\mathbf{T}}$ denote the j^{th} row of $\mathbf{A}$ and the l^{th} row of $\mathbf{B}$, respectively. Then*

1. *rank*$(\mathbf{C}) \leq \min(p, q) - \sum_{k=1}^{r} \mathbf{1}\{\|\tilde{\mathbf{a}}_k\|_2 + \|\tilde{\mathbf{b}}_k\|_2 = 0\}$.
2. $\{j : \|\mathbf{c}_j\|_2 \neq 0\} \subset \{j : \|\mathbf{a}_j\|_2 \neq 0\}$.
3. $\{l : \|\tilde{\mathbf{c}}_l\|_2 \neq 0\} \subset \{l : \|\mathbf{b}_l\|_2 \neq 0\}$.

The first statement in the above lemma suggests the following rank-reducing prior,

$$\pi^{\mathrm{RR}}(\mathbf{A}, \mathbf{B} \mid \lambda_1) \;\propto\; \exp\left[-\frac{1}{2}\lambda_1 \sum_{k=1}^{r} \mathbf{1}\left\{\|\tilde{\mathbf{a}}_k\|_2 + \|\tilde{\mathbf{b}}_k\|_2 \neq 0\right\}\right], \tag{5.7}$$

where $\lambda_1 > 0$, which induces sparsity on columns of $\mathbf{A}$ and $\mathbf{B}$ simultaneously and thus reduces the rank of $\mathbf{C}$. Similarly, Lemma 5.1 suggests the following row-wise

and column-wise sparsity-inducing priors,

$$\pi^{\mathrm{RS}}(\mathbf{A} \mid \lambda_2) \;\propto\; \exp\left[-\frac{1}{2}\lambda_2 \sum_{j=1}^{p} \mathbf{1}\left\{\|\mathbf{a}_j\|_2 \neq 0\right\}\right], \tag{5.8}$$

$$\pi^{\mathrm{CS}}(\mathbf{B} \mid \lambda_3) \;\propto\; \exp\left[-\frac{1}{2}\lambda_3 \sum_{l=1}^{q} \mathbf{1}\left\{\|\mathbf{b}_l\|_2 \neq 0\right\}\right], \tag{5.9}$$

where $\lambda_2 > 0$ and $\lambda_3 > 0$. Combining (5.7), (5.8) and (5.9) leads us to the following prior,

$$\pi(\mathbf{A}, \mathbf{B} \mid \boldsymbol{\lambda}) \;\propto\; \pi^{\mathrm{RR}}(\mathbf{A}, \mathbf{B} \mid \lambda_1)\pi^{\mathrm{RS}}(\mathbf{A} \mid \lambda_2)\pi^{\mathrm{CS}}(\mathbf{B} \mid \lambda_3), \tag{5.10}$$

where $\lambda = (\lambda_1, \lambda_2, \lambda_3)$.

The discontinuity problem still presents in (5.10) due to use of the ℓ_0 norm. We address this problem by approximating ℓ_0 by a well-behaved smooth function. Let $\mathbf{D}$ be a $m \times n$ matrix. Define

$$\mathcal{P}_{\lambda,\omega}(\mathbf{D}) = \frac{1}{2}\lambda \sum_{i=1}^{m} \frac{\mathbf{d}_i^{\mathbf{T}}\mathbf{d}_i}{\left(\omega + \mathbf{d}_i^{\mathbf{T}}\mathbf{d}_i\right)^{1-\frac{\beta}{2}}}, \tag{5.11}$$

where $0 \leq \beta \leq 1$ and $\omega > 0$. We have

$$\frac{\mathbf{d}_i^{\mathbf{T}}\mathbf{d}_i}{\left(\omega + \mathbf{d}_i^{\mathbf{T}}\mathbf{d}_i\right)^{1-\frac{\beta}{2}}} \longrightarrow \left(\|\mathbf{d}_i\|_2\right)^{\beta},$$

as $\omega \to 0$, where we define $(0)^0 = 0$. This implies that the proposed penalty approximates the group ℓ_β-norm penalty when ω is sufficiently small. In particular, when $\beta = 0$ and ω is chosen to be a small positive constant ω_0, (5.11) gives an approximate group ℓ_0-norm penalty and produces approximately sparse solutions, while it is con-

tinuous as well as differentiable with respect to $\mathbf{D}$. In all our numerical studies, we set $\omega_0 = 10^{-10}$ and utilize tolerance level 10^{-5} to determine zero estimates. Figure 5.1 shows the plots of $f(x) = (x^2)/(x^2 + \omega_0)$ for varying $\omega_0 = 10^{-k}$, $k = 4, 6, 8, 10$. Indeed, when $w_0 = 10^{-10}$, the function closely mimics the ℓ_0 penalty.

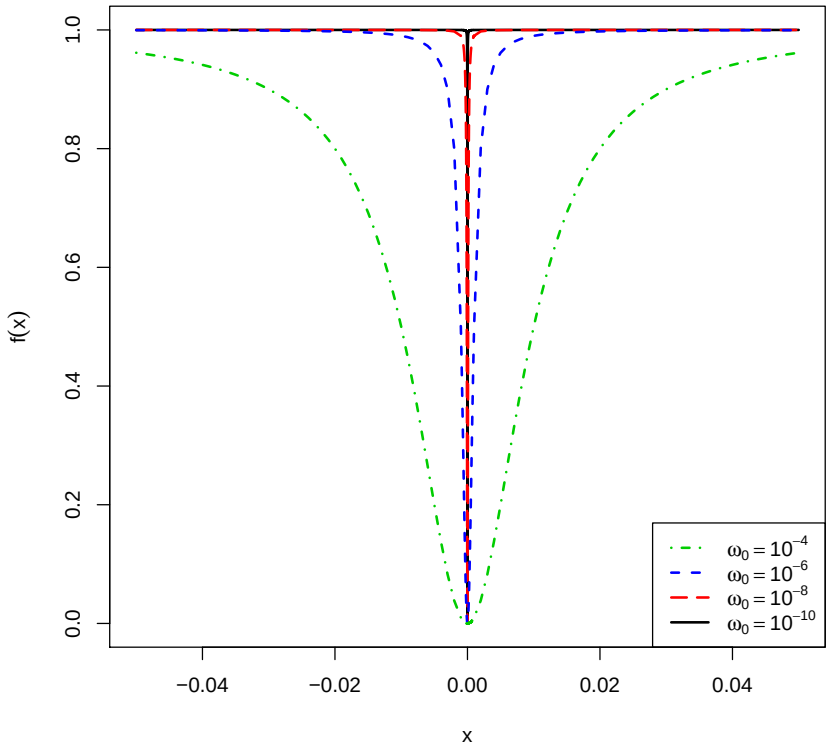

FIGURE 5.1: Plot of $f(x) = \frac{x^2}{x^2+\omega_0}$ for $\omega_0 = 10^{-k}$, where $k = 4, 6, 8, 10$.

We now propose the following prior distribution for our Bayesian Sparse and Reduced-rank Regression (BSRR) method,

$$\pi^{\text{BSRR}}(\mathbf{A}, \mathbf{B} \mid \boldsymbol{\lambda}) \propto \exp\left(-\frac{1}{2}\lambda_1 \sum_{k=1}^{r} \frac{\tilde{\mathbf{a}}_k^{\mathbf{T}}\tilde{\mathbf{a}}_k + \tilde{\mathbf{b}}_k^{\mathbf{T}}\tilde{\mathbf{b}}_k}{\omega_0 + \tilde{\mathbf{a}}_k^{\mathbf{T}}\tilde{\mathbf{a}}_k + \tilde{\mathbf{b}}_k^{\mathbf{T}}\tilde{\mathbf{b}}_k}\right) \times \exp\left(-\frac{1}{2}\lambda_2 \sum_{j=1}^{p} \frac{\mathbf{a}_j^{\mathbf{T}}\mathbf{a}_j}{\omega_0 + \mathbf{a}_j^{\mathbf{T}}\mathbf{a}_j} - \frac{1}{2}\lambda_3 \sum_{l=1}^{q} \frac{\mathbf{b}_l^{\mathbf{T}}\mathbf{b}_l}{\omega_0 + \mathbf{b}_l^{\mathbf{T}}\mathbf{b}_l}\right) \quad (5.12)$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ and $\lambda_i \geq 0$ $(i = 1, 2, 3)$. The BSRR posterior is then given as

$$\pi^{\text{BSRR}}(\mathbf{A}, \mathbf{B} \mid \mathbf{Y}, \boldsymbol{\lambda}) \;\; \propto \;\; f\left(\mathbf{Y} \mid \mathbf{C} = \mathbf{A}\mathbf{B}^{\mathbf{T}}\right) \pi^{\text{BSRR}}(\mathbf{A}, \mathbf{B} \mid \boldsymbol{\lambda}), \tag{5.13}$$

and the MAP estimate $(\hat{\mathbf{A}}_{\text{map}}, \hat{\mathbf{B}}_{\text{map}})$ is defined as

$$(\hat{\mathbf{A}}_{\text{map}}, \hat{\mathbf{B}}_{\text{map}}) = \arg\max_{\mathbf{A},\mathbf{B}} \left\{\pi(\mathbf{A}, \mathbf{B} \mid \mathbf{Y}, \boldsymbol{\lambda})\right\}.$$

Since $\mathbf{C} = \mathbf{A}\mathbf{B}^{\mathbf{T}}$, the MAP estimator for $\mathbf{C}$, named BSRR estimator, is given by $\hat{\mathbf{C}}_{\text{BSRR}} = \hat{\mathbf{A}}_{\text{map}} \left(\hat{\mathbf{B}}_{\text{map}}\right)^{\mathbf{T}}$.

The following lemma tells us the BSRR model can be expressed as a hierarchical Bayesian model by introducing auxiliary variables.

Lemma 5.2. *Define* $\mathbf{d} = (d_{1,1}, \ldots, d_{1,r}, d_{2,1}, \ldots, d_{2,p}, d_{3,1}, \ldots, d_{3,q})$. *Let* $\pi(\mathbf{A}, \mathbf{B}, \mathbf{d} \mid \boldsymbol{\lambda})$ *be a density function of* $(\mathbf{A}, \mathbf{B}, \mathbf{d})$ *such that*

$$\begin{aligned}
\pi(\mathbf{A}, \mathbf{B}, \mathbf{d} \mid \boldsymbol{\lambda}) \;\; &\propto \;\; \pi(\mathbf{A}, \mathbf{B} \mid \mathbf{d}) \pi(\mathbf{d} \mid \boldsymbol{\lambda}) \\
&\propto \;\; \exp\left\{-\frac{1}{2}\left(\sum_{k=1}^{r} d_{1,k}(\tilde{\mathbf{a}}_k^{\mathbf{T}}\tilde{\mathbf{a}}_k + \tilde{\mathbf{b}}_k^{\mathbf{T}}\tilde{\mathbf{b}}_k)\right)\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}\left(\sum_{j=1}^{p} d_{2,j}(\mathbf{a}_j^{\mathbf{T}}\mathbf{a}_j) + \sum_{l=1}^{q} d_{3,l}(\mathbf{b}_l^{\mathbf{T}}\mathbf{b}_l)\right)\right\} \\
&\quad \times \prod_{i=1}^{3}\left[\prod_{j=1}^{m_i}\left\{(d_{i,j})^{\frac{\lambda_i}{2}} \exp\left(-\frac{\omega_0}{2} d_{i,j}\right)\right\}\right],
\end{aligned} \tag{5.14}$$

where $m_1 = r$, $m_2 = p$ *and* $m_3 = q$. *Let* $\pi^{BSRR}(\mathbf{A}, \mathbf{B} \mid \boldsymbol{\lambda})$ *denote the BSRR prior defined in (5.12). Then, for any positive* $\boldsymbol{\lambda}$ *and* ω_0, *we have that*

$$\pi^{BSRR}(\mathbf{A}, \mathbf{B} \mid \boldsymbol{\lambda}) \propto \max_{\mathbf{d}} \left\{\pi(\mathbf{A}, \mathbf{B}, \mathbf{d} \mid \boldsymbol{\lambda})\right\}.$$

The proof of Lemma 5.2 can be shown by differentiating Eq. (5.14) with respect to $d_{i,j}$'s, letting them to be zero, and finding the solutions of the equations. Based on Lemma 5.2, we introduce the following hierarchical Bayesian representation of the sparse reduced-rank regression model (HBSRR),

$$
\begin{aligned}
f(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}) &\propto \exp\left(-\frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{A}\mathbf{B}^{\mathbf{T}}\|_F^2\right), && (5.15)\\
\pi(\mathbf{A}, \mathbf{B} \mid \mathbf{d}) &\propto \exp\left\{-\frac{1}{2}\left(\|\mathbf{A}\mathbf{D}_1^{1/2}\|_F^2 + \|\mathbf{D}_2^{1/2}\mathbf{A}\|_F^2\right)\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}\left(\|\mathbf{B}\mathbf{D}_1^{1/2}\|_F^2 + \|\mathbf{D}_3^{1/2}\mathbf{B}\|_F^2\right)\right\}, && (5.16)\\
\mathbf{D}_i^{1/2} &= \mathbf{diag}(\sqrt{d_{i,1}}, \ldots, \sqrt{d_{i,m_i}}),\ i = 1, 2, 3, \\
\pi(\mathbf{d} \mid \boldsymbol{\lambda}) &\propto \prod_{i=1}^{3}\left[\prod_{j=1}^{m_i}\left\{(d_{i,j})^{\frac{\lambda_i}{2}} \exp\left(-\frac{\omega_0}{2}d_{i,j}\right)\right\}\right], && (5.17)
\end{aligned}
$$

where $m_1 = r$, $m_2 = p$, $m_3 = q$, $\mathbf{d} = (\mathbf{diag}(\mathbf{D}_1), \mathbf{diag}(\mathbf{D}_2), \mathbf{diag}(\mathbf{D}_3))$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$, and $d_{i,m_i} > 0$ $(i = 1, 2, 3)$. Let $(\hat{\mathbf{A}}_{\text{mode}}, \hat{\mathbf{B}}_{\text{mode}}, \hat{\mathbf{d}}_{\text{mode}})$ be the mode of the induced posterior $\pi(\mathbf{A}, \mathbf{B}, \mathbf{d} \mid \mathbf{Y}, \boldsymbol{\lambda})$ from the above hierarchical model. Recall that $\hat{\mathbf{C}}_{\text{BSRR}}$ indicates the BSRR estimator. Then, using Lemma 5.2, it is straightforward to show that $\hat{\mathbf{C}}_{\text{BSRR}} = \hat{\mathbf{A}}_{\text{mode}}\left(\hat{\mathbf{B}}_{\text{mode}}\right)^{\mathbf{T}}$, almost surely. This enables us to easily find $\hat{\mathbf{C}}_{\text{BSRR}}$ using the HBSRR. Since all full conditional distributions of the HBSRR are well-known distributions such as Gaussian and gamma, the estimation procedure can be conducted by standard Bayesian estimation algorithms.

5.4 Bayesian analysis

Since our posterior distribution is complex, the Bayesian inference procedure requires the implementation of iterated conditional modes (ICM) algorithm (Besag, 1986) or

Markov chain Monte Carlo (MCMC) sampling techniques, to obtain Bayes estimators such as posterior mode, posterior mean, or credible set. We derive full conditional distributions from the joint posterior of HBSRR. Then, we describe the implementation of ICM and MCMC, and discuss the determination of the tuning parameter from a Bayesian perspective.

5.4.1 Full conditionals

To derive the full conditionals of the HBSRR, we write

$$\begin{aligned}
\|\mathbf{Y}-\mathbf{XAB^T}\|_F^2 &= \mathbf{tr}\left\{\left(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B^T}\right)\left(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B^T}\right)^{\mathbf{T}}\right\} \\
&\quad +\mathbf{tr}\left(\tilde{\mathbf{x}}_j\mathbf{a}_j^{\mathbf{T}}\mathbf{B^T}\mathbf{B}\mathbf{a}_j\tilde{\mathbf{x}}_j^{\mathbf{T}}\right)-2\,\mathbf{tr}\left\{\tilde{\mathbf{x}}_j\mathbf{a}_j^{\mathbf{T}}\mathbf{B^T}\left(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B^T}\right)^{\mathbf{T}}\right\} \\
&= \mathbf{tr}\left\{\left(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B^T}\right)\left(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B^T}\right)^{\mathbf{T}}\right\} \\
&\quad +\mathbf{a}_j^{\mathbf{T}}\mathbf{B^T}\mathbf{B}\mathbf{a}_j\tilde{\mathbf{x}}_j^{\mathbf{T}}\tilde{\mathbf{x}}_j-2\mathbf{a}_j^{\mathbf{T}}\mathbf{B^T}\left(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B^T}\right)^{\mathbf{T}}\tilde{\mathbf{x}}_j. \qquad (5.18)
\end{aligned}$$

Here we use the notation $\mathbf{C}_{(j)}$ to denote the submatrix of a generic matrix $\mathbf{C}$ by deleting its j^{th} row, and $\mathbf{C}_{(\tilde{j})}$ by deleting its j^{th} column. Using (5.15), (5.16) and (5.18), the full conditional distribution of $\mathbf{a}_j$ $(j=1,\ldots,p)$ is determined to be

$$\mathbf{a}_j \mid \text{Others} \overset{ind}{\sim} \mathrm{N}_r\left(\boldsymbol{\mu}_j^{\mathbf{A}}, \boldsymbol{\Sigma}_j^{\mathbf{A}}\right), \qquad (5.19)$$

where

$$\begin{aligned}
\boldsymbol{\mu}_j^A &= \boldsymbol{\Sigma}_j^A\mathbf{B^T}\left(\mathbf{Y}-\mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B^T}\right)^{\mathbf{T}}\tilde{\mathbf{x}}_j, \\
\boldsymbol{\Sigma}_j^A &= \left(\mathbf{B^T}\mathbf{B}\tilde{\mathbf{x}}_j^{\mathbf{T}}\tilde{\mathbf{x}}_j+\mathbf{D}_1+d_{2,j}\mathbf{I}_r\right)^{-1},
\end{aligned}$$

with $\mathbf{I}_r$ denoting the $r \times r$ identity matrix. Similar to (5.18), we have

$$\begin{aligned} \|\mathbf{Y} - \mathbf{XAB}^{\mathbf{T}}\|_F^2 &= \mathbf{tr}\left\{\left(\mathbf{Y}_{(\tilde{l})} - \mathbf{XA}(\mathbf{B}_{(\tilde{l})})^{\mathbf{T}}\right)^{\mathbf{T}}\left(\mathbf{Y}_{(\tilde{l})} - \mathbf{XA}(\mathbf{B}_{(\tilde{l})})^{\mathbf{T}}\right)\right\} \\ &\quad + \mathbf{b}_l^{\mathbf{T}}(\mathbf{XA})^{\mathbf{T}}\mathbf{XA}\mathbf{b}_l - 2\mathbf{b}_l^{\mathbf{T}}(\mathbf{XA})^{\mathbf{T}}\tilde{\mathbf{y}}_l + \tilde{\mathbf{y}}_l^{\mathbf{T}}\tilde{\mathbf{y}}_l. \end{aligned} \tag{5.20}$$

The full conditional distribution of $\mathbf{b}_l$, for $l = 1, \ldots, q$, is given by

$$\mathbf{b}_l \mid \text{Others} \overset{ind}{\sim} \mathrm{N}_r\left(\boldsymbol{\mu}_l^{\mathbf{B}}, \boldsymbol{\Sigma}_l^{\mathbf{B}}\right), \tag{5.21}$$

where

$$\begin{aligned} \boldsymbol{\mu}_l^B &= \boldsymbol{\Sigma}_l^{\mathbf{B}}(\mathbf{XA})^{\mathbf{T}}\tilde{\mathbf{y}}_l, \\ \boldsymbol{\Sigma}_l^B &= \left((\mathbf{XA})^{\mathbf{T}}\mathbf{XA} + \mathbf{D}_1 + d_{3,l}\mathbf{I}_r\right)^{-1}. \end{aligned}$$

From (5.16) and (5.17), it is straightforward to show that the full conditionals for elements of $\mathbf{D}$ are written as

$$d_{1,k} \mid \text{Others} \overset{ind}{\sim} \text{Gamma}\left(\frac{\lambda_1}{2}, \frac{\omega_0 + \tilde{\mathbf{a}}_k^T\tilde{\mathbf{a}}_k + \tilde{\mathbf{b}}_k^T\tilde{\mathbf{b}}_k}{2}\right), \tag{5.22}$$

$$d_{2,j} \mid \text{Others} \overset{ind}{\sim} \text{Gamma}\left(\frac{\lambda_2}{2}, \frac{\omega_0 + \mathbf{a}_j^T\mathbf{a}_j}{2}\right), \tag{5.23}$$

$$d_{3,l} \mid \text{Others} \overset{ind}{\sim} \text{Gamma}\left(\frac{\lambda_3}{2}, \frac{\omega_0 + \mathbf{b}_l^T\mathbf{b}_l}{2}\right), \tag{5.24}$$

where $k = 1, \ldots, r$, $j = 1, \ldots, p$, and $l = 1, \ldots, q$.

5.4.2 Iterated conditional modes

All full conditionals in Section 5.4.1 are well-known distributions (normal or gamma distribution), and the modes are thus well-known. Consequently, using the full conditionals, we construct the following ICM algorithm to find the BSRR estimate $\hat{\mathbf{C}}_{\text{BSRR}}$:

Algorithm 5.3 (ICM algorithm for $\hat{\mathbf{C}}_{\text{BSRR}}$)**.**

Set *initial values* $\left(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{d}}\right) = \left(\mathbf{A}^{(0)}, \mathbf{B}^{(0)}, \mathbf{d}^{(0)}\right)$.

Update $\left(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{d}}\right) = \left(\mathbf{A}^{(t+1)}, \mathbf{B}^{(t+1)}, \mathbf{d}^{(t+1)}\right)$ *by*

$$
\begin{aligned}
\mathbf{a}_j^{(t+1)} &\leftarrow \left((\mathbf{B}^{(t)})^{\mathbf{T}}\mathbf{B}^{(t)}\tilde{\mathbf{x}}_j^{\mathbf{T}}\tilde{\mathbf{x}}_j + \mathbf{D}_1^{(t)} + d_{2,j}^{(t)}\mathbf{I}_r\right)^{-1} \\
&\quad \times (\mathbf{B}^{(t)})^{\mathbf{T}}\left(\mathbf{Y} - \mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}^{(t)}(\mathbf{B}^{(t)})^{\mathbf{T}}\right)^{\mathbf{T}}\tilde{\mathbf{x}}_j, \\
\mathbf{b}_l^{(t+1)} &\leftarrow \left(\left(\mathbf{X}\mathbf{A}^{(t+1)}\right)^{\mathbf{T}}\mathbf{X}\mathbf{A}^{(t+1)} + \mathbf{D}_1^{(t)} + d_{3,l}^{(t)}\mathbf{I}_r\right)^{-1}(\mathbf{X}\mathbf{A}^{(t+1)})^{\mathbf{T}}\tilde{\mathbf{y}}_l, \\
d_{1,k}^{(t+1)} &\leftarrow \lambda_1\left(\omega_0 + (\tilde{\mathbf{a}}_k^{(t+1)})^{\mathbf{T}}\tilde{\mathbf{a}}_k^{(t+1)} + (\tilde{\mathbf{b}}_k^{(t+1)})^{\mathbf{T}}\tilde{\mathbf{b}}_k^{(t+1)}\right)^{-1}, \\
d_{2,j}^{(t+1)} &\leftarrow \lambda_2\left[\omega_0 + (\mathbf{a}_j^{(t+1)})^{\mathbf{T}}\mathbf{a}_j^{(t+1)}\right]^{-1}, \\
d_{3,l}^{(t+1)} &\leftarrow \lambda_3\left[\omega_0 + (\mathbf{b}_l^{(t+1)})^{\mathbf{T}}\mathbf{b}_l^{(t+1)}\right]^{-1},
\end{aligned}
$$

for $k = 1, \ldots, r$, $j = 1, \ldots, p$ *and* $l = 1, \ldots, q$.

Repeat *until convergence.*

Return $\hat{\mathbf{C}}_{BSRR} = \hat{\mathbf{A}}\hat{\mathbf{B}}^{\mathbf{T}}$.

To set an initial value $(\mathbf{A}^{(0)}, \mathbf{B}^{(0)}, \mathbf{d}^{(0)})$, we propose to utilize the OLS estimate $\hat{\mathbf{C}}_{\text{ols}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$. Let $\hat{\mathbf{C}}_{\text{ols}} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathbf{T}}$ be the sigular value decomposition of $\hat{\mathbf{C}}_{\text{ols}}$.

Then, $(\mathbf{A}^{(0)}, \mathbf{B}^{(0)}, \mathbf{d}^{(0)})$ can be defined as

$$\begin{aligned}
\mathbf{A}^{(0)} &= \mathbf{U}\mathbf{S}^{1/2}, \quad \mathbf{B}^{(0)} = \mathbf{V}\mathbf{S}^{1/2}, \\
d_{1,k}^{(0)} &= \lambda_1 \left(\omega_0 + (\tilde{\mathbf{a}}_k^{(0)})^{\mathbf{T}} \tilde{\mathbf{a}}_k^{(0)} + (\tilde{\mathbf{b}}_k^{(0)})^{\mathbf{T}} \tilde{\mathbf{b}}_k^{(0)} \right)^{-1}, \\
d_{2,j}^{(0)} &= \lambda_2 \left[\omega_0 + (\mathbf{a}_j^{(0)})^{\mathbf{T}} \mathbf{a}_j^{(0)} \right]^{-1}, \\
d_{3,l}^{(0)} &= \lambda_3 \left[\omega_0 + (\mathbf{b}_l^{(0)})^{\mathbf{T}} \mathbf{b}_l^{(0)} \right]^{-1},
\end{aligned}$$

for $k = 1, \ldots, r$, $j = 1, \ldots, p$ and $l = 1, \ldots, q$.

5.4.3 Posterior sampling

Using the HBSRR, we introduce indirect sampling method to obtain the posterior samples for $\mathbf{C}$, so that they can be used to construct the credible set for the BSRR estimate $\hat{\mathbf{C}}_{\text{BSRR}}$. Recall that $(\hat{\mathbf{A}}_{\text{mode}}, \hat{\mathbf{B}}_{\text{mode}}, \hat{\mathbf{d}}_{\text{mode}})$ denotes the posterior mode of the HBSRR. Then,

$$\begin{aligned}
\hat{\mathbf{C}}_{\text{BSRR}} &= \arg \max_{\mathbf{C}=\mathbf{A}\mathbf{B}^{\mathbf{T}}} \left\{ \max_{\mathbf{d}} \pi(\mathbf{A}, \mathbf{B}, \mathbf{d} \mid \mathbf{Y}, \boldsymbol{\lambda}) \right\} \\
&= \arg \max_{\mathbf{C}=\mathbf{A}\mathbf{B}^{\mathbf{T}}} \left\{ f(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}) \pi(\mathbf{A}, \mathbf{B} \mid \hat{\mathbf{d}}_{\text{mode}}) \right\}.
\end{aligned}$$

Consequently, we can obtain the posterior sample of $\mathbf{C}$ from the following posterior

$$\pi(\mathbf{A}, \mathbf{B} \mid \mathbf{Y}, \hat{\mathbf{d}}_{\text{mode}}) \propto f(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}) \pi(\mathbf{A}, \mathbf{B} \mid \hat{\mathbf{d}}_{\text{mode}}).$$

Note that $\hat{\mathbf{d}}_{\text{mode}}$ can be obtained by the proposed ICM algorithm in the previous section. First, to generate MCMC samples from the above posterior distribution of $(\mathbf{A}, \mathbf{B})$, we consider a Gibbs sampler that iterates through the following steps:

(1) update $\mathbf{a}_j$ for $j = 1, \ldots, p$;

(2) update $\mathbf{b}_l$ for $l = 1, \ldots, q$.

The explicit forms of full conditionals of $\mathbf{a}_j$ and $\mathbf{b}_l$, respectively, are given in (5.19) and (5.21). In each Gibbs step, we update $\mathbf{a}_j$ and $\mathbf{b}_l$ by generating samples from

$$\begin{aligned}
\mathbf{a}_j \mid \mathbf{Y}, \mathbf{A}_{(j)}, \mathbf{B}, \hat{\mathbf{d}}_{\text{mode}} &\sim \mathrm{N}_r\left(\boldsymbol{\mu}_{\mathbf{A},j}, \boldsymbol{\Sigma}_{\mathbf{A},j}\right), \\
\boldsymbol{\mu}_{\mathbf{A},j} &= \boldsymbol{\Sigma}_{\mathbf{A},j}\mathbf{B}^{\mathbf{T}}\left(\mathbf{Y} - \mathbf{X}_{(\tilde{j})}\mathbf{A}_{(j)}\mathbf{B}^{\mathbf{T}}\right)^{\mathbf{T}}\tilde{\mathbf{x}}_j, \\
\boldsymbol{\Sigma}_{\mathbf{A},j} &= \left(\mathbf{B}^{\mathbf{T}}\mathbf{B}\tilde{\mathbf{x}}_j^{\mathbf{T}}\tilde{\mathbf{x}}_j + \hat{\mathbf{D}}_1 + \hat{d}_{2,j}\mathbf{I}_r\right)^{-1} \quad \text{for } j = 1, \ldots, p, \\
\mathbf{b}_l \mid \mathbf{Y}, \mathbf{A}, \hat{\mathbf{d}}_{\text{mode}} &\sim \mathrm{N}_r\left(\boldsymbol{\mu}_{\mathbf{B},l}, \boldsymbol{\Sigma}_{\mathbf{B},l}\right), \\
\boldsymbol{\mu}_{\mathbf{B},l} &= \boldsymbol{\Sigma}_{\mathbf{B},l}(\mathbf{X}\mathbf{A})^{\mathbf{T}}\tilde{\mathbf{y}}_l, \\
\boldsymbol{\Sigma}_{\mathbf{B},l} &= \left((\mathbf{X}\mathbf{A})^{\mathbf{T}}\mathbf{X}\mathbf{A} + \hat{\mathbf{D}}_1 + \hat{d}_{3,j}\mathbf{I}_r\right)^{-1} \quad \text{for } l = 1, \ldots, q,
\end{aligned}$$

where $\hat{\mathbf{d}}_{\text{mode}} = (\mathbf{diag}(\hat{\mathbf{D}}_1), \mathbf{diag}(\hat{\mathbf{D}}_2), \mathbf{diag}(\hat{\mathbf{D}}_3))$. Let $\{\mathbf{A}^i, \mathbf{B}^i\}_{i=1}^N$ be a set of obtained MCMC samples from the above sampling procedure. Then a set of posterior samples for $\mathbf{C}$ can be obtained by $\mathcal{S} = \left\{\mathbf{C}^i : \mathbf{C}^i = \mathbf{A}^i\left(\mathbf{B}^i\right)^{\mathbf{T}}\right\}_{i=1}^N$.

5.4.4 Tuning parameter selection

In practice, we are usually interested in selecting a tuning parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ (or hyperparameter) from the set of candidates $\mathcal{L} = \{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_K\}$. We assume that there is no preferred model, i.e., $\pi(\boldsymbol{\lambda}_k) = 1/K$ for $k = 1, \ldots, K$. In general, the tuning parameter is determined via a grid search strategy from a lower bound $\boldsymbol{\lambda}_L = (0^+, 0^+, 0^+)$ to a given upper bound $\boldsymbol{\lambda}_U$ which is the smallest value to induce the marginal null model (i.e., all estimates are zero). Hence, the set $\mathcal{L}$ is well-defined.

Let $m(\mathbf{Y} \mid \boldsymbol{\lambda}_k) = \int f(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}) \Pi(d\mathbf{A}, d\mathbf{B} \mid \boldsymbol{\lambda}_k)$ be a marginal likelihood for a given $\boldsymbol{\lambda}_k$. Then, we can show that $m(\mathbf{Y} \mid \boldsymbol{\lambda}_k)$ is proportional to the posterior probability of $\boldsymbol{\lambda}_k$ given $\mathbf{Y}$, that is

$$\Pi(\boldsymbol{\lambda}_k \mid \mathbf{Y}) = \frac{m(\mathbf{Y} \mid \boldsymbol{\lambda}_k)\pi(\boldsymbol{\lambda}_k)}{\sum_{k=1}^{K} m(\mathbf{Y} \mid \boldsymbol{\lambda}_k)\pi(\boldsymbol{\lambda}_k)} \propto m(\mathbf{Y} \mid \boldsymbol{\lambda}_k). \tag{5.25}$$

From the above viewpoint, we define the optimal $\boldsymbol{\lambda}_{k^*}$ such that

$$\boldsymbol{\lambda}_{k^*} = \arg\max_{\boldsymbol{\lambda}_k \in \mathcal{L}} \{m(\mathbf{Y} \mid \boldsymbol{\lambda}_k)\}.$$

Let $\mathbf{C}$ be the $p \times q$ coefficient matrix with $\text{rank}(\mathbf{C}) = r^*$. Without loss of generality, suppose that the first p^* rows and q^* columns of $\mathbf{C}$ are non-zero and the remaining rows and columns of $\mathbf{C}$ are zero. Then, the matrix $\mathbf{C}$ can be decomposed as

$$\mathbf{C} = \left[\begin{array}{c} \mathbf{I}_{r^*} \\ \hline \mathbf{C}_A \\ \hline \mathbf{O}_A \end{array}\right] \left[\begin{array}{c|c} \mathbf{C}_B & \mathbf{O}_B \end{array}\right],$$

where $\mathbf{I}_{r^*}$ denotes the identity matrix of order r^*, $\mathbf{C}_A$ is a $(p^* - r^*) \times r^*$ nonzero matrix, $\mathbf{C}_B$ is a $r^* \times q^*$ nonzero matrix, and $\mathbf{O}_A$ is the $(p - p^*) \times r^*$ zero matrix, and $\mathbf{O}_B$ is the $r^* \times (q - q^*)$ zero matrix. The key to above parameterization of $\mathbf{C}$ is that the matrix $\mathbf{C}_A$ and $\mathbf{C}_B$ are uniquely determined. It can be seen that for given (r^*, p^*, q^*), the number of free parameters in $\mathbf{C}$ is $\dim(\mathbf{C}_A) + \dim(\mathbf{C}_B) = r^*(p^* + q^* - r^*)$. Suppose that a given tuning parameter $\boldsymbol{\lambda} = \boldsymbol{\lambda}_k$ results in an estimator with (r_k, p_k, q_k). Define $\boldsymbol{\theta} = [vec(\mathbf{C}_A)^{\mathbf{T}}, vec(\mathbf{C}_B)^{\mathbf{T}}]^{\mathbf{T}}$ such that $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{r_k(p_k + q_k - r_k)}$, where $vec(\cdot)$ denotes

the vectorization of a matrix. Then, the marginal likelihood can be rewritten as

$$\begin{aligned} m(\mathbf{Y} \mid \boldsymbol{\lambda}_k) &= \int_{\Theta} f(\mathbf{Y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}_k) d\boldsymbol{\theta} \\ &= \int_{\Theta} \exp\left\{(nq_k) g_n(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\lambda}_k)\right\} d\boldsymbol{\theta}, \end{aligned} \tag{5.26}$$

where $g_n(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\lambda}) = (nq_k)^{-1} \{\log f(\mathbf{Y} \mid \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}_k)\}$. Let $\hat{\boldsymbol{\theta}}$ be the mode of $g_n(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\lambda})$. By the Laplace approximation, the marginal likelihood in (5.26) can be expressed as

$$m(\mathbf{Y} \mid \boldsymbol{\lambda}) = \frac{(2\pi/nq_k)^{\frac{r_k(p_k+q_k-r_k)}{2}}}{\left|-G_n(\hat{\boldsymbol{\theta}})\right|^{1/2}} \exp\left\{(nq_k) g_n(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}, \boldsymbol{\lambda}_k)\right\} \left\{1 + O_p(n^{-1})\right\}, \tag{5.27}$$

where

$$G_n(\hat{\boldsymbol{\theta}}) = \left. \frac{\partial^2 g_n(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\lambda})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathbf{T}}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

By taking the logarithm of the formula (5.27) and ignoring the term of $O(1)$ and higher order terms, we have the following approximation of log marginal likelihood

$$\log\{m(\mathbf{Y} \mid \boldsymbol{\lambda}_k)\} \approx \log f(\mathbf{Y} \mid \hat{\boldsymbol{\theta}}) - \frac{r_k(p_k + q_k - r_k)}{2} \log(nq_k). \tag{5.28}$$

Let $\hat{\mathbf{C}}_k$ be the BRRR estimate for given $\boldsymbol{\lambda}_k$. Then, by substituting it in (5.28) and multiplying by -2, (5.28) reduces to the BRRR version of Bayesian information criterion (Schwarz, 1978),

$$\mathrm{BIC}(\boldsymbol{\lambda}_k) = -2 \log f(\mathbf{Y} \mid \hat{\mathbf{C}}_k) + r_k(p_k + q_k - r_k) \log(nq_k). \tag{5.29}$$

According (5.25), we know that minimizing the BIC corresponds to maximizing the posterior probability of $\boldsymbol{\lambda}_k$ given $\mathbf{Y}$. Hence, we regard the tuning parameter $\boldsymbol{\lambda}_*$ as the optimum if $\boldsymbol{\lambda}_* = \arg\min_{\boldsymbol{\lambda}_k \in \mathcal{L}} \mathrm{BIC}(\boldsymbol{\lambda}_k)$.

5.5 Posterior consistency

In Bayesian analysis, the posterior consistency assures that the posterior converges to point mass at the true parameter as more data are collected (Diaconis and Freedman, 1986; Ghosh et al., 2006; Choi and Ramamoorthi, 2008). Here, we discuss the posterior consistency for the proposed BSRR method, following Armagan et al. (2013). We allow the number of predictors p to grow with sample size n, and the number of true non-zero coefficients p^* is assumed to be finite. Henceforth, we denote p as p_n. Similarly the response matrix $\mathbf{Y}$ and predictor matrix $\mathbf{X}$ are denoted by $\mathbf{Y}_n$ and $\mathbf{X}_n$, respectively. Unlike p_n, the number of response variables q is assumed to be fixed in our analysis.

Suppose that, given $\mathbf{X}_n$ and $\mathbf{C}^*$, $\mathbf{Y}_n$ is generated from

$$\mathbf{Y}_n = \mathbf{X}_n \mathbf{C}^* + \mathbf{E},$$

where $\mathbf{e}_i \overset{\text{i.i.d.}}{\sim} N_q(\mathbf{0}, \boldsymbol{\Sigma})$ with a positive definite matrix $\boldsymbol{\Sigma}$ (assumed to be known) and $\mathbf{C}^*$ is a $(p_n \times q)$ matrix such that $\mathbf{card}\{j : \mathbf{c^*}_j^{\mathbf{T}} \mathbf{c^*}_j \neq 0\} = p^*$, $\mathbf{card}\{l : \tilde{\mathbf{c^*}}_l^{\mathbf{T}} \tilde{\mathbf{c^*}}_l \neq 0\} = q^*$ and $\mathrm{rank}(\mathbf{C}^*) = r^*$. Further, we make the following assumptions.

I. $p_n = o(n)$, but $p^* < \infty$ and $q^* \leq q < \infty$.

II. $0 < S_{\min} < \liminf_{n\to\infty} \frac{S_{n,\min}}{\sqrt{n}} \leq \limsup_{n\to\infty} \frac{S_{n,\max}}{\sqrt{n}} < S_{\max} < \infty$, where $S_{n,\min}$

and $S_{n,\max}$ denote the smallest and the largest singular values of $\mathbf{X}$, respectively.

III. $\sup_{(j,l)}(c^*_{jl}) < \infty$, where c^*_{jl} indicates the $(j,l)^{th}$ element of $\mathbf{C}^*$.

Our main results are presented in Theorems 5.4 and 5.5 below.

Theorem 5.4. *Under assumptions I and II, if the prior* $\Pi(\mathbf{A},\mathbf{B})$ *satisfies the following condition:*

$$\Pi\left\{(\mathbf{A},\mathbf{B}) : \|\mathbf{AB^T} - \mathbf{C}^*\|_F < \frac{\Delta}{n^{\frac{\rho}{2}}}\right\} > \exp(-dn),$$

for all $0 < \Delta < \frac{\epsilon^2 S^2_{\min}}{48 S^2_{\max}}$ *and* $0 < d < \frac{\epsilon S^2_{\min}}{32\tau_{\max}} - \frac{3\Delta S_{\max}}{2\tau_{\min}}$ *and some* $\rho > 0$, *where* $\tau_{\min}$ *and* $\tau_{\max}$ *denote, respectively, the smallest and the largest eigenvalue of* $\mathbf{\Sigma}$, *then the posterior of* ($\mathbf{A}$,$\mathbf{B}$) *induced by the prior* $\Pi(\mathbf{A},\mathbf{B})$ *is strongly consistent, i.e., for any* $\epsilon > 0$,

$$\Pi\left\{(\mathbf{A},\mathbf{B}) : \|\mathbf{C} - \mathbf{C}^*\|_F > \epsilon, \mathbf{C} = \mathbf{AB^T} \mid \mathbf{Y}_n\right\} \to 0 \text{ almost surely},$$

as $n \to \infty$.

Theorem 5.5. *Under assumptions I, II and III, the prior defined in (5.12) yields a strongly consistent posterior if* $\lambda_i = \delta_i n^{\rho/2}\sqrt{p_n}\log n$ *for finite* $\delta_i > 0$, $i = 1, 2, 3$.

In Theorem 5.4, we establish a sufficient condition on a prior distribution in order to achieve posterior consistency. Theorem 5.5 then shows that our BSRR prior in (5.12) satisfies the sufficient condition in Theorem 5.4, and consequently, our BSRR method possesses the desirable posterior consistency property.

5.6 Simulation studies

To examine the performance of our BSRR method, we conduct Monte Carlo experiments under several possible scenarios. For purposes of comparison, we also consider the following two reduced priors:

$$\pi^{RR}(\mathbf{A}, \mathbf{B} \mid \lambda_1, \lambda_2) := \pi^{\mathrm{BSRR}}(\mathbf{A}, \mathbf{B} \mid \boldsymbol{\lambda}, \lambda_3 = 0), \tag{5.30}$$

$$\pi^{RC}(\mathbf{A}, \mathbf{B} \mid \lambda_2, \lambda_3) := \pi^{\mathrm{BSRR}}(\mathbf{A}, \mathbf{B} \mid \boldsymbol{\lambda}, \lambda_1 = 0). \tag{5.31}$$

We denote the Bayesian methods using (5.30) and (5.31) as RR (Row-wise-sparse and Reduced-rank) method and RC (Row-and-Column-wise sparse) method, respectively. Our BSRR method aims to recover all the low-dimensional structures in A1–A3, but RR and RC methods, respectively, do not consider the column-wise sparsity of $\mathbf{C}$ in A3 and the reduced rank structure of $\mathbf{C}$ in A1. Therefore, the RR method is analogous to the joint rank and predictor selection methods proposed by Chen and Huang (2012) and Bunea et al. (2012). The RR and RC methods can be derived from BSRR method with setting $\lambda_3 = 0$ and $\lambda_1 = 0$ in (5.13), respectively. Hence, the BSRR estimate $\hat{\mathbf{C}}_{\mathrm{BSRR}}$ as well as RR and RC estimates, $\hat{\mathbf{C}}_{\mathrm{RR}}$ and $\hat{\mathbf{C}}_{\mathrm{RC}}$, are obtained by the proposed algorithm in Section 5.4.2. Similarly, the unknown tuning parameter $\boldsymbol{\lambda}$ for each model is estimated by the proposed BIC in (5.29).

We generate data from the multivariate regression model $\mathbf{Y} = \mathbf{XC} + \mathbf{E}$. For the $n \times p$ design matrix $\mathbf{X}$, its n rows are independently generated from $N_p(\mathbf{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = \{\Gamma_{ij}\}_{p\times p}$ and $\Gamma_{ij} = (0.5)^{|i-j|}$. The $p \times q$ coefficient matrix $\mathbf{C}$ is defined as $\mathbf{C} = \sum_{k=1}^{r^*} s_k \mathbf{C}_k$, where $s_k = 5 + (k-1)\lceil \frac{15}{r^*} \rceil$; the entries of $\mathbf{C}_k$ are all zero expect in its upper left $p^* \times q^*$ submatrix, which is generated by $\mathbf{z}_1\mathbf{z}_2^{\mathbf{T}}/\|\mathbf{z}_1\|_2/\|\mathbf{z}_2\|_2$, where $\mathbf{z}_1 \in \mathbb{R}^{p^*}$,

$\mathbf{z}_2 \in \mathbb{R}^{q^*}$, and all their entries are i.i.d samples from uniform($[-1,-0.3]\cup[0.3,1]$). The entries of the noise matrix $\mathbf{E}$ are independently generated from $N(0,\sigma^2)$, where σ^2 is chosen according to the signal to noise ratio (SNR) defined by $s_{r^*}(\mathbf{XC})/s_1(\mathbf{P}_X\mathbf{E})$ with $\mathbf{P}_X = \mathbf{X}(\mathbf{X^T X})^{-}\mathbf{X^T}$.

In the first scenario, we generate models of moderate dimensions (i.e., $p, q < n$) in three setups:

(a1) $p = q = 25$, $n = 50$, $r^* = 3$, $p^* = 10$, $q^* = 10$. This setup favors our BSRR method.

(a2) $p = q = 25$, $n = 50$, $r^* = 3$, $p^* = 10$, $q^* = 25$. As all the responses are revelent in the model, this setup favors the RR method.

(a3) $p = q = 25$, $n = 50$, $r^* = 10$, $p^* = 10$, $q^* = 10$. This favors the RC method, which does not enforce rank reduction.

In the second scenario, we generate high-dimensional data (i.e., $p, q > n$) using similar settings as above,

(b1) $p = 200$, $q = 170$, $n = 50$, $r^* = 3$, $p^* = 10$, $q^* = 10$.

(b2) $p = 200$, $q = 170$, $n = 50$, $r^* = 3$, $p^* = 10$, $q^* = 170$.

(b3) $p = 200$, $q = 170$, $n = 50$, $r^* = 10$, $p^* = 10$, $q^* = 10$.

The estimation accuracy is measured by the following three mean squared errors (MSEs):

$$\begin{aligned} \mathrm{MSE}_{\mathrm{est}} &= 100\|\hat{\mathbf{C}} - \mathbf{C}\|_F^2/(pq), \\ \mathrm{MSE}_{\mathrm{pred}} &= 100\|\mathbf{X}\hat{\mathbf{C}} - \mathbf{XC}\|_F^2/(nq), \\ \mathrm{MSE}_{\mathrm{dim}} &= 100\|\mathbf{s}(\hat{\mathbf{C}}) - \mathbf{s}(\mathbf{C})\|^2/\min(p,q), \end{aligned}$$

where $\mathbf{s}(\mathbf{C})$ denotes the vector of singular values for a matrix $\mathbf{C}$. To assess the variable selection performance, we use false positive rate (FPR) and false negative rate (FNR) such that $\text{FPR}\% = 100\text{FP}/(\text{TN} + \text{FP})$ and $\text{FNR}\% = 100\text{FN}/(\text{TP} + \text{FN})$, where TP, FP, TN and FN denote the numbers of true nonzeros, false nonzeros, true zeros and false zeros, respectively. The rank selection performance is evaluated by the percentage of correct rank identification (CRI%). All measurements are estimated by the Monte Carlo method with 500 replications.

Tables 5.1 and 5.2 summarize the simulation results. As expected, in the cases (a1) and (b1), where rank reduction, predictor selection and response selection are all preferable, the BSRR method performs much better than the other two reduced methods. In cases (a2) and (b2), rank reduction and predictor selection are preferable while response selection is not necessary. The performance of the BSRR method is very similar to the RR method which assumes the correct model structure. In the cases (a3) and (b3), rank reduction becomes unnecessary when response and predictor selections are performed. While the BSRR method slightly underestimates the true rank ($r^* = 10$), its variable selection performance (FPR and FNR) is comparable to that of the RC method which assumes the correct model structure. The results are consistent for different SNR levels. Therefore, our BSRR approach provides a flexible and unified way for simultaneously exploring rank reduction, predictor selection and response selection.

TABLE 5.1: Summary of the simulation results for examples (a1)–(a3).

Case	SNR	Method	MSE_{est}	MSE_{pred}	MSE_{dim}	FPR%	FPR%	CRI%	$\widehat{r}$
(a1)	0.50	**BSRR**	10.37	152.13	29.32	0.95	0.48	99.80	3.00
		RR	18.81	283.62	71.93	29.27	0.50	99.60	3.00
		RC	22.80	294.02	288.71	3.18	0.06	0.00	9.84
	0.75	**BSRR**	3.82	55.36	7.60	0.11	0.16	100.00	3.00
		RR	7.00	109.79	16.43	28.60	0.16	99.80	3.00
		RC	7.85	99.55	92.91	0.39	0.04	0.00	9.49
	1.00	**BSRR**	2.05	29.66	3.72	0.03	0.10	100.00	3.00
		RR	3.59	57.63	6.18	28.56	0.10	100.00	3.00
		RC	4.17	52.64	48.88	0.08	0.04	0.00	9.30
(a2)	0.50	BSRR	20.69	312.02	74.81	0.00	1.60	100.00	3.00
		RR	21.51	320.58	80.66	0.00	1.20	100.00	3.00
		RC	52.39	640.71	763.88	0.00	0.00	0.00	10.00
	0.75	BSRR	7.63	122.94	13.76	0.00	1.10	100.00	3.00
		RR	7.75	123.49	15.90	0.00	0.80	100.00	3.00
		RC	22.70	279.66	326.48	0.00	0.00	0.00	10.00
	1.00	BSRR	3.96	65.09	4.94	0.00	0.49	100.00	3.00
		RR	4.02	65.35	5.72	0.00	0.40	100.00	3.00
		RC	12.40	154.14	179.63	0.00	0.01	0.00	10.00
(a3)	0.50	BSRR	0.04	0.52	0.16	0.27	0.00	18.40	9.11
		RR	0.10	1.17	0.15	28.79	0.00	25.00	9.18
		RC	0.04	0.50	0.09	2.45	0.00	97.20	9.99
	0.75	BSRR	0.02	0.24	0.11	0.66	0.00	18.60	9.11
		RR	0.04	0.53	0.10	28.79	0.00	22.20	9.15
		RC	0.02	0.21	0.04	5.13	0.00	98.40	9.98
	1.00	BSRR	0.01	0.15	0.09	1.45	0.00	19.40	9.12
		RR	0.03	0.31	0.08	29.18	0.00	21.00	9.14
		RC	0.01	0.12	0.02	6.49	0.00	98.60	9.99

5.7 Yeast cell cycle data

Transcription factors (TFs), also called sequence-specific DNA binding proteins, regulate the transcription of genes from DNA to mRNA by binding specific DNA sequences. In order to understand the regulatory mechanism, it is important to reveal the network structure between TFs and their target genes. The network structure can be formulated using the multivariate regression model in (5.1), where the row and column of the response matrix, respectively, correspond to genes and samples

TABLE 5.2: Summary of the simulation results for examples (b1)–(b3).

Case	SNR	Method	MSE_{est}	MSE_{pred}	MSE_{dim}	FPR%	FPR%	CRI%	$\widehat{r}$
(b1)	0.50	**BSRR**	0.03	4.00	0.50	0.00	0.02	99.00	3.00
		RR	0.42	54.15	25.00	4.72	0.04	40.00	3.60
		RC	0.07	7.79	7.04	0.01	0.00	0.00	9.32
	0.75	**BSRR**	0.02	1.75	0.24	0.00	0.00	99.60	3.00
		RR	0.18	23.89	10.39	4.72	0.02	29.20	3.71
		RC	0.03	2.92	2.49	0.00	0.00	0.00	8.94
	1.00	**BSRR**	0.01	1.00	0.17	0.00	0.00	99.80	3.00
		RR	0.09	12.45	4.25	4.72	0.00	52.20	3.48
		RC	0.01	1.58	1.24	0.00	0.00	0.00	8.62
(b2)	0.50	BSRR	0.37	55.54	7.92	0.00	0.42	99.60	3.00
		RR	0.38	56.00	8.91	0.00	0.03	99.60	3.00
		RC	1.67	164.79	220.20	0.00	0.01	0.00	10.00
	0.75	BSRR	0.15	23.44	1.27	0.00	0.24	99.80	3.00
		RR	0.15	23.49	1.57	0.00	0.02	100.00	3.00
		RC	0.73	72.09	95.14	0.00	0.02	0.00	10.00
	1.00	BSRR	0.08	13.02	0.35	0.00	0.11	99.80	3.00
		RR	0.08	13.00	0.46	0.00	0.00	99.80	3.00
		RC	0.40	39.83	52.55	0.00	0.02	0.00	10.00
(b3)	0.50	BSRR	0.00	0.02	0.01	0.04	0.00	19.80	9.14
		RR	0.00	0.27	0.02	4.69	0.00	29.40	9.24
		RC	0.00	0.02	0.00	0.00	0.00	98.00	9.98
	0.75	BSRR	0.00	0.01	0.01	0.06	0.00	20.00	9.14
		RR	0.00	0.12	0.01	4.68	0.00	25.20	9.19
		RC	0.00	0.01	0.00	0.00	0.00	98.20	9.98
	1.00	BSRR	0.00	0.01	0.01	0.04	0.00	20.00	9.14
		RR	0.00	0.07	0.01	4.67	0.00	23.60	9.18
		RC	0.00	0.00	0.00	0.00	0.00	98.60	9.99

(arrays, tissue types, time points), and the design matrix includes the binding information representing the strength of interaction between TFs and the target genes. The regression coefficient matrix then describes actual transcription factor activities of TFs for genes. In practice, many TFs are not actually related to the genes and there exists dependency among the samples due to the design of experiment.

Here, we analyze an *Yeast cell cycle* data (Chun and Keles, 2010) using BSRR. The dataset is available in the **spls** package in **R**. The response matrix $\mathbf{Y}$ consists of 542 cell-cycle-regulated genes from an α factor arrest method, where mRNA levels

are measured at every 7 minutes during 119 minutes, i.e., $n = 542$ and $q = 18$. The 542×106 predictor matrix $\mathbf{X}$ contains the binding information of the target genes for a total of 106 TFs, where Chromatin immunoprecipitation (ChIP) for the 542 genes was performed on each of these 106 TFs. In our analyses, $\mathbf{Y}$ and $\mathbf{X}$ are centered.

We apply the BSRR method to the dataset. We use the proposed BIC to choose the tuning parameter and obtain $\hat{\lambda}_1 = 5$, $\hat{\lambda}_2 = 1.5$ and $\hat{\lambda}_3 = 1.5$. As a result, 26 TFs are identified at 17 time points (105 min is eliminated) with the estimated rank $\widehat{r} = 4$. Figure 5.2 displays the obtained parameter estimates and 95% credible bands for randomly selected 4 TFs among the 26 TFs. The same data set was also analyzed by the adaptive SRRR method of Chen and Huang (2012). In the adaptive SRRR, 32 TFs were identified at 18 time points with the optimal rank $\widehat{r} = 4$ determined by a cross validation method. Among their selected 32 TFs, 21 TFs were also identified by our BSRR method. To compare variable selection performance between two methods, we define the following two models:

$$M^1 : \quad \mathbf{Y} = \mathbf{X}_1\mathbf{C}_1 + \mathbf{E}; \tag{5.32}$$

$$M^2 : \quad \mathbf{Y} = \mathbf{X}_2\mathbf{C}_2 + \mathbf{E}; \tag{5.33}$$

where $\mathbf{X}_1$ contains the information of the 542 genes for the 32 TFs identified by the adaptive SRRR, $\mathbf{X}_2$ contains the information of the 542 genes for the 26 TFs identified by BSRR, $\mathbf{C}_1$ is the 32×18 matrix with $\text{rank}(\mathbf{C}_1) = 4$, and $\mathbf{C}_2$ is the 26×18 matrix with $\text{rank}(\mathbf{C}_2) = 4$. To conduct a fair comparison, we consider the following reduced

rank priors in Geweke (1996) for the models M^1 and M^2, respectively:

$$\pi^1(\mathbf{A}_1, \mathbf{B}_1) \propto \exp\left\{-\frac{\tau}{2}\left(\|\mathbf{A}_1\|_F^2 + \|\mathbf{B}_1\|_F^2\right)\right\} \quad \text{s.t.} \quad \mathbf{C}_1 = \mathbf{A}_1\mathbf{B}_1^{\mathbf{T}}, \tag{5.34}$$

$$\pi^2(\mathbf{A}_2, \mathbf{B}_2) \propto \exp\left\{-\frac{\tau}{2}\left(\|\mathbf{A}_2\|_F^2 + \|\mathbf{B}_2\|_F^2\right)\right\} \quad \text{s.t.} \quad \mathbf{C}_2 = \mathbf{A}_2\mathbf{B}_2^{\mathbf{T}}, \tag{5.35}$$

where $\mathbf{A}_1$ is a 32×4 matrix, $\mathbf{B}_1$ is an 18×4 matrix, $\mathbf{A}_2$ is an 26×4 matrix, $\mathbf{B}_2$ is a 18×4 matrix and we set $\tau = 0.0001$ to be a non-informative (flat) prior, so that the parameter estimates are determined nearly by the observations $(\mathbf{Y}_1, \mathbf{X}_1)$ and $(\mathbf{Y}_2, \mathbf{X}_2)$. As the Bayesian model selection criterion, using the priors in (5.34) and (5.35), we utilize the deviance information criteria (DIC) defined by

$$\begin{aligned}
\mathrm{DIC}_1 &= -4E_{\mathbf{A}_1,\mathbf{B}_1|\mathbf{Y},\mathbf{X}_1}\left[\log\left\{f\left(\mathbf{Y} \mid \mathbf{X}_1, \mathbf{C}_1 = \mathbf{A}_1\mathbf{B}_1^{\mathbf{T}}\right)\right\}\right] \\
&\quad +2\log\left\{f\left(\mathbf{Y} \mid \mathbf{X}_1, \mathbf{C}_1 = \overline{\mathbf{A}_1\mathbf{B}_1^{\mathbf{T}}}\right)\right\}, \\
\mathrm{DIC}_2 &= -4E_{\mathbf{A}_2,\mathbf{B}_2|\mathbf{Y},\mathbf{X}_2}\left[\log\left\{f\left(\mathbf{Y} \mid \mathbf{X}_2, \mathbf{C}_2 = \mathbf{A}_2\mathbf{B}_2^{\mathbf{T}}\right)\right\}\right] \\
&\quad +2\log\left\{f\left(\mathbf{Y} \mid \mathbf{X}_2, \mathbf{C}_2 = \overline{\mathbf{A}_2\mathbf{B}_2^{\mathbf{T}}}\right)\right\},
\end{aligned}$$

where $\overline{\mathbf{AB}^{\mathbf{T}}}$ denotes the posterior mean. If $\mathrm{DIC}_1 > \mathrm{DIC}_2$, then it implies that the model M^2 is more strongly supported by the given data than the model M^1. Let $\{\mathbf{A}_1^i, \mathbf{B}_1^i\}_{i=1}^N$ and $\{\mathbf{A}_2^i, \mathbf{B}_2^i\}_{i=1}^N$ be MCMC samples from the posteriors $\pi^1(\mathbf{A}_1, \mathbf{B}_1 \mid \mathbf{Y}, \mathbf{X}_1)$ and $\pi^2(\mathbf{A}_2, \mathbf{B}_2 \mid \mathbf{Y}, \mathbf{X}_2)$, respectively. Note that the MCMC samples can be easily generated from multivariate normal distributions by using the Gibbs sampler. Define $\left\{\mathbf{C}_m^i : \mathbf{C}_m^i = \mathbf{A}_m^i(\mathbf{B}_m^i)^{\mathbf{T}}\right\}_{i=1}^N$, for $m = 1, 2$. Then the DIC can be estimated by

the following Monte Carlo estimator:

$$\begin{aligned}
\widehat{\mathrm{DIC}}_1 &= -4\left[\frac{1}{N}\sum_{i=1}^{N}\log f\left(\mathbf{Y} \mid \mathbf{X}_1, \mathbf{C}_1^i\right)\right] + 2\log f\left(\mathbf{Y} \mid \mathbf{X}_1, \frac{1}{N}\sum_{i=1}^{N}\mathbf{C}_1^i\right), \\
\widehat{\mathrm{DIC}}_2 &= -4\left[\frac{1}{N}\sum_{i=1}^{N}\log f\left(\mathbf{Y} \mid \mathbf{X}_2, \mathbf{C}_2^i\right)\right] + 2\log f\left(\mathbf{Y} \mid \mathbf{X}_2, \frac{1}{N}\sum_{i=1}^{N}\mathbf{C}_2^i\right).
\end{aligned}$$

Based on 1,000 MCMC samples (after 1,000 burn-in iterations) with 100 replication, we obtain $\widehat{\mathrm{DIC}}_1 = 19824.46$ and $\widehat{\mathrm{DIC}}_2 = 19784.03$ with Monte Carlo errors 1.29 and 1.14, respectively. Since $\widehat{\mathrm{DIC}}_1 > \widehat{\mathrm{DIC}}_2$, this result supports the model M^2. Consequently, this implies that our BSRR method has better variable selection performance than the adaptive SRRR for Yeast cell cycle data. Recall that the response at 105 min was eliminated in the BSRR method. Table 5.3 displays the parameter estimates and 95% credible intervals (CIs) at 105 min from the model M^2. Since all CIs include zero, this demonstrates that the response elimination at 105 min in the BSRR is valid. In other words, none of TFs activates at 105 min.

5.8 Discussion

We have developed a Bayesian sparse and low rank regression method, which achieves simultaneous rank reduction and predictor/response selection. There are many directions for future research. We have mainly focused on the ℓ_0 type sparsity-inducing penalties to construct prior distribution. The method can be extended to use other forms of penalties for inducing diverse lower-dimensional structures. The low-rank structure induces dependency among the response variables, and hence the error correlation structure is not explicitly considered in the current work. Incorporating the

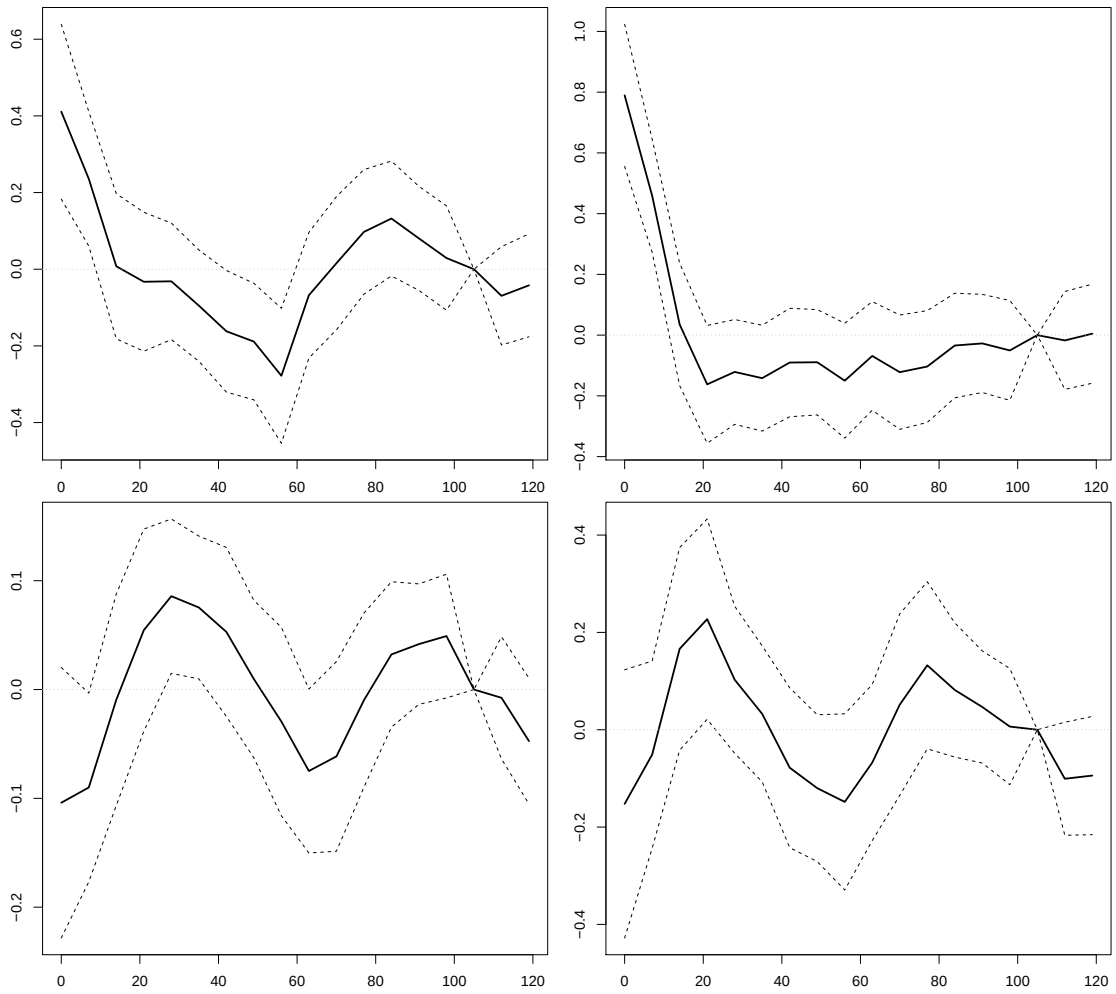

FIGURE 5.2: The parameter estimates and 95% credible bands for randomly selected 4 TFs (ACE2, STE12, SWI4, ZAP1)from the BSRR, where x-axis indicates time (min) and y-axis indicates estimated coefficients.

variance component into our model might improve the efficiency of the coefficient estimation. In a Bayesian framework, this can be accomplished by assigning an appropriate prior on the variance component; the choice of the prior should be carefully treated due to the lack of unimodality of the posterior (Park and Casella, 2008). In practice, the response variables could be binary or counts. It is thus pressing to utilize a general likelihood function with the proposed BSRR prior. The proposed ICM

TABLE 5.3: Parameter estimates (Est) with 95% credible intervals (CIs) at 105 min from the model M^2.

TF	Est	CIs	TF	Est	CIs
ACE2	-0.03	(-0.16 , 0.10)	RME1	0.05	(-0.08 , 0.17)
ARG81	-0.05	(-0.19 , 0.09)	RTG3	0.01	(-0.09 , 0.11)
FKH2	0.06	(-0.01 , 0.13)	SFP1	-0.05	(-0.15 , 0.05)
HIR1	0.10	(-0.06 , 0.26)	SOK2	-0.03	(-0.07 , 0.02)
HIR2	0.07	(-0.06 , 0.21)	STB1	0.01	(-0.05 , 0.06)
IME4	-0.02	(-0.13 , 0.08)	STE12	-0.04	(-0.21 , 0.12)
MBP1	-0.04	(-0.13 , 0.05)	SWI4	0.02	(-0.03 , 0.08)
MCM1	-0.03	(-0.11 , 0.04)	SWI5	-0.06	(-0.21 , 0.09)
MET4	0.06	(-0.02 , 0.13)	SWI6	-0.01	(-0.08 , 0.07)
NDD1	0.05	(-0.08 , 0.17)	YAP7	0.03	(-0.04 , 0.10)
NRG1	0.02	(-0.06 , 0.11)	YFL044C	0.04	(-0.06 , 0.15)
PHD1	-0.02	(-0.09 , 0.04)	YJL206C	0.08	(-0.05 , 0.22)
REB1	0.03	(-0.03 , 0.09)	ZAP1	-0.03	(-0.15 , 0.09)

algorithm converges relatively fast, and in each iteration the main cost is to inverse a matrix of dimension $\min(n, p)$ owning to Woodbury matrix identity. However, this approach would still be inefficient when both p and n are extremely large. One way is to conduct some pre-screening procedure (Fan and Lv, 2008; Fan and Song, 2010) before implementation of the proposed method. It would also be interesting to study online learning and the divide-and-conquer strategies of the proposed model. We have established the posterior consistency of the proposed sparse and low-rank estimation method under a high-dimensional asymptotic regime, which characterizes the behavior of the posterior distribution when the number of predictors p_n increases with the sample size n. The theoretical analysis of the Bayesian (point) estimator itself could be of interest rather than the entire posterior distribution (Alquier, 2013).

5.9 Extensions

In practice, it is possible that the response variables are binary, counted, or even mixed. Furthermore, some observations would be unobserved, i.e., some y_{il}'s are missing. The proposed BSRR, however, is applicable when all responses are continuous and fully observed. To overcome these limitations, as an extension of BSRR, we introduce a new approach, called Generalized Bayesian Sparse and Reduced-rank Regression (GBSRR).

Consider a parametric model,

$$E(\mathbf{Y}) = G(\mathbf{M} + \mathbf{XC}), \tag{5.36}$$

where $\mathbf{M} = \mathbf{1}_n(\mu_1, \ldots, \mu_q)$ indicates the intercept matrix and $G(\mathbf{C}) = [g_{il}(c_{il})]_{n\times q}$ denotes a link function matrix with the known univariate link function $g_{il}(\cdot)$ for $i = 1, \ldots, n$ and $l = 1, \ldots, q$. For the coefficient matrix $\mathbf{C}$ in model (5.36), our underlying assumptions A1-A3 in Section 5.1 are still conditioned.

To develop the likelihood function for (5.36), we propose to use the Bregman divergence as the negative likelihood function. To address the challenge of missing data, we introduce an auxiliary variable $\mathbf{\Delta} = [\Delta_{il}]_{n\times q}$ as follows. Let Δ_{il} be the observing indicator for y_{il}, that is $\Delta_{il} = 1$ if y_{il} is observed and $\Delta_{il} = 0$, otherwise. Then, we define the likelihood by

$$f(\mathbf{Y} \mid \mathbf{C}, \boldsymbol{\mu}) \;\propto\; \exp\left[-\sum_{i=1}^{n}\sum_{l=1}^{q} \Delta_{il}\mathrm{BD}_{\psi}\left\{y_{il}, g_{il}(\mu_l + \mathbf{x}_i^{\mathbf{T}}\tilde{\mathbf{c}}_l)\right\}\right],$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_q)^{\mathbf{T}}$.

Now, we introduce the GBSRR,

$$
\begin{aligned}
f(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}, \mathbf{m}) &\propto \exp\left(-\mathrm{BD}_{\psi}^{\Delta}\{\mathbf{Y}, G(\mathbf{M} + \mathbf{X}\mathbf{A}\mathbf{B}^{\mathbf{T}})\}\right), \\
\pi(\mathbf{A}, \mathbf{B} \mid \mathbf{d}) &\propto \exp\left\{-\frac{1}{2}\left(\|\mathbf{A}\mathbf{D}_1^{1/2}\|_F^2 + \|\mathbf{D}_2^{1/2}\mathbf{A}\|_F^2\right)\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}\left(\|\mathbf{B}\mathbf{D}_1^{1/2}\|_F^2 + \|\mathbf{D}_3^{1/2}\mathbf{B}\|_F^2\right)\right\}, \\
\mathbf{D}_i^{1/2} &= \mathbf{diag}(\sqrt{d_{i,1}}, \ldots, \sqrt{d_{i,m_i}}),\ i = 1, 2, 3, \\
\pi(\mathbf{d} \mid \boldsymbol{\lambda}) &\propto \prod_{i=1}^{3}\left[\prod_{j=1}^{m_i}\left\{(d_{i,j})^{\frac{\lambda_i}{2}} \exp\left(-\frac{\omega_0}{2} d_{i,j}\right)\right\}\right], \\
\pi(\boldsymbol{\mu} \mid \sigma_m^2) &\propto \exp\left(-\frac{1}{2\sigma_m^2}\boldsymbol{\mu}^{\mathbf{T}}\boldsymbol{\mu}\right),
\end{aligned}
$$

where $m_1 = r,\ m_2 = p,\ m_3 = q,\ \mathbf{d} = (\mathbf{diag}(\mathbf{D}_1), \mathbf{diag}(\mathbf{D}_2), \mathbf{diag}(\mathbf{D}_3)),\ \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3),\ d_{i,m_i} > 0\ (i = 1, 2, 3)$, and

$$
\mathrm{BD}_{\psi}^{\Delta}\{\mathbf{Y}, G(\mathbf{M} + \mathbf{X}\mathbf{A}\mathbf{B}^{\mathbf{T}})\} = \sum_{i=1}^{n}\sum_{l=1}^{q} \Delta_{il} \mathrm{BD}_{\psi}\left\{y_{il}, g_{il}(m_l + \mathbf{x}_i^{\mathbf{T}}\mathbf{A}\mathbf{b}_l)\right\}.
$$

Note that using local quadratic approximation (Fan and Li, 2001), the MAP estimation can be easily implemented by the ICM algorithm as in the BSRR, but we will not discuss the details in this chapter.

Chapter 6

Sparse Functional Estimation of Regression Coefficients using Bregman Clustering

6.1 Introduction

In living organisms, a *transcription factor* is a special protein that binds to DNA and controls which genes are turned on or off. Owing to the action of transcription factor, different genes are expressed in different cells and thus various cells can function differently, while they contain exactly the same DNA. For instance, genomes in all the cells are identical, but the gene expression levels in lung cells are different than that in skin cells because the transcription was controlled by the DNA-binding proteins.

In biological research, a study of relationship between transcription factors and their target genes during a biological process has been a subject of intensive investigation in order to reveal the transcriptional regulatory networks (Spellman et al., 1998;

Lee et al., 2002; Boulesteix and Strimmer, 2005). Understanding the transcriptional regulatory mechanisms, however, involves the following difficulties:

1. It is hard to directly measure the actual activity of transcription factors due to the lack of technology (Boulesteix and Strimmer, 2005).

2. A small number of transcription factors, that are significantly related to a given biological process, should be identified from a large number of candidates, known as *sparse high-dimensional problem*. In addition, the transcription factors could be highly correlated, called *multicollinearity* (Chun and Keles, 2010).

3. Since a biological process is dynamic, it requires a time-course investigation into the temporal behavior of transcription factors during the biological process rather than at a single time point (Luan and Li, 2003; Wang et al., 2007).

In an attempt to overcome the first challenge, an integrative analysis of gene expression data and Chromatin Immunoprecipitation (ChIP) data has played a key role, where the ChIP data represent the connectivity information (or the strength of binding interaction) between transcription factors and their target genes (Liao et al., 2003; Li and Chan, 2004). Gao et al. (2004), for example, utilized a multivariate linear regression model along with backward variable elimination. In the regression model, gene expression data and ChIP data were considered as the response variable and the predictor variable, respectively, so that the hidden transcription factor activities were captured by the coefficient estimate of selected predictors from the backward variable selection. However, due to the high-dimensionality and multicollinearity, this method leads to inaccurate results (Fan and Lv, 2010). To tackle this problem, Chun and Keles (2010) introduced a sparse partial least squares (PLS) regression that simultaneously achieves the estimation and the variable selection by inducing sparsity

(or many zero values) on the high-dimensional coefficient matrix. Consequently, both high-dimensionality and multicollinearity are remedied through the sparse PLS. While the sparse PLS works reasonably well at the observed time points, it cannot explain the dynamic transcription factor activities at unobserved time points. Meanwhile, Wang et al. (2007) suggested to apply the group SCAD penalty (Fan and Li, 2001) to the functional response model with time-varying coefficients to capture the dynamic biological process. In the SCAD penalized estimation procedure, a time-varying coefficient function reduces to a spline function spanned by natural cubic B-spline basis if the corresponding predictor is considered as the relevant variable. Otherwise, it becomes the zero function, i.e., it takes the value zero at any time points. Consequently, this approach allows understanding the entire activity patterns of identified transcription factors over time.

It is common for gene expression data to present heterogeneity in gene expression profiles, possibly due to unobserved genetic and environmental factors. In fact, the presence of heterogeneity for gene expression data has been discussed in many biology or bioinformatics literatures (Spellman et al., 1998; Yeung et al., 2001; Luan and Li, 2003). Nevertheless, none of the previous studies (Gao et al., 2004; Boulesteix and Strimmer, 2005; Wang et al., 2007; Chun and Keles, 2010) has taken this important natural characteristic of gene expression data into consideration for the inference procedure of transcriptional regulatory mechanisms. However, the ignored heterogeneity can lead to mis-identification of relevant transcription factors and inaccurate estimation of the transcription factor activities. This aspect motivates us to develop a new method that handles the heterogeneity as well as the aforementioned three challenges for reconstructing the transcriptional regulatory networks.

In this chapter, we propose a novel functional coefficient estimation for a finite

mixture functional regression model from a Bayesian perspective. In the proposed method, for each gene, we consider an observed gene expression level as a realization of functional response from a general finite mixture of the functional response models with time-varying coefficients, so that both functional clustering and functional coefficient estimation are accomplished together. In order to develop a generalized estimation procedure of functions, we assume that all unknown coefficient functions including intercept functions belong to a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950; Wahba, 1990). To handle various types of data, we construct the distance-based-likelihood using a general class of divergence measures (or loss functions), called *Bregman divergence*.

To simultaneously estimate the unknown coefficient functions and eliminate irrelevant predictors, we introduce a well-behaved sparsity-inducing prior, proposed by (Goh et al., 2014), that closely mimics a ℓ_0-norm penalty under maximum a posteriori (MAP) estimation. In the previous studies, if the measurement values were missing at any time point of gene expression data, then the corresponding gene was entirely excluded from the analysis. Unlike the existing methods, we take full-account of all genes using a simple trick, regardless of the missing values. Therefore, our proposed method enables us to use the data more efficiently.

Some remarks are due on the notation used throughout this chapter. For a generic matrix $\mathbf{A}$, we use $\mathbf{A}_{[i,\ ]}$, $\mathbf{A}_{[\ ,j]}$ and $\mathbf{A}_{[i,j]}$ to denote the i^{th} row, the j^{th} column and the $(i,j)^{th}$ element of $\mathbf{A}$, respectively, where $\mathbf{A}_{[i,\ ]}$ and $\mathbf{A}_{[\ ,j]}$ are defined in a form of column vector.

6.2 Model setup

For the i^{th} gene, we observe the expression level $y_i(t)$ at time $t(> 0)$ with the time invariant covariate vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathbf{T}}$ that contains binding information of p transcription factors on the target gene, that is x_{ij} representing the strength of binding interaction between the j^{th} transcription factor and the i^{th} gene, for $i = 1, \ldots, n$. To take the hidden heterogeneity into account, we assume that each gene belongs to one of the gene clusters, where the number of clusters $(= k^*)$ is fixed, but unknown. Define the cluster indicator for the i^{th} gene as $z_i \in \{1, \ldots, k^*\}$, i.e., $z_i = k$ if the i^{th} gene belongs to the k^{th} cluster. To formulate the transcriptional regulatory networks, we consider the following functional response regression model: for $i = 1, \ldots, n$,

$$y_i(t) = \mu_{z_i}(t) + \mathbf{x}_i^{\mathbf{T}}\mathbf{b}(t) + \epsilon_i(t), \tag{6.1}$$

where $\mathbf{b}(t) = (b_1(t), \cdots, b_p(t))^{\mathbf{T}}$ is the p-dimensional vector of unknown coefficient functions that presents the actual transcription factor activities at time t, $\epsilon_i(t)$ denotes a realization of a zero-mean stochastic process, and $\mu_k(t)$ is the underlying shape function of the k^{th} cluster, describing how the mean of gene expression level changes over time without the effect of transcriptional factors within the k^{th} cluster.

In (6.1), we assume that all unknown coefficient functions and shape functions belong to the RKHS. Then, a function in the RKHS can be expressed as an infinite linear combination of reproducing kernels. To formulate, let $\kappa(\cdot, \cdot)$ denote a kernel, e.g., $\kappa(t, s) = \exp\{-\tau(t - s)^2\}$, called a "Gaussian kernel", and $\kappa(t, s) = \exp(ts + 1)^{\tau}$, called a "polynomial kernel". Suppose that $g(\cdot)$ is a function in the RKHS. Then, the function g can be expressed as $g(\cdot) = \sum_{j=1}^{\infty} \alpha_j \kappa(\cdot, t_j)$, where $\alpha_1, \alpha_2, \ldots \in \mathbb{R}$; see

Aronszajn (1950); Wahba (1990) for more details on RKHS. Suppose that, for each gene, the measurements are recorded at $t = t_1, t_2, \ldots, t_q$. Then, from the representer theorem (Wahba, 1990), each of the functions admits a representation of the form: $\mu_k(t) = \sum_{l=1}^{q} \alpha_{kl}\kappa(t, t_l)$ and $b_j(t) = \sum_{l=1}^{q} \beta_{jl}\kappa(t, t_l)$ for $j = 1, \ldots, p$ and $k = 1, \ldots, k^*$. Consequently, owing to the representer theorem, the infinite dimensional parameter space reduces to the finite dimensional space. For notational convenience, we define $\mathbf{y}_i = (y_{i1}, \ldots y_{iq})^{\mathbf{T}} = (y_i(t_1), \ldots y_i(t_q))^{\mathbf{T}}$, $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{kq})^{\mathbf{T}}$, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jq})^{\mathbf{T}}$ and $\mathbf{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p]^{\mathbf{T}}$ for $i = 1, \ldots, n$, $j = 1, \ldots, p$ and $k = 1, \ldots, k^*$. Let $\mathbf{K}$ denote the Gram matrix of $\kappa(\cdot, \cdot)$ with respect to $t_1, \ldots, t_q$, i.e., $\mathbf{K}_{[i,j]} = \kappa(t_i, t_j)$. Then, from (6.1), the conditional expectation of $\mathbf{y}_i$ is

$$E(\mathbf{y}_i \mid \boldsymbol{\alpha}_{z_i}, \mathbf{B}) = \mathbf{K}\boldsymbol{\alpha}_{z_i} + \mathbf{K}\mathbf{B}^{\mathbf{T}}\mathbf{x}_i = \mathbf{c}_{z_i} + \mathbf{D}^{\mathbf{T}}\mathbf{x}_i, \tag{6.2}$$

for $i = 1, \ldots, n$, where $\mathbf{c}_k = \mathbf{K}\boldsymbol{\alpha}_k$ and $\mathbf{D} = \mathbf{B}\mathbf{K}$ for $k = 1, \ldots, k^*$. Note that since the Gram matrix $\mathbf{K}$ is positive definite, the defined linear transformations through $\mathbf{K}$ are one-to-one. From this aspect, our strategy for the parameter estimation is as follows: first, we determine the estimate of $\mathbf{c}_k$ and $\mathbf{D}$, say $\widehat{\mathbf{c}_k}$ and $\widehat{\mathbf{D}}$; and then obtain the estimate of our target parameters by using $\widehat{\boldsymbol{\alpha}_k} = \mathbf{K}^{-1}\widehat{\mathbf{c}_k}$ and $\widehat{\mathbf{B}} = \widehat{\mathbf{D}}\mathbf{K}^{-1}$, for $k = 1, \ldots, k^*$. Hence, $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_{k^*}]^{\mathbf{T}}$ and $\mathbf{D}$ will be treated as the parameter of interest. Note that we are also required to handle the hidden parameter $\mathbf{z} = (z_1, \ldots, z_n)$.

6.3 Bayesian modeling

6.3.1 Likelihood

Define $\mathbf{y}_i^D = \mathbf{y}_i - \mathbf{D}^{\mathbf{T}}\mathbf{x}_i$ for $i = 1, \ldots, n$, where $\mathbf{y}_i^D$ can be considered as the i^{th} adjusted observation vector by the corresponding covariates. Then, the parametric model (6.2) can be restated as

$$E(\mathbf{y}_i^D \mid \mathbf{c}_{z_i}, \mathbf{D}) = \mathbf{c}_{z_i}, \quad i = 1, \ldots, n. \tag{6.3}$$

If we consider $\mathbf{y}_i^D$'s as observations, then (6.3) can be viewed as a parametric clustering model in a machine learning context. From this perspective, we *temporarily* assume that $\mathbf{y}_1^D, \ldots, \mathbf{y}_n^D$ are observed and then develop the likelihood using model-based clustering approach.

To generalize various clustering methods, Banerjee et al. (2005) introduced *Bregman hard clustering* and *Bregman soft clustering* via maximization of the following criteria, respectively:

$$Hard(\mathbf{C}, \mathbf{z}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{k^*} \mathbf{1}_{\{z_i = k\}} \exp\left\{ -\mathrm{BD}_\psi\left(\mathbf{y}_i^D, \mathbf{c}_k\right) \right\} \right]; \tag{6.4}$$

$$Soft(\mathbf{C}, \mathbf{p}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{k^*} p_k \exp\left\{ -\mathrm{BD}_\psi\left(\mathbf{y}_i^D, \mathbf{c}_k\right) \right\} \right], \tag{6.5}$$

where $\mathrm{BD}_\psi(\cdot, \cdot)$ indicates the Bregman divergence defined in (1.1), $\mathbf{1}_{\{\cdot\}}$ denotes an indicator function, and $\mathbf{p} = (p_1, \ldots, p_{k^*})$ denotes a mixing probability vector such that $p_k \in [0, 1]$ and $\sum_{k=1}^{k^*} p_k = 1$. Note that in the hard clustering, each observation is assigned to a solitary cluster, whereas in the soft clustering, we assign the member-

ship probabilities of clusters to each observation. Since Bregman divergence includes many well-known loss functions, such as squared Euclidean distance, Kullback-Leibler (KL) divergence, Itakura-Saito distance (Itakura and Saito, 1970) and Mahalanobis distance, the Bregman hard clustering unifies many existing hard clustering methods. For example, if $\psi(\mathbf{x}) = \mathbf{x}^{\mathbf{T}}\mathbf{x}$, then (6.4) reduces to the *k-means clustering*, which is one of the most popular and simple clustering algorithms. In addition, Banerjee et al. (2005) showed that every natural exponential family distribution corresponds to a unique and distinct Bregman divergence. Owing to this bijection property, the Bregman soft clustering can be viewed as a general extension of the exponential family mixture models.

To combine all attractive aspects of the Bregman clustering, we introduce a novel Bregman mixture clustering criterion,

$$Mix\left(\mathbf{C}, \mathbf{Z}, \mathbf{p}\right) \;=\; \prod_{i=1}^{n}\left[\sum_{k=1}^{k^*}\nu(z_i, p_k, \omega)\exp\left\{-\mathrm{BD}_{\psi}\left(\mathbf{y}_i^D, \mathbf{c}_k\right)\right\}\right], \tag{6.6}$$

where $\nu(z_i, p_k, \omega) = \omega\mathbf{1}_{\{z_i=k\}} + (1-\omega)p_k$ and $\omega \in [0,1]$ is a deterministic tuning parameter. Note that, $\nu(z_i, p_1, \omega), \ldots, \nu(z_i, p_{k^*}, \omega)$ are mixing probabilities due to the fact that $\nu(z_i, p_k, \omega) \in [0,1]$ and $\sum_{k=1}^{k^*}\nu(z_i, p_k, \omega) = 1$ for any given z_i and ω. However, unlike the classical soft clustering that assigns the same set of mixing probabilities to all observations, we assign a distinct set of mixing probabilities for each observation, which is a very important feature of our proposed methodology. It is worth noting that our proposed clustering method in (6.6) indeed unifies the Bregman clustering of Banerjee et al. (2005).

Remark 6.1. If $\omega = 0$, then (6.6) is identical to the Bregman soft clustering criterion in (6.4). If $\omega = 1$, then (6.6) is identical to the Bregman hard clustering criterion in

(6.5).

We now remove the temporary assumption for $\mathbf{y}_1^D, \ldots, \mathbf{y}_n^D$. Hence, from now on we regard $\mathbf{y}_1, \ldots, \mathbf{y}_n$ as observations. Consequently, by transforming (6.3) back to (6.2), $\mathbf{y}_i^D$ and $\mathbf{c}_k$ are respectively replaced by $\mathbf{y}_i$ and $\mathbf{c}_k + \mathbf{D}^{\mathbf{T}}\mathbf{x}_i$ in (6.6). In practice, we could have incomplete observations, i.e., some y_{il}'s are missing. To address this difficulty, we introduce an auxiliary variable $\bar{\mathbf{\Delta}} = \left[\bar{\Delta}_{il}\right]_{n\times q}$ with $\bar{\Delta}_{il} = \Delta_{il}/(\sum_{l=1}^{q} \Delta_{il})$, where Δ_{il} denotes the observing indicator for y_{il}, i.e., $\Delta_{il} = 1$ if y_{il} is observed and $\Delta_{il} = 0$, otherwise. Now, we define the likelihood as

$$f\left(\mathbf{Y} \mid \mathbf{C}, \mathbf{D}, \mathbf{z}, \mathbf{p}\right) \;\propto\; \prod_{i=1}^{n}\left[\sum_{k=1}^{k^*} \nu(z_i, p_k, \omega) \exp\left\{-\overline{\mathrm{BD}}_{\psi}\left(\mathbf{y}_i, \mathbf{c}_k + \mathbf{D}^{\mathbf{T}}\mathbf{x}_i\right)\right\}\right], \quad (6.7)$$

where

$$\begin{aligned} \nu(z_i, p_k, \omega) &= \omega \mathbf{1}_{\{z_i = k\}} + (1-\omega)p_k, \\ \overline{\mathrm{BD}}_{\psi}\left(\mathbf{y}_i, \mathbf{c}_k + \mathbf{D}^{\mathbf{T}}\mathbf{x}_i\right) &= \sum_{l=1}^{q} \bar{\Delta}_{il} \mathrm{BD}_{\psi}\left(y_{il}, \mathbf{C}_{[k,l]} + \mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}\right). \end{aligned} \quad (6.8)$$

Strictly speaking, the defined $\overline{\mathrm{BD}}_{\psi}$ in (6.8) is not the Bregman divergence defined in (1.1), since missing values induce an empty convex subset. However, owing to $\bar{\mathbf{\Delta}}$, all missing values are entirely eliminated in $\overline{\mathrm{BD}}_{\psi}$ and, thus, it can be viewed as the Bregman divergence with respect to the observed values. Consequently, all attractive properties of the Bregman divergence are preserved in our proposed method. To complete our Bayesian approach, we need to specify the prior for $(\mathbf{C}, \mathbf{D}, \mathbf{z}, \mathbf{p})$, i.e., $\pi(\mathbf{C}, \mathbf{D}, \mathbf{z}, \mathbf{p})$, and this will be discussed in the following section.

6.3.2 Priors

To define an appropriate prior $\pi(\mathbf{C}, \mathbf{D}, \mathbf{z}, \mathbf{p})$, we consider independent priors as follows:

$$\pi(\mathbf{C}, \mathbf{D}, \mathbf{z}, \mathbf{p}) \quad \propto \quad \pi(\mathbf{C})\pi(\mathbf{D})\pi(\mathbf{z})\pi(\mathbf{p}). \tag{6.9}$$

Since we have no information on $(\mathbf{C}, \mathbf{z}, \mathbf{p})$, it is valid to consider the following non-informative (or flat) priors,

$$\begin{aligned}
\pi(\mathbf{C}) &\propto \exp(-\frac{1}{2\sigma_c^2}\|\mathbf{C}\|_F^2), \\
\pi(\mathbf{p}) &\propto \prod_{k=1}^{k^*} p_k^{d_k-1}, \\
\pi(\mathbf{z}) &\propto \prod_{i=1}^{n} \left(\sum_{k=1}^{k^*} v_k \mathbf{1}_{\{z_i=k\}} \right),
\end{aligned}$$

where $\sigma_c^2 = 10^{-10}$, $d_1 = \cdots = d_{k^*} = 1$, $v_1 = \cdots = v_{k^*} = 1/k^*$, and $\|\mathbf{C}\|_F = \sqrt{\mathbf{tr}(\mathbf{C}^\mathbf{T}\mathbf{C})}$ denotes the Frobenius norm. In order to remove insignificant predictors from the model (6.7), we need to utilize a row-wise sparsity-inducing prior for $\mathbf{D}$. Consequently, the irrelevant predictors are eliminated by forcing the corresponding rows of $\mathbf{D}$ to be entirely zeros. However, the sparsity-inducing prior should be carefully determined due to the fact that the misspecified prior can lead to inconsistent variable selection (Zou, 2006). In this chapter, by introducing auxiliary variable $\boldsymbol{\gamma}$, we introduce a row-wise sparsity-inducing prior,

$$\pi(\mathbf{D}) \quad \propto \quad \exp\left\{ -\lambda \sum_{j=1}^{p} \frac{(\mathbf{D}_{[j,\,]})^\mathbf{T}\mathbf{D}_{[j,\,]}}{\tau^2 + (\mathbf{D}_{[j,\,]})^\mathbf{T}\mathbf{D}_{[j,\,]}} \right\}, \tag{6.10}$$

where λ and τ are prefixed hyperparameters. Note that

$$(\mathbf{D}_{[j,\,]})^{\mathbf{T}}\mathbf{D}_{[j,\,]}/\left(\tau + (\mathbf{D}_{[j,\,]})^{\mathbf{T}}\mathbf{D}_{[j,\,]}\right) \longrightarrow \mathbf{1}\left\{(\mathbf{D}_{[j,\,]})^{\mathbf{T}}\mathbf{D}_{[j,\,]} \neq 0\right\} \quad \text{as} \quad \tau \to 0.$$

Hence, when τ is chosen to be sufficiently small, our prior closely mimics the ℓ_0-norm penalty which is free from the inconsistency. Hence, under MAP estimation, this prior induces row-wise sparse estimate of $\mathbf{D}$. We throughout this chapter define $\tau = 10^{-20}$ and this works very well in our numerical studies. The hyperparameter $\lambda(> 0)$ controls degrees of sparsity. From a perspective of Bayesian predictive variable selection, we propose to determine the optimal λ based on the Bayesian Information Criterion (BIC), which is the second order approximation to the prior predictive density (Schwarz, 1978).

Remark 6.2. Recall that our final target parameter is $\mathbf{B}(= \mathbf{D}\mathbf{K}^{-1})$ in (6.2). Since a Gram matrix $\mathbf{K}$ is positive definite, the induced sparsity on $\mathbf{D}$ via (6.10) is also preserved on $\mathbf{B}$ regardless of $\mathbf{K}$, i.e., $(\mathbf{D}_{[j,\,]})^{\mathbf{T}}\mathbf{D}_{[j,\,]} = 0 \;\Leftrightarrow\; (\mathbf{B}_{[j,\,]})^{\mathbf{T}}\mathbf{B}_{[j,\,]} = 0$ for $j = 1, \ldots, p$. Hence, our variable selection is invariant to a choice of the reproducing kernel $\kappa(\cdot, \cdot)$.

6.4 Posterior computation

In order to induce the sparsity on $\mathbf{D}$, the parameter esimation should be based on MAP. To find the MAP estimate, we discuss *Iterated conditional modes* (ICM) algorithm (Besag, 1986) that iteratively maximizes each full conditional distribution.

6.4.1 Conditional mode of z

Let $\xi_1, \ldots, \xi_n$ be independent auxiliary variables with probability $\pi(\xi_i) = \omega \mathbf{1}_{\{\xi_i=1\}} + (1-\omega)\mathbf{1}_{\{\xi_i=0\}}$ for $i = 1, \ldots, n$. Define

$$f(\mathbf{y}_i|\mathbf{C}, \mathbf{D}, z_i, \mathbf{p}, \xi_i) = \begin{cases} \sum_{k=1}^{k^*} \mathbf{1}_{\{z_i=k\}} f\left(\mathbf{y}_i|\mathbf{c}_k, \mathbf{D}\right) & \text{if} \quad \xi_i = 1 \\ \sum_{k=1}^{k^*} p_k f\left(\mathbf{y}_i|\mathbf{c}_k, \mathbf{D}\right) & \text{if} \quad \xi_i = 0 \end{cases},$$

where $f\left(\mathbf{y}_i|\mathbf{c}_k, \mathbf{D}\right) = \exp\left\{-\overline{\mathrm{BD}}_\psi\left(\mathbf{y}_i, \mathbf{c}_k + \mathbf{D}^{\mathbf{T}}\mathbf{x}_i\right)\right\}$. Then, it is easy to check that the likelihood in (6.7) can be expressed as

$$f\left(\mathbf{Y}|\mathbf{C}, \mathbf{D}, \mathbf{z}, \mathbf{p}\right) \propto \prod_{i=1}^{n} \left\{ \int_{\xi_i \in \{0,1\}} f(\mathbf{y}_i|\mathbf{C}, \mathbf{D}, z_i, \mathbf{p}, \xi_i) d^{\pi}(\xi_i) \right\}. \tag{6.11}$$

From (6.11), using Jensen's inequality, we have that

$$\begin{aligned} \log f\left(\mathbf{Y}|\mathbf{C}, \mathbf{D}, \mathbf{z}, \mathbf{p}\right) &\geq \sum_{i=1}^{n} \left\{ \int_{\xi_i \in \{0,1\}} \log f(\mathbf{y}_i|\mathbf{C}, \mathbf{D}, z_i, \mathbf{p}, \xi_i) d^{\pi}(\xi_i) \right\} + c \\ &= \omega \sum_{i=1}^{n} \log\left(\sum_{k=1}^{k^*} \mathbf{1}_{\{z_i=k\}} f\left(\mathbf{y}_i|\mathbf{c}_k, \mathbf{D}\right) \right) \\ &\quad + (1-\omega) \sum_{i=1}^{n} \log\left(\sum_{k=1}^{k^*} p_k f\left(\mathbf{y}_i|\mathbf{c}_k, \mathbf{D}\right) \right) + c, \end{aligned} \tag{6.12}$$

where c indicates a constant. Note that z_i is only related to the first term in (6.12). Since $\pi(z_i) \propto \sum_{k=1}^{k^*} \mathbf{1}_{\{z_i=k\}}$. Hence, it is easy to check that the full conditional modes of $z_1, \ldots, z_n$ are given by

$$\arg\max_{z_i \in \mathcal{I}} \pi\left(z_i \mid \text{rest}\right) = \arg\min_{z_i \in \mathcal{I}} \left\{ \sum_{l=1}^{q} \bar{\Delta}_{il} \mathrm{BD}_\psi\left(y_{il}, \mathbf{C}_{[z_i,l]} + \mathbf{x}_i^{\mathbf{T}} \mathbf{D}_{[\ ,l]}\right) \right\}, \quad i = 1, \ldots, n, \tag{6.13}$$

where $\mathcal{I} = \{1, \ldots, k^*\}$.

6.4.2 Conditional mode of p

To find the conditional modes of $p_1 \ldots, p_{k^*}$, we use the Expectation-Maximization (EM) algorithm for the soft clustering (or the finite mixture model). Let $\eta_1, \ldots, \eta_n$ be independent auxiliary variables with probability

$$\pi\left(\eta_i \mid \boldsymbol{\theta}, \mathcal{D}_i\right) = \sum_{m=1}^{k^*} \frac{p_{\eta_i} \exp\left\{-\overline{\mathrm{BD}}_\psi\left(\mathbf{y}_i, \mathbf{c}_{\eta_i} + \mathbf{D}^{\mathbf{T}}\mathbf{x}_i\right)\right\}}{\sum_{k=1}^{k^*} p_k \exp\left\{-\overline{\mathrm{BD}}_\psi\left(\mathbf{y}_i, \mathbf{c}_k + \mathbf{D}^{\mathbf{T}}\mathbf{x}_i\right)\right\}} \mathbf{1}_{\{\eta_i = m\}}, \tag{6.14}$$

for $i = 1, \ldots, n$, where $\boldsymbol{\theta} = (\mathbf{C}, \mathbf{D}, \mathbf{z}, \mathbf{p})$ and $\mathcal{D}_i = (\mathbf{x}_i, \mathbf{y}_i)$. Then, from the EM algorithm, the full conditional mode of p_k is given as

$$\arg\max_{p_k} \pi\left(p_k \mid \text{rest}\right) \quad = \quad \frac{\sum_{i=1}^{n} \pi\left(\eta_i = k | \boldsymbol{\theta}^{(t)}, \mathcal{D}_i\right) + d_k - 1}{\sum_{k=1}^{k^*} \sum_{i=1}^{n} \pi\left(\eta_i = k | \boldsymbol{\theta}^{(t)}, \mathcal{D}_i\right) + \sum_{k=1}^{k^*} d_k - k^*}, \tag{6.15}$$

where $\boldsymbol{\theta}^{(t)} = (\mathbf{C}^{(t)}, \mathbf{D}^{(t)}, \mathbf{z}^{(t)}, \mathbf{p}^{(t)})$ denotes the current estimate of $\boldsymbol{\theta}$ in the ICM update.

6.4.3 Conditional mode of C and D

Similar to previous section, let $u_1, \ldots, u_n$ be independent auxiliary variables with probability

$$\pi\left(u_i \mid \boldsymbol{\theta}, \mathcal{D}_i\right) = \sum_{m=1}^{k^*} \frac{\nu(z_i, p_{u_i}, \omega) f\left(\mathbf{y}_i | \mathbf{c}_{u_i}, \mathbf{D}\right)}{\sum_{k=1}^{k^*} \nu(z_i, p_k, \omega) f\left(\mathbf{y}_i | \mathbf{c}_k, \mathbf{D}\right)} \mathbf{1}_{\{u_i = m\}},$$

for $i = 1, \ldots, n$, where $f\left(\mathbf{y}_i | \mathbf{c}_k, \mathbf{D}\right) = \exp\left\{-\overline{\mathrm{BD}}_\psi\left(\mathbf{y}_i, \mathbf{c}_k + \mathbf{D}^{\mathbf{T}}\mathbf{x}_i\right)\right\}$ and $\nu(z_i, p_k, \omega) = \omega \mathbf{1}_{\{z_i = k\}} + (1 - \omega) p_k$. Note that if $\omega = 1$, then we have $\pi\left(u_i | \boldsymbol{\theta}, \mathcal{D}_i\right) = \mathbf{1}_{\{u_i = z_i\}}$ and

if $\omega = 1$ then u_i is identical to η_i in (6.14) with probability one. Using Jensen's inequality, we have that

$$\begin{aligned}
\log f\left(\mathbf{Y}|\mathbf{C},\mathbf{D},\mathbf{z},\mathbf{p}\right) &= \sum_{i=1}^{n}\log\left\{\sum_{k=1}^{k^*}\pi(u_i=k|\boldsymbol{\theta}^{(t)},\mathcal{D}_i)\frac{\nu(z_i,p_k,\omega)f\left(\mathbf{y}_i|\mathbf{c}_k,\mathbf{D}\right)}{\pi(u_i=k|\boldsymbol{\theta}^{(t)},\mathcal{D}_i)}\right\}+c\\
&\geq \sum_{i=1}^{n}\sum_{k=1}^{k^*}\left[\pi(u_i=k|\boldsymbol{\theta}^{(t)},\mathcal{D}_i)\log\left\{\frac{\nu(z_i,p_k,\omega)f\left(\mathbf{y}_i|\mathbf{c}_k,\mathbf{D}\right)}{\pi(u_i=k|\boldsymbol{\theta}^{(t)},\mathcal{D}_i)}\right\}\right]+c\\
&= \sum_{i=1}^{n}\sum_{k=1}^{k^*}\left[\pi(u_i=k|\boldsymbol{\theta}^{(t)},\mathcal{D}_i)\log\left\{\frac{\nu(z_i,p_k,\omega)}{\pi(u_i=k|\boldsymbol{\theta}^{(t)},\mathcal{D}_i)}\right\}\right]\\
&\quad+\sum_{i=1}^{n}\sum_{k=1}^{k^*}\left[\pi(u_i=k|\boldsymbol{\theta}^{(t)},\mathcal{D}_i)\left\{-\overline{\mathrm{BD}}_{\psi}\left(\mathbf{y}_i,\mathbf{c}_k+\mathbf{D}^{\mathbf{T}}\mathbf{x}_i\right)\right\}\right]+c, \qquad (6.16)
\end{aligned}$$

where c indicates a constant. In (6.16), $\mathbf{C}$ and $\mathbf{D}$ are only related to the second term and thus maximizing this term maximizes the likelihood $f\left(\mathbf{Y}|\mathbf{C},\mathbf{D},\mathbf{z},\mathbf{p}\right)$ with respect to $\mathbf{C}$ and $\mathbf{D}$ for given $\mathbf{z}$ and $\mathbf{p}$. Consequently, the full conditional mode for $\mathbf{C}$ and $\mathbf{D}$ can be obtained by maximizing the following pseudo conditional of $\mathbf{C}$ and $\mathbf{D}$:

$$\widetilde{\pi}(\mathbf{C},\mathbf{D}|\cdots) \;\propto\; \exp\left\{-Q(\mathbf{C},\mathbf{D}|\boldsymbol{\theta}^{(t)},\mathcal{D})\right\}\pi(\mathbf{C})\pi(\mathbf{D}), \qquad (6.17)$$

where $Q(\mathbf{C},\mathbf{D}|\boldsymbol{\theta}^{(t)},\mathcal{D})=\sum_{i=1}^{n}\sum_{k=1}^{k^*}\left\{\pi(u_i=k|\boldsymbol{\theta}^{(t)},\mathcal{D}_i)\overline{\mathrm{BD}}_{\psi}\left(\mathbf{y}_i,\mathbf{c}_k+\mathbf{D}^{\mathbf{T}}\mathbf{x}_i\right)\right\}$.

First, let us discuss finding the conditional mode of $\mathbf{C}$. Define

$$Q_k^C(\mathbf{c}_k)=\sum_{i=1}^{n}\sum_{l=1}^{q}\pi(u_i=k|\boldsymbol{\theta}^{(t)},\mathcal{D}_i)\left\{\bar{\Delta}_{il}\mathrm{BD}_{\psi}\left(y_{il},c_{kl}+\mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}\right)\right\},$$

for $k=1\ldots,k^*$. Note that $\sum_{k=1}^{k^*}Q_k^C(\mathbf{c}_k)=Q(\mathbf{C},\mathbf{D}|\boldsymbol{\theta}^{(t)},\mathcal{D})$. Let $\mathbf{c}_k^{(t)}$ be the current estimate of $\mathbf{c}_k$ in the ICM update. Then using Taylor expansion at the current

estimate, we define a quadratic approximation $\tilde{Q}_k^C(\cdot|\mathbf{c}_k^{(t)})$ to $Q_k^C(\cdot)$ as

$$\tilde{Q}_k^C(\mathbf{c}_k|\mathbf{c}_k^{(t)}) = (\mathbf{c}_k)^{\mathbf{T}} \left\{\nabla Q_k^C(t)\right\} + \frac{1}{2}(\mathbf{c}_k - \mathbf{c}_k^{(t)})^{\mathbf{T}} \left\{\nabla^2 Q_k^C(t)\right\} (\mathbf{c}_k - \mathbf{c}_k^{(t)}), \qquad (6.18)$$

where $\nabla Q_k^C(t)$ and $\nabla^2 Q_k^C(t)$ respectively denote the gradient vector and the Hessian matrix of Q_k^C at $\mathbf{c}_k^{(t)}$ and they can be explicitly expressed as

$$\begin{aligned}
\left[\nabla Q_k^C(t)\right]_l &= \sum_{i=1}^{n} \pi(u_i = k|\boldsymbol{\theta}^{(t)}, \mathcal{D}_i)\bar{\Delta}_{il} \left(c_{kl}^{(t)} + \mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}^{(t)} - y_{il}\right) \psi''\left(c_{kl}^{(t)} + \mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}\right), \\
\left[\nabla^2 Q_k^C(t)\right]_{[l,s]} &= \sum_{i=1}^{n} \pi(u_i = k|\boldsymbol{\theta}^{(t)}, \mathcal{D}_i)\bar{\Delta}_{il} \left\{\psi''\left(c_{kl}^{(t)} + \mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}\right)\right. \\
&\quad \left. - \left(y_{il} - c_{kl}^{(t)} - \mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}^{(t)}\right) \psi'''\left(c_{kl}^{(t)} + \mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}\right)\right\} \mathbf{1}_{\{l=s\}}.
\end{aligned}$$

Since $\pi(\mathbf{C}) \propto \prod_{k=1}^{k^*} \exp\left(-\frac{1}{2\sigma_c^2}\mathbf{c}_k^{\mathbf{T}}\mathbf{c}_k\right)$, using (6.18) we have

$$\begin{aligned}
[\mathbf{c}_k \mid \text{rest}] \approx\ & N_q\left(\left[\nabla^2 Q_k^C(t) + \frac{1}{\sigma_c^2}\mathbf{I}_q\right]^{-1} \left\{\nabla^2 Q_k^C(t)\mathbf{c}_k^{(t)} - \nabla Q_k^C(t)\right\}, \right. \\
& \left. \left[\nabla^2 Q_k^C(t) + \frac{1}{\sigma_c^2}\mathbf{I}_q\right]^{-1}\right).
\end{aligned}$$

Hence, the full conditional mode of $\mathbf{c}_1, \ldots, \mathbf{c}_{k^*}$ are given as

$$\arg\max_{\mathbf{c}_k} \pi\left(\mathbf{c}_k|\cdots\right) = \left[\nabla^2 Q_k^C(t) + \frac{1}{\sigma_c^2}\mathbf{I}_q\right]^{-1} \left\{\nabla^2 Q_k^C(t)\mathbf{c}_k^{(t)} - \nabla Q_k^C(t)\right\}. \qquad (6.19)$$

Now, we discuss finding the conditional mode of $\mathbf{D}$. Define

$$Q_l^D(\mathbf{D}_{[\ ,l]}) = \sum_{i=1}^{n}\sum_{k=1}^{k^*} \pi(u_i = k|\boldsymbol{\theta}^{(t)}, \mathcal{D}_i) \left\{\bar{\Delta}_{il}\mathrm{BD}_\psi\left(y_{il}, c_{kl} + \mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}\right)\right\},$$

for $l = 1 \ldots, q$. Note that $\sum_{l=1}^{q} Q_l^D(\mathbf{D}_{[\ ,l]}) = Q(\mathbf{C}, \mathbf{D}|\boldsymbol{\theta}^{(t)}, \mathcal{D})$. Let $\mathbf{D}_{[\ ,l]}^{(t)}$ be the current estimate of $\mathbf{D}_{[\ ,l]}$ in the ICM iteration. Then using Taylor expansion at the current estimate, we define a quadratic approximation $\tilde{Q}_l^D(\cdot|\mathbf{D}^{(t)})$ to $Q_l^D(\cdot)$ such that

$$\tilde{Q}_l^D(\mathbf{D}_{[\ ,l]}|\mathbf{D}_{[\ ,l]}^{(t)}) = (\mathbf{D}_{[\ ,l]})^{\mathbf{T}} \left\{\nabla Q_l^D(t)\right\} + \frac{1}{2}(\mathbf{D}_{[\ ,l]} - \mathbf{D}_{[\ ,l]}^{(t)})^{\mathbf{T}} \left\{\nabla^2 Q_l^D(t)\right\} (\mathbf{D}_{[\ ,l]} - \mathbf{D}_{[\ ,l]}^{(t)}),$$

where $\nabla Q_l^D(t)$ and $\nabla^2 Q_l^D(t)$ respectively denote the gradient vector and the Hessian matrix of Q_l^D at $\mathbf{D}_{[\ ,l]}^{(t)}$ and they can be explicitly expressed as

$$\begin{aligned}
\nabla Q_l^D(t) &= \sum_{i=1}^{n}\sum_{k=1}^{k^*} \pi(u_i = k|\boldsymbol{\theta}^{(t)}, \mathcal{D}_i)\bar{\Delta}_{il} \left(c_{kl} + \mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}^{(t)} - y_{il}\right) \psi''\left(c_{kl} + \mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}^{(t)}\right) \mathbf{x}_i, \\
\nabla^2 Q_l^D(t) &= \sum_{i=1}^{n}\sum_{k=1}^{k^*} \pi(u_i = k|\boldsymbol{\theta}^{(t)}, \mathcal{D}_i)\bar{\Delta}_{il} \left\{\psi''\left(c_{kl} + \mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}\right)\right. \\
&\quad \left. - \left(y_{il} - c_{kl} - \mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}^{(t)}\right) \psi'''\left(c_{kl} + \mathbf{x}_i^{\mathbf{T}}\mathbf{D}_{[\ ,l]}\right)\right\} \mathbf{x}_i\mathbf{x}_i^{\mathbf{T}}.
\end{aligned}$$

Define $\gamma_j = \frac{\lambda}{\tau^2 + (\mathbf{D}_{[j,\ ]})^{\mathbf{T}}\mathbf{D}_{[j,\ ]}}$. Then we can easily show that

$$\lambda \sum_{j=1}^{p} \frac{(\mathbf{D}_{[j,\ ]})^{\mathbf{T}}\mathbf{D}_{[j,\ ]}}{\tau^2 + (\mathbf{D}_{[j,\ ]})^{\mathbf{T}}\mathbf{D}_{[j,\ ]}} = \sum_{j=1}^{p} \gamma_j (\mathbf{D}_{[j,\ ]})^{\mathbf{T}}\mathbf{D}_{[j,\ ]} = \mathbf{tr}\left\{\mathbf{D}^{\mathbf{T}}\,\mathbf{Diag}(\boldsymbol{\gamma})\mathbf{D}\right\},$$

where $\boldsymbol{\gamma} = (\gamma_1 \ldots, \gamma_p)^{\mathbf{T}}$. This implies that $\pi(\mathbf{D}) \propto \prod_{l=1}^{q} \exp\left(-\mathbf{D}_{[\ ,l]}^{\mathbf{T}}\,\mathbf{Diag}(\boldsymbol{\gamma})\mathbf{D}_{[\ ,l]}\right)$. Hence, the full conditional posterior mode of $\mathbf{D}_{[\ ,1]}, \ldots, \mathbf{D}_{[\ ,q]}$ can be determined by

$$\begin{aligned}
&\arg\max_{\mathbf{D}_{[\ ,l]}} \pi\left(\mathbf{D}_{[\ ,l]}|\cdots\right) = \arg\min_{\mathbf{D}_{[\ ,l]}} \left[\tilde{Q}_l^D(\mathbf{D}_{[\ ,l]}|\mathbf{D}_{[\ ,l]}^{(t)}) + (\mathbf{D}_{[\ ,l]})^{\mathbf{T}}\,\mathbf{Diag}(\boldsymbol{\gamma}^{(t)})\mathbf{D}_{[\ ,l]}\right] \\
&= \left[\nabla^2\tilde{Q}_l^D(t) + 2\,\mathbf{Diag}(\boldsymbol{\gamma})\right]^{-1} \left\{\nabla^2\tilde{Q}_l^D(t)\mathbf{D}_{[\ ,l]}^{(t)} - \nabla\tilde{Q}_l^D(t)\right\}, \qquad (6.20)
\end{aligned}$$

where $\boldsymbol{\gamma}^{(t)} = (\gamma_1^{(t)} \ldots, \gamma_p^{(t)})^{\mathbf{T}}$ with $\gamma_j^{(t)} = \frac{\lambda}{\tau^2 + (\mathbf{D}_{[j,\ ]}^{(t)})^{\mathbf{T}}\mathbf{D}_{[j,\ ]}^{(t)}}$.

6.5 Future works

To demonstrate the applicability of proposed method, we will conduct real data analysis using *Yeast cell cycle* data (Spellman et al., 1998) and *Yeast ChIP* data (Lee et al., 2002); each data set is publicly available at `http://genome-www.stanford.edu` and `http://younglab.wi.mit.edu/datadownload.htm`, respectively. In addition, simulation studies will be performed to verify the validity of our proposed method.

Appendix A

Proofs

A.1 Proof of theorem 2.1

$$\begin{aligned} D_\psi\left(f_1/f_2, 1\right) &= \int \left\{\psi\left(f_1/f_2\right) - \psi(1) - \left(f_1/f_2 - 1\right)\psi'(1)\right\} d\nu \\ &= \int \left\{\psi\left(\frac{f_1(x)}{f_2(x)}\right) - \psi(1) - \left(\frac{f_1(x)}{f_2(x)} - 1\right)\psi'(1)\right\} f_2(x)dx \\ &= \int \psi\left(\frac{f_1(x)}{f_2(x)}\right) f_2(x)dx - \psi(1) \\ &= \Phi_\psi\left(f_1, f_2\right) - \psi(1), \end{aligned}$$

which completes the proof.

A.2 Proof of theorem 2.6

$$
\begin{aligned}
& \arg\min_{(\boldsymbol{\mu},\boldsymbol{\Sigma})} D_{\psi}\left(f, g(\boldsymbol{\mu},\boldsymbol{\Sigma})\right) \\
= \; & \arg\min_{(\boldsymbol{\mu},\boldsymbol{\Sigma})} \int f(\mathbf{x}) \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}\right) d\mathbf{x} \\
= \; & \arg\max_{(\boldsymbol{\mu},\boldsymbol{\Sigma})} \int f(\mathbf{x}) \log\left\{g(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\right\} d\mathbf{x} \\
= \; & \arg\max_{(\boldsymbol{\mu},\boldsymbol{\Sigma})} \int \log\left\{\frac{1}{(2\pi)^{\frac{p}{2}}\left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right)\right\} f(\mathbf{x}) d\mathbf{x} \\
= \; & \arg\max_{(\boldsymbol{\mu},\boldsymbol{\Sigma})} \int \left\{\log\left(\left|\boldsymbol{\Sigma}^{-1}\right|\right) - (\mathbf{x}-\boldsymbol{\eta}+\boldsymbol{\eta}-\boldsymbol{\mu})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\eta}+\boldsymbol{\eta}-\boldsymbol{\mu})\right\} f(\mathbf{x}) d\mathbf{x} \\
= \; & \arg\max_{(\boldsymbol{\mu},\boldsymbol{\Sigma})} \left\{\log\left(\left|\boldsymbol{\Sigma}^{-1}\right|\right) - \mathbf{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{V}\right) - (\boldsymbol{\mu}-\boldsymbol{\eta})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\eta})\right\} \\
= \; & \left\{(\boldsymbol{\mu},\boldsymbol{\Sigma}) \;:\; \frac{\partial h\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)}{\partial \boldsymbol{\mu}} = \mathbf{0} \text{ and } \left.\frac{\partial h\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)}{\partial \boldsymbol{\Sigma}}\right|_{\boldsymbol{\mu}=\boldsymbol{\eta}} = \mathbf{0}\right\} \\
= \; & (\boldsymbol{\eta}, \mathbf{V}),
\end{aligned}
$$

where $h(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \log\left(|\boldsymbol{\Sigma}^{-1}|\right) - \mathbf{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{V}\right) - (\boldsymbol{\mu}-\boldsymbol{\eta})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\eta})$.

A.3 Proof of theorem 3.9

When $\{\mathbf{y}_{-s}\} = \emptyset$, the left hand side of (3.8) is

$$
\left[\frac{1}{N}\sum_{j=1}^{N}\frac{g(\boldsymbol{\theta}^{j})}{f(\mathbf{y}|\boldsymbol{\theta}^{j})\pi(\boldsymbol{\theta}^{j})}\right]^{-1},
$$

which converges, as $N \to \infty$, to $m(\mathbf{y})$ almost surely (Chen, 1994). Assume $\{\mathbf{y}_{-s}\} \neq \emptyset$, and then

$$\begin{aligned} p(\mathbf{y}_s|\mathbf{y}_{-s}) &= \frac{m(\mathbf{y})}{m(\mathbf{y}_{-s})} \\ &= \frac{m(\mathbf{y})}{\int_\Theta f(\mathbf{y}_{-s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &= \frac{m(\mathbf{y})}{\int_\Theta \frac{f(\mathbf{y}|\boldsymbol{\theta})}{f(\mathbf{y}_s|\mathbf{y}_{-s},\boldsymbol{\theta})}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &= \left[\int_\Theta \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{f(\mathbf{y}_s|\mathbf{y}_{-s},\boldsymbol{\theta})}d\boldsymbol{\theta}\right]^{-1} \\ &\overset{\text{a.s.}}{=} \lim_{N\to\infty}\left[\frac{1}{N}\sum_{j=1}^{N}\frac{1}{f(\mathbf{y}_s|\mathbf{y}_{-s},\boldsymbol{\theta}^j)}\right]^{-1}, \end{aligned}$$

here the last part of proof can be done by the pointwise ergodic theorem.

A.4 Proof of theorem 3.11

For given i, let $\hat{p}_i = \left[\frac{1}{N}\sum_{q=1}^{N}\frac{1}{f(\mathbf{y}_{i:n}|\mathbf{y}_{1:i-1},\boldsymbol{\theta}^q)}\right]^{-1}$. Note that

$$\begin{aligned} \sum_{j=1}^{N}\frac{\mathbf{1}(t^j \le y)}{N} &= \frac{\hat{p}_i}{\hat{p}_i}\left\{\sum_{j=1}^{N}\frac{\mathbf{1}(t^j \le y)}{N}\right\} \\ &= \hat{p}_i\left\{\frac{1}{N^2}\sum_{q=1}^{N}\sum_{j=1}^{N}\frac{\mathbf{1}(t^j \le y)}{f(\mathbf{y}_{i:n}|\mathbf{y}_{1:i-1},\boldsymbol{\theta}^q)}\right\}, \end{aligned}$$

and $\hat{p}_i$ converges to $p(\mathbf{y}_{i:n}|\mathbf{y}_{1:i-1})$ with probability 1 by Theorem 3.9. By the pointwise ergodic theorem and Fubini's theorem,

$$\begin{aligned}
\lim_{N\to\infty}\sum_{j=1}^{N}\frac{\mathbf{1}(t^j \le y)}{N} &\overset{\text{a.s.}}{=} p(\mathbf{y}_{i:n}|\mathbf{y}_{1:i-1})\int_{\mathbb{R}}\int_{\Theta}\frac{\mathbf{1}(t\le y)}{f(\mathbf{y}_{i:n}|\mathbf{y}_{1:i-1},\boldsymbol{\theta})}f_{Y_i}(t|\mathbf{y}_{1:i-1},\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}dt \\
&= \int_{\mathbb{R}}\mathbf{1}(t\le y)\int_{\Theta}f_{Y_i}(t|\mathbf{y}_{1:i-1},\boldsymbol{\theta})\frac{p(\mathbf{y}_{i:n}|\mathbf{y}_{1:i-1})\pi(\boldsymbol{\theta}|\mathbf{y})}{f(\mathbf{y}_{i:n}|\mathbf{y}_{1:i-1},\boldsymbol{\theta})}d\boldsymbol{\theta}dt \\
&= \int_{-\infty}^{y}\int_{\Theta}f_{Y_i}(t|\mathbf{y}_{1:i-1},\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}_{1:i-1})d\boldsymbol{\theta}dt \\
&= \int_{-\infty}^{y}p_{Y_i}(t|\mathbf{y}_{1:i-1})dt \\
&= P(Y_i\le y|\mathbf{y}_{1:i-1}).
\end{aligned}$$

This completes our proof.

A.5 Proof of lemma 4.3

Since $\sup_{\boldsymbol{\beta}\in\mathbb{R}^p} f(\mathbf{y}|\boldsymbol{\beta})d\boldsymbol{\beta} < \infty$, there exists a constant $m(<\infty)$ such that $f(\mathbf{y}|\boldsymbol{\beta}) < m$ for any $\boldsymbol{\beta}$. Hence, we have $\int f(\mathbf{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})d\boldsymbol{\beta} \le m\int\pi(\boldsymbol{\beta})d\boldsymbol{\beta} < \infty$ and this implies the propriety of the posterior.

A.6 Proof of theorem 5.4

To establish Theorem 5.4, we start by extending the first lemma in Armagan et al. (2013) to the multivariate case.

Lemma A.1. *Let* $\mathcal{C}_\epsilon = \{(\mathbf{A},\mathbf{B}) : \|\mathbf{C}-\mathbf{C}^*\|_F > \epsilon,\ \mathbf{C}=\mathbf{A}\mathbf{B}^{\mathbf{T}}\}$, *where* $\mathbf{C}^*$ *denotes the true coefficient matrix. Define* $\Phi_n = I(\mathbf{Y}_n \in \mathcal{Y}_n)$, *where* $\mathcal{Y}_n = \{\mathbf{Y}_n : \|\hat{\mathbf{C}}_n - \mathbf{C}^*\|_F >$

$\epsilon/2\}$ *and* $\hat{\mathbf{C}}_n = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}_n$. *Then, under assumptions I and II, for sufficiently large* n,

$$E_{\mathbf{Y}_n|\mathbf{C}^*}(\Phi_n) \leq \exp\left(-\frac{\epsilon^2 n S_{\min}^2}{16\tau_{\max}}\right), \tag{A.1}$$

$$\sup_{(\mathbf{A},\mathbf{B})\in\mathcal{C}_\epsilon} E_{\mathbf{Y}_n|\mathbf{A},\mathbf{B}}(1-\Phi_n) \leq \exp\left(-\frac{\epsilon^2 n S_{\min}^2}{16\tau_{\max}}\right), \tag{A.2}$$

where $\tau_{\max}$ *denotes the largest eigenvalue of* $\mathbf{\Sigma}$.

Proof of lemma A.1

Using assumption II, we have

$$\begin{aligned}
E_{\mathbf{Y}_n|\mathbf{C}^*}(\Phi_n) &= P_{\mathbf{Y}_n|\mathbf{C}^*}\left(\mathbf{Y}_n : \|\hat{\mathbf{C}}_n - \mathbf{C}^*\|_F > \epsilon/2\right) \\
&\leq P_{\mathbf{Y}_n|\mathbf{C}^*}\left(\mathbf{Y}_n : \mathbf{tr}\{(\hat{\mathbf{C}}_n - \mathbf{C}^*)^\mathbf{T}(\hat{\mathbf{C}}_n - \mathbf{C}^*)\mathbf{\Sigma}^{-1}\} > \epsilon^2 n/(4\tau_{\max})\right) \\
&\leq P_{\mathbf{Y}_n|\mathbf{C}^*}\Big(\mathbf{Y}_n : \mathbf{tr}\{\mathbf{\Sigma}^{-1/2}(\hat{\mathbf{C}}_n - \mathbf{C}^*)^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{X}(\hat{\mathbf{C}}_n - \mathbf{C}^*)\mathbf{\Sigma}^{-1/2}\} \\
&\qquad > \epsilon^2 S_{n,\min}^2/(4\tau_{\max})\Big) \\
&\leq P\left(\chi^2_{p_n q} > \epsilon^2 n S_{\min}^2/(4\tau_{\max})\right), \qquad (A.3)
\end{aligned}$$

where χ^2_m denotes a chi-squared random variable with m degrees of freedom. Similary,

$$\begin{aligned}
&\sup_{(\mathbf{A},\mathbf{B})\in\mathcal{C}_\epsilon} E_{\mathbf{Y}_n|\mathbf{A},\mathbf{B}}(1-\Phi_n) = \sup_{(\mathbf{A},\mathbf{B})\in\mathcal{C}_\epsilon} P_{\mathbf{Y}_n|\mathbf{A},\mathbf{B}}\left(\mathbf{Y}_n : \|\hat{\mathbf{C}}_n - \mathbf{C}^*\|_F \le \epsilon/2\right) \\
&\le \sup_{(\mathbf{A},\mathbf{B})\in\mathcal{C}_\epsilon} P_{\mathbf{Y}_n|\mathbf{A},\mathbf{B}}\left(\mathbf{Y}_n : |\|\hat{\mathbf{C}}_n - \mathbf{C}\|_F - \|\mathbf{C}-\mathbf{C}^*\|_F| \le \epsilon/2\right) \\
&\le \sup_{(\mathbf{A},\mathbf{B})\in\mathcal{C}_\epsilon} P_{\mathbf{Y}_n|\mathbf{A},\mathbf{B}}\left(\mathbf{Y}_n : \|\hat{\mathbf{C}}_n - \mathbf{C}\|_F \ge -\epsilon/2 + \|\mathbf{C}-\mathbf{C}^*\|_F\right) \\
&= P_{\mathbf{Y}_n|\mathbf{C}=\mathbf{A}\mathbf{B}^\mathbf{T}}\left(\mathbf{Y}_n : \|\hat{\mathbf{C}}_n - \mathbf{C}\|_F \ge \epsilon/2\right) \\
&\le P\left(\chi^2_{p_n q} > \epsilon^2 n S^2_{\min}/(4\tau_{\max})\right). \qquad \text{(A.4)}
\end{aligned}$$

According to Armagan et al. (2013), we note that

$$P(\chi^2_m \ge x) \le \exp(-x/4), \quad \text{if } x \ge 8m. \qquad \text{(A.5)}$$

Under assumption I, for sufficiently large n, (A.3) and (A.4) thus imply (A.1) and (A.2), respectively.

We now show the proof of Theorem 5.4 using similar technique as in Armagan et al. (2013). Given Lemma A.1, the posterior probability of $\mathcal{C}_\epsilon$ can be bounded as follows:

$$\Pi(\mathcal{C}_\epsilon|\mathbf{Y}_n) = \frac{\int_{\mathcal{C}_\epsilon}\left\{\frac{f(\mathbf{Y}_n|\mathbf{A},\mathbf{B})}{f(\mathbf{Y}_n|\mathbf{C}^*)}\right\}\Pi(d\mathbf{A}d\mathbf{B})}{\int\left\{\frac{f(\mathbf{Y}_n|\mathbf{A},\mathbf{B})}{f(\mathbf{Y}_n|\mathbf{C}^*)}\right\}\Pi(d\mathbf{A}d\mathbf{B})} \le \Phi_n + \frac{(1-\Phi_n)J_{\mathcal{C}_\epsilon}}{J_n}, \qquad \text{(A.6)}$$

where $J_{\mathcal{C}_\epsilon} = \int_{\mathcal{C}_\epsilon}\left\{\frac{f(\mathbf{Y}_n|\mathbf{A},\mathbf{B})}{f(\mathbf{Y}_n|\mathbf{C}^*)}\right\}\Pi(d\mathbf{A}d\mathbf{B})$ and $J_n = \int\left\{\frac{f(\mathbf{Y}_n|\mathbf{A},\mathbf{B})}{f(\mathbf{Y}_n|\mathbf{C}^*)}\right\}\Pi(d\mathbf{A}d\mathbf{B})$.

Define $I_1 = \Phi_n$ and $I_2 = (1-\Phi_n)J_{\mathcal{C}_\epsilon}$. Then the inequality (A.6) can be written as

$$\Pi(\mathcal{C}_\epsilon \mid \mathbf{Y}_n) \le I_1 + I_2/J_n. \qquad \text{(A.7)}$$

Let $b = \epsilon^2 S_{\min}^2/(16\tau_{\max})$. For sufficiently large n, using Markov's inequality and (A.1) of Lemma A.1, we have

$$P_{\mathbf{Y}_n|\mathbf{C}^*}\left\{I_1 \geq \exp(-bn/2)\right\} \leq \exp(bn/2) E_{\mathbf{Y}_n|\mathbf{C}^*}(I_1) \leq \exp(-bn/2). \tag{A.8}$$

This implies $\sum_n P_{\mathbf{Y}_n|\mathbf{C}^*}\left\{I_1 \geq \exp(-bn/2)\right\} < \infty$. Hence, using the Borel-Cantelli lemma, we have $I_1 < \exp(-bn/2)$ almost surely. From (A.2) of Lemma A.1, we have

$$\begin{aligned}
E_{\mathbf{Y}_n|\mathbf{C}^*}(I_2) &= E_{\mathbf{Y}_n|\mathbf{C}^*}\left\{(1-\Phi_n)J_{\mathcal{C}_\epsilon}\right\} \\
&= E_{\mathbf{Y}_n|\mathbf{C}^*}\left\{(1-\Phi_n)\int_{\mathcal{C}_\epsilon}\frac{f(\mathbf{Y}_n \mid \mathbf{A},\mathbf{B})}{f(\mathbf{Y}_n \mid \mathbf{C}^*)}\Pi(d\mathbf{A}d\mathbf{B})\right\} \\
&= \int_{\mathcal{C}_\epsilon}\left\{\int(1-\Phi_n)f(\mathbf{Y}_n \mid \mathbf{A},\mathbf{B})d\mathbf{Y}_n\right\}\Pi(d\mathbf{A}d\mathbf{B}) \\
&\leq \int_{\mathcal{C}_\epsilon}\sup_{(\mathbf{A},\mathbf{B})\in\mathcal{C}_\epsilon}\left\{\int(1-\Phi_n)f(\mathbf{Y}_n \mid \mathbf{A},\mathbf{B})d\mathbf{Y}_n\right\}\Pi(d\mathbf{A}d\mathbf{B}) \\
&= \Pi(\mathcal{C}_\epsilon)\sup_{(\mathbf{A},\mathbf{B})\in\mathcal{C}_\epsilon}\left\{E_{\mathbf{Y}_n|\mathbf{A},\mathbf{B}}(1-\Phi_n)\right\} \\
&\leq \exp(-bn).
\end{aligned}$$

Hence, for sufficiently large n, $P_{\mathbf{Y}_n|\mathbf{C}^*}\left\{I_2 \geq \exp(-bn/2)\right\} \leq \exp(-bn/2)$. It implies that $\sum_n P_{\mathbf{Y}_n|\mathbf{C}^*}\left\{I_2 \geq \exp(-bn/2)\right\} < \infty$. By the Borel-Cantelli lemma, we have $I_2 < \exp(-bn/2)$ almost surely.

Now, to establish posterior consistency, it suffices to show $\exp(bn/2)J_n \to \infty$ almost surely as $n \to \infty$. Define a set $\mathcal{D}_{n,\nu}$ such that

$$\mathcal{D}_{n,\nu} = \left\{(\mathbf{A},\mathbf{B}) : n^{-1}\log\left\{f(\mathbf{Y}_n|\mathbf{C}^*)/f(\mathbf{Y}_n|\mathbf{A},\mathbf{B})\right\} < \nu\right\},$$

for $0 < \nu < b/2$. Then we have

$$\begin{aligned}
\exp(bn/2)J_n &= \exp(bn/2)\int \exp\left\{-n\frac{1}{n}\log\frac{f(\mathbf{Y}_n \mid \mathbf{C}^*)}{f(\mathbf{Y}_n \mid \mathbf{A},\mathbf{B})}\right\}\Pi(d\mathbf{A}d\mathbf{B}) \\
&\geq \exp(bn/2)\int_{\mathcal{D}_{n,\nu}} \exp\left\{-n\frac{1}{n}\log\frac{f(\mathbf{Y}_n \mid \mathbf{C}^*)}{f(\mathbf{Y}_n \mid \mathbf{A},\mathbf{B})}\right\}\Pi(d\mathbf{A}d\mathbf{B}) \\
&\geq \exp\left\{(b/2-\nu)n\right\}\Pi(\mathcal{D}_{n,\nu}). \qquad \text{(A.9)}
\end{aligned}$$

Define $\kappa_n = n^{\frac{1+\rho}{2}}$, where $0 < \rho < 1$. Using (A.5), for sufficiently large n, it is easy to show that $P_{\mathbf{Y}_n|\mathbf{C}^*}(\mathbf{Y}_n : \|\mathbf{Y}_n - \mathbf{X}\mathbf{C}^*\|_F^2 > \kappa_n^2) \leq P(\chi_{nq}^2 > \kappa_n^2/\tau_{\max}) \leq \exp\{-\kappa_n^2/(4\tau_{\max})\}$, and this implies $\sum_n P_{\mathbf{Y}_n|\mathbf{C}^*}(\mathbf{Y}_n : \|\mathbf{Y}_n - \mathbf{X}\mathbf{C}^*\|_F^2 > \kappa_n^2) < \infty$. Hence, by the Borel-Cantelli lemma, we have

$$\|\mathbf{Y}_n - \mathbf{X}\mathbf{C}^*\|_F \leq \kappa_n, \quad \text{almost surely.} \qquad \text{(A.10)}$$

Note that

$$\begin{aligned}
\mathcal{D}_{n,\nu} &= \big\{(\mathbf{A},\mathbf{B}) : n^{-1}(\|(\mathbf{Y}_n - \mathbf{X}\mathbf{C})\mathbf{\Sigma}^{-1/2}\|_F^2 - \|(\mathbf{Y}_n - \mathbf{X}\mathbf{C}^*)\mathbf{\Sigma}^{-1/2}\|_F^2) < 2\nu, \\
&\qquad \mathbf{C} = \mathbf{A}\mathbf{B}^{\mathbf{T}}\big\} \\
&\supseteq \left\{(\mathbf{A},\mathbf{B}) : n^{-1}\left|\|\mathbf{Y}_n - \mathbf{X}\mathbf{C}\|_F^2 - \|\mathbf{Y}_n - \mathbf{X}\mathbf{C}^*\|_F^2\right| < 2\tau_{\min}\nu, \mathbf{C} = \mathbf{A}\mathbf{B}^{\mathbf{T}}\right\}.
\end{aligned}$$

Using the facts that $|x^2 - y^2| = |\{2|y|(|x| - |y|) + (|x| - |y|)^2\}| \leq 2|y||x-y| + |x-y|^2$

and (A.10), for sufficiently large n, we have

$$\begin{aligned}
\Pi(\mathcal{D}_{n,\nu}) &\geq \Pi\left\{(\mathbf{A},\mathbf{B}) : n^{-1}\left(2\|\mathbf{Y}_n - \mathbf{X}\mathbf{C}^*\|_F\|\mathbf{X}(\mathbf{C}-\mathbf{C}^*)\|_F\right.\right. \\
&\quad \left.\left. +\|\mathbf{X}(\mathbf{C}-\mathbf{C}^*)\|_F^2\right) < 2\tau_{\min}\nu,\ \mathbf{C} = \mathbf{A}\mathbf{B}^{\mathbf{T}}\right\} \\
&\geq \Pi\left\{(\mathbf{A},\mathbf{B}) : n^{-1}2\kappa_n\|\mathbf{X}(\mathbf{C}-\mathbf{C}^*)\|_F < \frac{4\tau_{\min}\nu}{3},\right. \\
&\quad \left.\|\mathbf{X}(\mathbf{C}-\mathbf{C}^*)\|_F^2 < \frac{2\tau_{\min}\nu}{3},\ \mathbf{C} = \mathbf{A}\mathbf{B}^{\mathbf{T}}\right\} \\
&\geq \Pi\left\{(\mathbf{A},\mathbf{B}) : n^{-1}\|\mathbf{X}(\mathbf{C}-\mathbf{C}^*)\|_F < \frac{2\tau_{\min}\nu}{3\kappa_n},\ \mathbf{C} = \mathbf{A}\mathbf{B}^{\mathbf{T}}\right\} \\
&\geq \Pi\left\{(\mathbf{A},\mathbf{B}) : \sqrt{n}S_{\max}\|\mathbf{C}-\mathbf{C}^*\|_F < \frac{2\tau_{\min}\nu}{3\kappa_n}, \mathbf{C} = \mathbf{A}\mathbf{B}^{\mathbf{T}}\right\} \\
&= \Pi\left\{(\mathbf{A},\mathbf{B}) : \|\mathbf{C}-\mathbf{C}^*\|_F < \frac{\Delta}{n^{\frac{\rho}{2}}}, \mathbf{C} = \mathbf{A}\mathbf{B}^{\mathbf{T}}\right\}, \qquad \text{(A.11)}
\end{aligned}$$

where $\Delta = \frac{2\tau_{\min}\nu}{3S_{\max}}$. Therefore, if $\Pi\left\{(\mathbf{A},\mathbf{B}) : \|\mathbf{A}\mathbf{B}^{\mathbf{T}} - \mathbf{C}^*\|_F < \frac{\Delta}{n^{\frac{\rho}{2}}}\right\} > \exp(-dn)$ for all $0 < d < b/2 - \nu$, then we have $\exp(bn/2)J_n \to \infty$ almost surely as $n \to \infty$. This completes the proof.

A.7 Proof of theorem 5.5

Using the fact that $\frac{x^2}{(\omega_0+x^2)} \leq \frac{|x|}{2\sqrt{\omega_0}}$ for any x and $\omega_0 > 0$, it is easy to show that

$$\begin{aligned}
&\exp\left\{-\frac{1}{2}\left(\lambda_1\sum_{k=1}^{r}\frac{\tilde{\mathbf{a}}_k^{\mathbf{T}}\tilde{\mathbf{a}}_k + \tilde{\mathbf{b}}_k^{\mathbf{T}}\tilde{\mathbf{b}}_k}{\omega_0 + \tilde{\mathbf{a}}_k^{\mathbf{T}}\tilde{\mathbf{a}}_k + \tilde{\mathbf{b}}_k^{\mathbf{T}}\tilde{\mathbf{b}}_k} + \lambda_2\sum_{j=1}^{p_n}\frac{\mathbf{a}_j^{\mathbf{T}}\mathbf{a}_j}{\omega_0 + \mathbf{a}_j^{\mathbf{T}}\mathbf{a}_j} + \lambda_3\sum_{l=1}^{q}\frac{\mathbf{b}_l^{\mathbf{T}}\mathbf{b}_l}{\omega_0 + \mathbf{b}_l^{\mathbf{T}}\mathbf{b}_l}\right)\right\} \\
\geq\ &\exp\left\{-\frac{1}{2}\left((\lambda_1+\lambda_2)\sum_{j=1}^{p_n}\sum_{k=1}^{r}\frac{a_{jk}^2}{\omega_0 + a_{jk}^2} + (\lambda_1+\lambda_3)\sum_{l=1}^{q}\sum_{l=1}^{r}\frac{b_{lk}^2}{\omega_0 + b_{lk}^2}\right)\right\} \\
\geq\ &\exp\left\{-\frac{1}{4\sqrt{\omega_0}}\left((\lambda_1+\lambda_2)\sum_{j=1}^{p_n}\sum_{k=1}^{r}|a_{jk}| + (\lambda_1+\lambda_3)\sum_{l=1}^{q}\sum_{l=1}^{r}|b_{lk}|\right)\right\}.
\end{aligned}$$

Define

$$\pi(\mathbf{A},\mathbf{B}) \propto \exp\left\{-\frac{1}{4\sqrt{\omega_0}}\left((\lambda_1+\lambda_2)\sum_{j=1}^{p_n}\sum_{k=1}^{r}|a_{jk}| + (\lambda_1+\lambda_3)\sum_{l=1}^{q}\sum_{l=1}^{r}|b_{lk}|\right)\right\}.$$

Then, according to Theorem 5.4, it is enough to show that

$$\Pi\left\{(\mathbf{A},\mathbf{B}) : \|\mathbf{A}\mathbf{B}^{\mathbf{T}} - \mathbf{C}^*\|_F < \frac{\Delta}{n^{\frac{\rho}{2}}}\right\} > \exp(-dn),$$

for sufficiently large n and any $d > 0$. Let $\mathcal{N}_r = \{j : (\mathbf{c}_j^*)^{\mathbf{T}}\mathbf{c}_j^* = 0, j = 1,\ldots,p_n\}$ be the index set of zero rows in $\mathbf{C}^*$. We can define $\mathbf{A}^* \in \mathbb{R}^{p_n\times r}$ and $\mathbf{B}^* \in \mathbb{R}^{q\times r}$ such that $\mathbf{A}^*(\mathbf{B}^*)^{\mathbf{T}} = \mathbf{C}^*$ and $(\mathbf{a}_j^*)^{\mathbf{T}}\mathbf{a}_j^* = 0$ for $j \in \mathcal{N}_r$, where $\mathbf{a}_j^*$ denotes the j^{th} row of $\mathbf{A}^*$.

Using the Cauchy-Schwarz inequality, we can show that

$$\begin{aligned}
&\Pi\left\{(\mathbf{A},\mathbf{B}) : \|\mathbf{A}\mathbf{B}^{\mathbf{T}} - \mathbf{C}^*\|_F < \frac{\Delta}{n^{\frac{\rho}{2}}}\right\}\\
&= \Pi\left\{(\mathbf{A},\mathbf{B}) : \|\mathbf{A}\mathbf{B}^{\mathbf{T}} - \mathbf{A}^*(\mathbf{B}^*)^{\mathbf{T}}\|_F < \frac{\Delta}{n^{\frac{\rho}{2}}}\right\}\\
&\geq \Pi\left\{(\mathbf{A},\mathbf{B}) : \|\mathbf{A}\mathbf{B}^{\mathbf{T}} - \mathbf{A}^*\mathbf{B}^{\mathbf{T}}\|_F + \|\mathbf{A}^*\mathbf{B}^{\mathbf{T}} - \mathbf{A}^*(\mathbf{B}^*)^{\mathbf{T}}\|_F < \frac{\Delta}{n^{\frac{\rho}{2}}}\right\}\\
&\geq \Pi\left\{(\mathbf{A},\mathbf{B}) : \|\mathbf{A}-\mathbf{A}^*\|_F\|\mathbf{B}\|_F + \|\mathbf{B}-\mathbf{B}^*\|_F\|\mathbf{A}^*\|_F < \frac{\Delta}{n^{\frac{\rho}{2}}}\right\}\\
&\geq \Pi\left\{(\mathbf{A},\mathbf{B}) : \|\mathbf{A}-\mathbf{A}^*\|_F\|\mathbf{B}\|_F < \frac{\Delta}{2n^{\frac{\rho}{2}}}, \|\mathbf{B}-\mathbf{B}^*\|_F\|\mathbf{A}^*\|_F < \frac{\Delta}{2n^{\frac{\rho}{2}}}\right\}.
\end{aligned}$$

Note that for given $\mathbf{B}^0$ such that $|b^0_{lk}| < \sup_{l,k}|b^*_{lk}|+1$ for $l = 1\ldots, q$ and $k = 1, \ldots, r$,

$$
\begin{aligned}
&\Pi\left\{\mathbf{A} : \|\mathbf{A}-\mathbf{A}^*\|_F < \frac{\Delta}{2n^{\frac{\rho}{2}}\|\mathbf{B}^0\|_F}\right\} \\
&\geq \Pi\left\{\mathbf{A} : \sum_{j=1}^{p_n}\sum_{k=1}^{r}(a_{jk}-a^*_{jk})^2 < \frac{\Delta^2}{4n^{\rho}qr(\sup_{l,k}|b^*_{lk}|+1)^2}\right\} \\
&\geq \Pi\left\{\mathbf{A} : \sum_{j\notin\mathcal{N}_r}\sum_{k=1}^{r}(a_{jk}-a^*_{jk})^2 < \frac{\Delta^2 p^*}{4n^{\rho}p_nqr(\sup_{l,k}|b^*_{lk}|+1)^2},\right. \\
&\quad \left.\sum_{j\in\mathcal{N}_r}\sum_{k=1}^{r}(a_{jk})^2 < \frac{\Delta^2(p_n-p^*)}{4n^{\rho}p_nqr(\sup_{l,k}|b^*_{lk}|+1)^2}\right\} \\
&\geq \left[\prod_{j\notin\mathcal{N}_r}\prod_{k=1}^{r}\Pi\left\{a_{jk} : |a_{jk}-a^*_{jk}| < \frac{\Delta}{2\sqrt{n^{\rho}p_nq}r(\sup_{l,k}|b^*_{lk}|+1)}\right\}\right] \\
&\quad \times\Pi\left\{(a_{jk}, j\in\mathcal{N}_c, k=1,\ldots,r) : \sum_{j\in\mathcal{N}_r}\sum_{k=1}^{r}(a_{jk})^2 < \frac{\Delta^2(p_n-p^*)}{4n^{\rho}p_nqr(\sup_{l,k}|b^*_{lk}|+1)^2}\right\}.
\end{aligned}
$$

Using the fact that $\pi(a_{jk}) = \frac{\lambda_1+\lambda_2}{8\sqrt{\omega_0}}\exp\left(-\frac{\lambda_1+\lambda_2}{4\sqrt{\omega_0}}|a_{jk}|\right)$ is a decreasing function in $|a_{jk}|$,

$$
\begin{aligned}
&\Pi\left\{a_{jk} : |a_{jk}-a^*_{jk}| < \frac{\Delta}{2\sqrt{n^{\rho}p_nq}r(\sup_{l,k}|b^*_{lk}|+1)}\right\} \\
&\geq \frac{\lambda_1+\lambda_2}{8\sqrt{\omega_0}}\frac{\Delta}{\sqrt{n^{\rho}p_nq}r(\sup_{l,k}|b^*_{lk}|+1)} \\
&\quad\times\exp\left\{-\frac{\lambda_1+\lambda_2}{4\sqrt{\omega_0}}\left(\sup_{j,k}|a^*_{jk}|+\frac{\Delta}{2\sqrt{n^{\rho}p_nq}r(\sup_{l,k}|b^*_{lk}|+1)}\right)\right\}.
\end{aligned}
\tag{A.12}
$$

Since $E^{\pi}(a_{ij}^2) = \frac{32\omega_0}{(\lambda_1+\lambda_2)^2}$, from the Markov's inequality, we have

$$\begin{aligned}
&\Pi\left\{(a_{jk}, j \in \mathcal{N}_c, k = 1, \ldots, r) : \sum_{j\in\mathcal{N}_r}\sum_{k=1}^{r}(a_{jk})^2 < \frac{\Delta^2(p_n - p^*)}{4n^{\rho}p_n qr(\sup_{l,k}|b_{lk}^*|+1)^2}\right\} \\
&\geq 1 - \frac{32\omega_0}{(\lambda_1+\lambda_2)^2}\frac{4n^{\rho}p_n qr^2(\sup_{l,k}|b_{lk}^*|+1)^2}{\Delta^2}. \qquad \text{(A.13)}
\end{aligned}$$

Similarly to (A.12), using $\pi(b_{lk}) = \frac{\lambda_1+\lambda_3}{8\sqrt{\omega_0}}\exp\left(-\frac{\lambda_1+\lambda_3}{4\sqrt{\omega_0}}|b_{lk}|\right)$, we have

$$\begin{aligned}
&\Pi\left\{\mathbf{B} : \|\mathbf{B}-\mathbf{B}^*\|_F < \frac{\Delta}{2n^{\frac{\rho}{2}}\|\mathbf{A}^*\|_F}\right\} \\
&\geq \Pi\left\{\mathbf{B} : \sum_{l=1}^{q}\sum_{k=1}^{r}(b_{lk}-b_{lk}^*)^2 < \frac{\Delta^2}{4n^{\rho}\|\mathbf{A}^*\|_F^2}\right\} \\
&\geq \prod_{l=1}^{q}\prod_{k=1}^{r}\Pi\left\{b_{lk} : |b_{lk}-b_{lk}^*| < \frac{\Delta}{2\sqrt{n^{\rho}qr}\|\mathbf{A}^*\|_F},\ |b_{lk}| < \sup_{l,k}|b_{lk}^*|+1\right\} \\
&= \prod_{l=1}^{q}\prod_{k=1}^{r}\Pi\left\{b_{lk} : |b_{lk}-b_{lk}^*| < \frac{\Delta}{2\sqrt{n^{\rho}qr}\|\mathbf{A}^*\|_F}\right\} \quad \text{for sufficiently large } n \\
&\geq \left(\frac{\lambda_1+\lambda_3}{8\sqrt{\omega_0}}\frac{\Delta}{\sqrt{n^{\rho}qr}\|\mathbf{A}^*\|_F}\right)^{qr}\exp\left\{-\frac{\lambda_1+\lambda_3}{4\sqrt{\omega_0}}\left(qr\sup_{l,k}|b_{lk}^*| + \frac{qr\Delta}{2\sqrt{n^{\rho}qr}\|\mathbf{A}^*\|_F}\right)\right\}. \qquad \text{(A.14)}
\end{aligned}$$

Due to the fact that (A.12) and (A.13) are free of $\mathbf{B}$, using (A.12)-(A.14), it is

straightforward to show that for sufficiently large n,

$$
\begin{aligned}
&\Pi\left\{(\mathbf{A},\mathbf{B}) : \|\mathbf{AB^T} - \mathbf{C}^*\|_F < \frac{\Delta}{n^{\frac{\rho}{2}}}\right\} \\
&\geq \int \left[\int \mathbf{1}\left\{\mathbf{A} : \|\mathbf{A}-\mathbf{A}^*\|_F\|\mathbf{B}\|_F < \frac{\Delta}{2n^{\frac{\rho}{2}}}\right\}\Pi(d\mathbf{A})\right] \\
&\quad \times \mathbf{1}\left\{\mathbf{B} : \|\mathbf{B}-\mathbf{B}^*\|_F\|\mathbf{A}^*\|_F < \frac{\Delta}{2n^{\frac{\rho}{2}}}\right\} \\
&\quad \times \mathbf{1}\left\{\mathbf{B} : |b_{lk}| < \sup_{l,k}|b^*_{lk}| + 1 \text{ for all } l,\, k\right\}\Pi(d\mathbf{B}) \\
&\geq \left(1 - \frac{32\omega_0}{(\lambda_1+\lambda_2)^2}\frac{4n^\rho p_n r^2(\sup_{l,k}|b^*_{lk}|+1)^2}{\Delta^2}\right) \\
&\quad \times \left(\frac{\lambda_1+\lambda_2}{8\sqrt{\omega_0}}\frac{\Delta}{\sqrt{n^\rho p_n}r(\sup_{l,k}|b^*_{lk}|+1)}\right)^{p^*r}\left(\frac{\lambda_1+\lambda_3}{8\sqrt{\omega_0}}\frac{\Delta}{\sqrt{n^\rho qr}\|\mathbf{A}^*\|_F}\right)^{qr} \\
&\quad \times \exp\left\{-\frac{\lambda_1+\lambda_2}{4\sqrt{\omega_0}}\left(p^*r\sup_{j,k}|a^*_{jk}| + \frac{p^*r\Delta}{2\sqrt{n^\rho p_n}r(\sup_{l,k}|b^*_{lk}|+1)}\right)\right\} \\
&\quad \times \exp\left\{-\frac{\lambda_1+\lambda_3}{4\sqrt{\omega_0}}\left(qr\sup_{l,k}|b_{lk}| + \frac{qr\Delta}{2\sqrt{n^\rho qr}\|\mathbf{A}^*\|_F}\right)\right\}. \qquad \text{(A.15)}
\end{aligned}
$$

Suppose that $\lambda_i = \delta_i\sqrt{n^\rho p_n}\log n$ with finite $\delta_i > 0$ for $i = 1, 2, 3$. Note that from assumption I, there exist finite constants ν_1, ν_2 and ν_3 such that $\sup_{j,k}|a^*_{jk}| < \nu_1$, $\|\mathbf{A}^*\|_F < \nu_2$ and $\sup_{l,k}|b_{lk}| < \nu_3$. Then, by taking the negative logarithm of (A.15),

for sufficiently large n,

$$
\begin{aligned}
&-\log \Pi \left\{ (\mathbf{A}, \mathbf{B}) : \|\mathbf{A}\mathbf{B}^{\mathbf{T}} - \mathbf{C}^*\|_F < \frac{\Delta}{n^{\frac{\rho}{2}}} \right\} \\
&\leq -\log \left(1 - \frac{32\omega_0}{(\delta_1+\delta_2)^2(\log n)^2} \frac{4r^2(\nu_3+1)^2}{\Delta^2} \right) \\
&\quad -p^* r \log \left(\frac{\delta_1+\delta_2}{8\sqrt{\omega_0}} \frac{\Delta}{r(\nu_3+1)} \right) - qr \log \left(\frac{(\delta_1+\delta_3)\sqrt{p_n}\log n}{8\sqrt{\omega_0}} \frac{\Delta}{\sqrt{qr}\nu_2} \right) \\
&\quad + \frac{\delta_1+\delta_2}{4\sqrt{\omega_0}} \frac{p^* r \Delta}{2r(\nu_3+1)} + \frac{(\delta_1+\delta_3)\sqrt{p_n}\log n}{4\sqrt{\omega_0}} \frac{\sqrt{qr}\Delta}{2\nu_2} \\
&\quad + \left(\frac{(\delta_1+\delta_2)}{4\sqrt{\omega_0}} p^* r \nu_1 + \frac{(\delta_1+\delta_3)}{4\sqrt{\omega_0}} qr\nu_3 \right) \sqrt{n^\rho p_n} \log n. \qquad \text{(A.16)}
\end{aligned}
$$

Note that (A.16) is dominated by the last term as $n \to \infty$. Hence, it implies that, for sufficiently large n, we have $-\log \Pi \left\{ (\mathbf{A}, \mathbf{B}) : \|\mathbf{A}\mathbf{B}^{\mathbf{T}} - \mathbf{C}^*\|_F < \frac{\Delta}{n^{\frac{\rho}{2}}} \right\} < dn$ for all $d > 0$, and this completes our proof.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669–679.

Alquier, P. (2013). Bayesian methods for low-rank matrix estimation: Short survey and theoretical study. In F. S. S. Jain, R. Munos, and T. Zeugmann (Eds.), *Algorithmic Learning Theory*, Volume 8139 of *Lecture Notes in Artificial Intelligence*, pp. 309–323. Springer-Verlag.

Amari, S. I. (2009). α–divergence is unique, belonging to both f–divergence and bregman divergence classes. *IEEE Transactions on Information Theory 55*, 4925–4931.

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics 22*, 327–351.

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

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society 68*, 337–404.

Babacan, S. D., M. Luessi, R. Molina, and A. K. Katsaggelos (2011). Low-rank matrix completion by variational sparse bayesian learning. In *IEEE International Conference on Audio, Speech and Signal Processing*, pp. 2188–2191.

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

Berger, J. O., B. Betró, E. Moreno, L. R. Pericchi, R. Ruggeri, G. Salinetti, and L. Wasserman (1988). In J. O. Berger, B. Betró, E. Moreno, L. R. Pericchi, R. Ruggeri, G. Salinetti, and L. Wasserman (Eds.), *Bayesian Robustness*. Hayward: IMS Lecture Notes–Monograph Seires.

Berger, J. O. and L. R. Pericchi (1996a). The intrinsic bayes factor for linear models. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 5*, pp. 25–44. Oxford: Oxford University Press.

Berger, J. O. and L. R. Pericchi (1996b). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association 91*, 109–122.

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

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B 48*, 259–302.

Boulesteix, A.-L. and K. Strimmer (2005). Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach. *Theoretical Biology and Medical Modelling 2*, 23.

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.

Breheny, P. and J. Huang (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface 2*, 369–380.

Brockwell, A. E. (2007). Universal residuals: A multivariate transformation. *Statistics & Probability Letters 77*, 1473–1478.

Bunea, F., Y. She, and M. Wegkamp (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics 39*(2), 1282–1309.

Bunea, F., Y. She, and M. Wegkamp (2012). Joint variable and rank selection for parsimonious estimation of high dimensional matrices. *Annals of Statistics 40*(5), 2359–2388.

Chen, K., K.-S. Chan, and N. C. Stenseth (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society Series B 74*(2), 203–221.

Chen, K., H. Dong, and K. S. Chan (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika 100*, 901–920.

Chen, L. and J. Z. Huang (2012). Sparse reduced–rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association 107*, 1533–1545.

Chen, M. H. (1994). Importance-weighted marginal bayesian posterior density estimation. *Journal of the American Statistical Association 89*, 818–824.

Chen, M. H. and Q. M. Shao (1999). Monte carlo estimation of bayesian credible and hpd intervals. *Journal of Computational and Graphical Statistics 8*, 69–92.

Chib, S. and E. Greenberg (1995). Understanding the metropolis-hastings algorithm. *The American Statistician 49*, 327–335.

Choi, T. and R. V. Ramamoorthi (2008). Remarks on consistency of posterior distributions. In S. Ghosal and B. Clarke (Eds.), *Pushing the limits of contemporary statistics: Contributions in honor of Jayanta K. Ghosh*, Volume 3 of *IMS Collections*, pp. 170–186.

Chun, H. and S. Keles (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society, Series B 72*, 3–25.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica 2*, 299–318.

Csiszár, I. (1995). Generalized projections for non-negative functions. *Acta Mathematica Hungarica 68*, 161–186.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series B 147*, 278–292.

Dey, D. K. and L. Birmiwal (1994). Robust bayesian analysis using entropy and divergence measures. *Statistics & Probability Letters 20*, 287–294.

Diaconis, P. and D. Freedman (1986). On the consistency of bayes estimates. *The Annals of Statistics 14*, 1–26.

Dicker, L., B. Huang, and X. Lin (2013). Variable selection and estimation with the seamless–l_0 penalty. *Statistica Sinica 23*, 929–962.

Eguchi, S. and Y. Kano (2001). Robustifying maximum likelihood estimation. Technical report, Institute of Statistical Mathematics, Tokyo, Japan.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

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

Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*, 101–148.

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

Frigyik, B. A., S. Srivastava, and M. R. Gupta (2008). Functional bregman divergence and bayesian estimation of distributions. *IEEE Transactions on Information Theory 54*, 5130–5139.

Gao, F., B. C. Foat, and H. J. Bussemaker (2004). Defining transcriptional networks through integrative modeling of mrna expression and transcription factor binding data. *BMC Bioinformatics 5*, 31.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association 70*, 320–328.

Geisser, S. and W. Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association 74*, 153–160.

Gelfand, A. E. and D. K. Dey (1991). On bayesian robustness of contaminated classes of priors. *Statistics and Decisions 9*, 63–80.

Gelfand, A. E. and D. K. Dey (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B 56*, 501–514.

Gelfand, A. E. and M. Ghosh (2000). Generalized linear models: A bayesian view. In D. K. Dey, S. K. Ghosh, and B. K. Mallick (Eds.), *Generalized Linear Models: A Bayesian Perspective*, pp. 3–22. New York: Marcel Dekker Press.

Gelfand, A. E., S. E. Hills, A. Racine-Poon, and A. F. M. Smith (1990). Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association 85*, 972–985.

Geweke, J. (1996). Bayesian reduced rank regression in econometrics. *Journal of Econometrics 75*, 121–146.

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

Ghosal, S., J. K. Ghosh, and T. Samanta (1995). On convergence of posterior distributions. *Annals of Statistics 23*, 2145–2152.

Ghosh, J. K., M. Delampady, and T. Samanta (2006). *An Introduction to Bayesian Analysis Theory and Methods*. New York: Springer.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B 69*, 243–268.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–378.

Goh, G., K. Chen, and D. K. Dey (2014). Bayesian sparse reduced rank multivariate regression. Technical Report 26, Department of Statistics, University of Connecticut, Storrs, CT, USA.

Goh, G. and D. K. Dey (2014). Bayesian model diagnostics using functional bregman divergence. *Journal of Multivariate Analysis 124*, 371–383.

Grünwald, P. D. and A. P. Dawid (2004). Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics 32*, 1367–1433.

Guttman, I. and D. Peña (1988). Outliers and influence: evaluation by posteriors of parameters in the linear model. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics 3*, pp. 631–640. Oxford: Oxford University Press.

Guttorp, P. and R. A. Lockhart (1988). Finding the location of a signal: A bayesian analysis. *Journal of the American Statistical Association 83*, 322–330.

Hennequin, R., B. David, and R. Badeau (2011). Beta-divergence as a subclass of bregman divergence. *IEEE Signal Processing Letters 18*, 83–86.

Huang, J., P. Breheny, and S. Ma (2012). A selective review of group selection in high dimensional models. *Statist. Sci. 27*(4), 481–499.

Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics 36*(2), 587–613.

Ibáñez, I., J. A. Silander Jr., J. A. Wilson, N. LaFleur, N. Tanaka, and I. Tsuyama (2009). Multivariate forecasts of potential distributions of invasive plant species. *Ecological Applications 19*, 359–375.

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics 17*, 295311.

Itakura, F. and S. Saito (1968). Analysis synthesis telephony based on the maximum likelihood method. In *Report of the 6th International Conference on Acoustics*.

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

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis 5*(2), 248–264.

Johnson, W. and S. Geisser (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *Journal of the American Statistical Association 78*, 137–144.

Kass, R. E. and S. Vaidyanathan (1992). Approximate bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society, Series B 54*, 129–144.

Kulis, B., M. A. Sustik, and I. S. Dhillon (2009). Low–rank kernel learning with bregman matrix divergences. *Journal of Machine Learning Research 10*, 341–376.

Kyung, M., J. Gilly, M. Ghosh, and G. Casella (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis 5*, 369–412.

Lee, T. I., N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thomson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young (2002). Transcriptional regulatory networks in saccharmomyces cerevisiae. *Science 298*, 799–804.

Li, Z. and C. Chan (2004). Extracting novel information from gene expression data. *Trends in Biotechnology 22*, 381–383.

Liao, J. C., R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury (2003). Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences of the United States of America 100*, 15522–15527.

Lim, Y. J. and Y. W. Teh (2007). Variational bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop*.

Ljung, G. M. and G. E. P. Box (1978). On a measure of a lack of fit in time series models. *Biometrika 65*, 297–303.

Luan, Y. and H. Li (2003). Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics 19*, 474–482.

Ma, Z. and T. Sun (2014, March). Adaptive Sparse Reduced-rank Regression. *ArXiv e-prints*.

McCulloch, R. E. (1989). Local model influence. *Journal of the American Statistical Association 84*, 473–478.

Mehrhoff, L. J., J. A. Silander Jr., S. A. Leicht, E. S. Mosher, and N. M. Tabak (2003). Ipane: invasive plant atlas of new england. Technical report, Department of Ecology & Evolutionary Biology, University of Connecticut, Storrs, CT.

Mukherjee, A. and J. Zhu (2011). Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining 4*(6), 612–622.

Negahban, S. and M. J. Wainwright (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics 39*(2), 1069–1097.

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

Peng, F. and D. K. Dey (1995). Bayesian analysis of outlier problems using divergence measures. *The Canadian Journal of Statistics 23*, 199–213.

Polson, N. G., J. G. Scott, and J. Windle (2014). The bayesian bridge. *Journal of the Royal Statistical Society, Series B 76*, 713–733.

Reinsel, G. C. and P. Velu (1998). *Multivariate reduced-rank regression: theory and applications*. New York: Springer.

Rohde, A. and A. Tsybakov (2011). Estimation of High-Dimensional Low-rank Matrices. *Annals of Statistics 39*(2), 887–930.

Salakhutdinov, R. and A. Mnih (2008). Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Smith, J. Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting 4*, 283–291.

Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, B. D, and B. Futcher (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell 9*, 3273–3297.

Stephens, M. A. (1969). Techniques for directional data. Technical Report 150, Department of Statistics, Stanford University, Stanford, CA.

Stone, M. (1974). Cross-validation choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B 36*, 111–147.

Taskar, B., S. Lacoste-Julien, and M. I. Jordan (2006). Structured prediction, dual extragradient and bregman projections. *Journal of Machine Learning Research 7*, 1627–1653.

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

Vemuri, B. C., M. Liu, S. I. Amari, and F. Nielsen (2011). Total bregman divergence

and its applications to dti analysis. *IEEE Transactions on Medical Imaging 30*, 475–483.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia, PA: CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics.

Wang, L., G. Chen, and H. Li (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics 23*, 1486–1494.

Wang, X. and D. K. Dey (2010). Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption. *Annals of Applied Statistics 4*, 2000–2023.

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

Yeung, K. Y., C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics 17*, 977–987.

Yuan, M., A. Ekici, Z. Lu, and R. Monteiro (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society, Series B 69*, 329–346.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B 68*(1), 49–67.

Zhang, C., Y. Jiang, and Y. Chai (2010). Penalized bregman divergence for large-dimensional regression and classification. *Biometrika 97*, 551–566.

Zhang, C., Y. Jiang, and Z. Shang (2009). New aspects of bregman divergence in regression and classification with parametric and nonparametric estimation. *The Canadian Journal of Statistics 37*, 119–139.

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

Zhou, M., C. Wang, M. Chen, J. Paisley, D. Dunson, and L. Carin (2010). Nonparametric bayesian matrix completion. In *IEEE Sensor Array and Multichannel Signal Processing Workshop*, pp. 213–216.

Zhu, H., Z. Khondker, Z. Lu, and J. G. Ibrahim (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association 109*, 997–990.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.