VARIATIONAL BAYES INFERENCE OF ISING MODELS AND THEIR APPLICATIONS

By

Minwoo Kim

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics – Doctor of Philosophy

2022

ABSTRACT

VARIATIONAL BAYES INFERENCE OF ISING MODELS AND THEIR APPLICATIONS

By

Minwoo Kim

Ising models originated in statistical physics have been widely used in modeling spatial data and computer vision problems. However, statistical inference of this model and its application to many practical fields remain challenging due to intractable nature of the normalizing constant in the likelihood. This dissertation consists of two main themes, (1) parameter estimation of Ising model and (2) structured variable selection based on the Ising model using variational Bayes (VB).

In Chapter 1, we review the background, research questions and development of Ising model, variational Bayes, and other statistical concepts. An Ising model basically deal with a binary random vector in which each component is dependent on its neighbors. There exist various versions of Ising model depending on parameterization and neighboring structure. In Chapter 2, with two-parameter Ising model, we describe a novel procedure for the parameter estimation based on VB which is computationally efficient and accurate compared to existing methods. Traditional pseudo maximum likelihood estimate (PMLE) can provide accurate results only for smaller number of neighbors. A Bayesian approach based on Markov chain Monte Carlo (MCMC) performs better even with a large number of neighbors. Computational costs of MCMC, however, are quite expensive in terms of time. Accordingly, we propose a VB method with two variational families, mean-field (MF) Gaussian family and bivariate normal (BN) family. Extensive simulation studies validate the efficacy of the families. Using our VB methods, computing times are remarkably decreased without deterioration in performance accuracy, or in some scenarios we get much more accurate output. In addition, we demonstrates theoretical properties of the proposed VB method under MF family. The main theoretical contribution of our work lies in establishing the consistency of

the variational posterior for the Ising model with the true likelihood replaced by the pseudo-likelihood. Under certain conditions, we first derive the rates at which the true posterior based on the pseudo-likelihood concentrates around the ε_n- shrinking neighborhoods of the true parameters. With a suitable bound on the Kullback-Leibler distance between the true and the variational posterior, we next establish the rate of contraction for the variational posterior and demonstrate that the variational posterior also concentrates around ε_n-shrinking neighborhoods of the true parameter.

In Chapter 3, we propose a Bayesian variable selection technique for a regression setup in which the regression coefficients hold structural dependency. We employ spike and slab priors on the regression coefficients as follows: (i) In order to capture the intrinsic structure, we first consider Ising prior on latent binary variables. If a latent variable takes one, the corresponding regression coefficient is active, otherwise, it is inactive. (ii) Employing spike and slab prior, we put Gaussian priors (slab) on the active coefficients and inactive coefficients will be zeros with probability one (spike).

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisors Dr. Tapabrata Maiti and Dr. Shrijita Bhattacharya for their guidance toward my researches and my life in US. They have always helped me make great progress with their insightful suggestions. I would also like to extend my appreciation to my dissertation committee members, Dr. Yimin Xiao and Dr. Jiliang Tang. Their feedback and comments are really beneficial for my research. I am also grateful to my previous advisor Dr. Chaeyoung Lim and my future advisor Dr. Marc G. Genton for giving me wonderful opportunities and encouraging me all the time. I made a lot of friends and I have never felt lonely because of them. I really appreciate my friends. Last but not least, I would like to express my sincere thanks to my family for their support and concerns.

During my five years at Michigan State University, I have learned a lot from the courses and seminars and I am able to apply the knowledge I learned to my research. Thanks to all professors and friends and family, I am incredibly confident in my academic success. I am very much looking forward to my life as a statistician.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

CHAPTER 1

INTRODUCTION

In this Chapter, we briefly introduce basic concepts without complete details related to our work, Ising model and its pseudo-likelihood (Section 1.1), variational Bayes (Section 1.2), Adam learning rates (Section 1.3), variable selection (Section 1.4), and posterior consistency (Section 1.5). In addition, previous studies that have produced significant results in those fields are introduced.

1.1 Ising model

A popular way of modeling a binary vector $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$, $x_i \in \{-1, 1\}$, in which elements are pairwise dependent is to take advantage of Ising model named after the physicist Ernst Ising Ising (1924) which has been used in a wide range of applications including spatial data analysis and computer vision. Many different versions of Ising model have emerged in the literature. Among them, the probability mass function of a general Ising model is of the form:

$$\mathbb{P}^{(n)}(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{Z_n(K_n, B_1, \ldots, B_n)} \exp\left(\boldsymbol{x}^\top K_n \boldsymbol{x} + \sum_{i=1}^{n} B_i x_i\right),$$

where $K_n \in \mathbb{R}^{n \times n}$ and $(B_1, \ldots, B_n) \in \mathbb{R}^n$ are model parameters and $Z_n(K_n, B_1, \ldots, B_n)$ is the normalizing constant that makes the sum of the probability mass function over all possible combinations of $\boldsymbol{x}$ equal to 1:

$$Z_n(K_n, B_1, \ldots, B_n) = \sum_{\boldsymbol{x} \in \{-1,1\}^n} \exp\left(\boldsymbol{x}^\top K_n \boldsymbol{x} + \sum_{i=1}^{n} B_i x_i\right).$$

The general Ising model can be reduced to two-parameter Ising model, assuming that all nonzero entries of K_n take the same value and $B_i = B$ for all i, which has an inverse temperature parameter $\beta > 0$ (also known as interaction parameter) and a magnetization parameter $B \neq 0$ (also known as threshold parameter). With a specified symmetric coupling

matrix $A_n \in \mathbb{R}^{n\times n}$, a probability mass function of two-parameter Ising model is:

$$\mathbb{P}^{(n)}_{\beta,B}(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{Z_n(\beta, B)} \exp\left(\frac{\beta}{2}\boldsymbol{x}^\top A_n \boldsymbol{x} + B\sum_{i=1}^{n} x_i\right), \tag{1.1}$$

In the two-parameter Ising model, β characterizes the strength of interactions among x_i's and B represents external influence on $\boldsymbol{x}$. In the first place, Ising model has been introduced for the relations between atom spins with the domain $\{-1, 1\}^n$ (Brush, 1967). While we work with the domain $\{-1, 1\}^n$, in many current applications, Ising model has been defined with another domain $\{0, 1\}^n$. One can read Haslbeck et al. (2021) for more details on two different domains.

An Ising model is usually represented by an undirected graph. Consider an undirected graph which has n vertices (nodes) x_i, $i = 1, \ldots, n$. Each vertex of the graph takes a value either -1 or 1, i.e., $x_i \in \{-1, 1\}$, and let $\mathcal{E} = \{(i, j) \mid i \sim j,\ 1 \le i, j \le n\}$ represent the set of edges in the graph where $i \sim j$ denotes that the vertices i and j are connected. Then, one common choice of the coupling matrix A_n is a scaled adjacency matrix of the underlying graph whose all diagonal elements are zeros and the other elements are non-negative:

Definition 1 (Scaled adjacency matrix)**.** *A scaled adjacency matrix for a graph G_n with n vertices is defined as:*

$$A_n(i, j) := \begin{cases} \frac{n}{2|G_n|} & \text{if } (i, j) \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases},$$

where $|G_n|$ denotes the number of edges in the graph G_n.

For a simple example, consider three nodes (x_1, x_2, x_3) and two edges between x_1 and x_2, and between x_2 and x_3 as in Figure 1.1. Then, $n = 3$, $|G_n| = 2$, and the scaled adjacency matrix is:

$$A_n = \begin{pmatrix} 0 & 0.75 & 0 \\ 0.75 & 0 & 0.75 \\ 0 & 0.75 & 0 \end{pmatrix}.$$

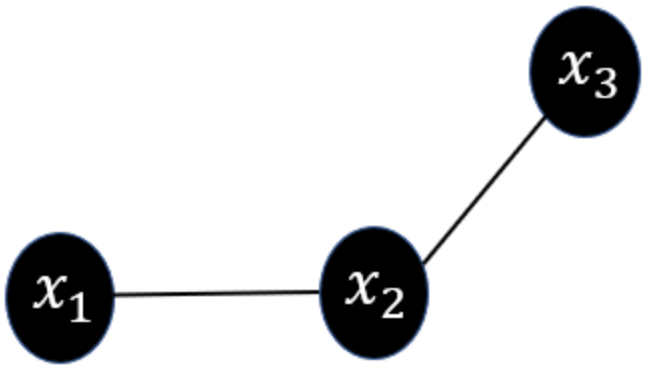

Figure 1.1: An undirected graph with three nodes

1.1.1 Pseudo-likelihood

One of the largest challenges in using Ising model is the unknown normalizing constant $Z_n(\beta, B)$ in the likelihood (1.1):

$$Z_n(\beta, B) = \sum_{\boldsymbol{x}\in\{-1,1\}^n} \exp\left(\frac{\beta}{2}\boldsymbol{x}^\top A_n \boldsymbol{x} + B\sum_{i=1}^{n} x_i\right)$$

One can notice that the exact calculation of the normalizing constant involves sum of 2^n terms, which is available only for small n. Due to the intractable nature of the normalizing constant, standard statistical methodologies based on the true likelihood are infeasible. One way to approximate the normalizing constant is importance sampling, see Geyer (1994); Gelman and Meng (1998); Molkaraie (2014). Another approach to handling the normalizing constant is to use a pseudo-likelihood. The conditional probability of x_i is easily calculated because x_i is binary:

$$\mathbb{P}^{(n)}_{\beta,B}\big(X_i = 1 | X_j, j \neq i\big) = \frac{e^{\beta m_i(\boldsymbol{x})+B}}{e^{\beta m_i(\boldsymbol{x})+B} + e^{-\beta m_i(\boldsymbol{x})-B}},$$

where $m_i(\boldsymbol{x}) = \sum_{j=1}^{n} A_n(i,j)x_j$. The pseudo-likelihood of Ising model corresponding to the true likelihood in (1.1), is defined as the product of one dimensional conditional distributions:

$$\begin{aligned}
&\prod_{i=1}^{n} \mathbb{P}_{\beta,B}^{(n)} \left(X_i = x_i \mid X_j, j \neq i\right) \\
&= 2^{-n} \exp\left(\sum_{i=1}^{n} \left(\beta x_i m_i(\boldsymbol{x}) + B x_i - \log\cosh(\beta m_i(\boldsymbol{x}) + B)\right)\right). \qquad (1.2)
\end{aligned}$$

Fauske (2009) is an example on empirical study using pseudo-likelihood and Ghosal et al. (2020) provides theoretical justification on use of pseudo-likelihood. For $\boldsymbol{v}, \boldsymbol{w} \in [-1,1]^n$, defining a function g as:

$$g(\boldsymbol{v}, \boldsymbol{w}) = \sum_{i=1}^{n} \frac{1+v_i}{2} \log \frac{1+w_i}{2} + \frac{1-v_i}{2} \log \frac{1-w_i}{2},$$

we point out that the pseudo-likelihood can be written as:

$$\prod_{i=1}^{n} \mathbb{P}_{\beta,B}^{(n)} \left(X_i = x_i \mid X_j, j \neq i\right) = e^{g(\boldsymbol{x}, b(\boldsymbol{x}))}, \qquad (1.3)$$

where $b(\boldsymbol{x}) = (b_1(\boldsymbol{x};\theta), \cdots, b_n(\boldsymbol{x};\theta))^\top$ and

$$b_i(\boldsymbol{x}) = E(X_i \mid X_j, j \neq i) = \tanh(\beta m_i(\boldsymbol{x}) + B),\ i = 1, \ldots, n.$$

1.2 Variational Bayes (VB)

Let θ be the set of parameters of interest with a prior distribution $p(\theta)$. In a Bayesian inference, a typical objective is to obtain posterior distribution given data $\mathcal{D}$. We can derive the exact posterior distribution $\pi(\theta \mid \mathcal{D})$ using Bayes' theorem:

$$\pi(\theta \mid \mathcal{D}) = \frac{p(\theta)p(\mathcal{D} \mid \theta)}{\int_\theta p(\theta)p(\mathcal{D} \mid \theta)d\theta}, \qquad (1.4)$$

where $p(\mathcal{D} \mid \theta)$ is a likelihood given parameter θ. The exact posterior in (1.4), however, is not typically available except for a few well-known examples. To get an approximated posterior, many statisticians have widely used sampling based Markov chain Monte Carlo (MCMC) methods but it is hardly scalable to high-dimensional cases. Beyond sampling methods,

variational Bayes (VB) also called variational inference or variational approximation (Jordan et al., 1999) has been popularized as an efficient alternative to MCMC. VB recasts the sampling problem as an optimization problem minimizing Kullback-Leibler (KL) divergence between a surrogate distribution (called a variational distribution) and the true posterior distribution (Blei et al., 2017).

Definition 2 (Kullback-Leibler (KL) divergence)**.** *For two probability measures P_1 and P_2 over a set Ω, the KL divergence between P_1 and P_2 is defined as:*

$$KL\left(P_1, P_2\right) = \mathbb{E}_{p_1}\left(\log p_1 - \log p_2\right) \tag{1.5}$$

$$= \int_{\Omega} \log\left(\frac{p_1(\omega)}{p_2(\omega)}\right) dP_1 \tag{1.6}$$

where p_1 and p_2 are corresponding densities to P_1 and P_2 respectively.

As the first step in building a VB algorithm, we need to define a family of distributions (called variational family) denoted by $\mathcal{Q}$ which contains candidates of the best approximation to the true posterior (1.4):

$$\mathcal{Q} = \{q(\theta; \boldsymbol{\nu}) : q \text{ is a probability density function that can be easily handled.}\},$$

where ν is a set of parameters (called variational parameters) that characterize variational distributions. For instance, if $\mathcal{Q}$ is a Gaussian family, then $\boldsymbol{\nu}$ includes mean (μ) and standard deviation (σ) of a Gaussian distribution. After an appropriate variational family is chosen, VB seeks the best surrogate function (called variational posterior) by minimizing KL divergence with the true posterior:

$$q^* = \underset{q \in \mathcal{Q}}{\arg\min}\ KL\left(Q, \Pi\left(\cdot \mid \mathcal{D}\right)\right). \tag{1.7}$$

Observe that:

$$\begin{aligned} KL\left(Q, \Pi\left(\cdot \mid \mathcal{D}\right)\right) &= \mathbb{E}_q\left(\log q(\theta) - \log \pi\left(\theta \mid \mathcal{D}\right)\right) \\ &= -\mathbb{E}_q\left(\log p\left(\theta, \mathcal{D}\right) - \log q(\theta)\right) + \log m\left(\mathcal{D}\right), \end{aligned} \tag{1.8}$$

where $m(\mathcal{D})$ is the marginal distribution of data which does not depend on θ. So, we can find the minimizer of KL divergence by maximizing $\mathbb{E}_q(\log p(\theta, \mathcal{D}) - \log q(\theta))$ which is called evidence lower bound (ELBO).

1.2.1 Black box variational inference (BBVI)

Black box variational inference (BBVI) is an stochastic gradient optimization technique suggested by Ranganath et al. (2014) to maximize ELBO using unbiased gradients. Consider the ELBO as a function of variational parameters $\boldsymbol{\nu}$:

$$\mathcal{L}(\boldsymbol{\nu}) = \mathbb{E}_q(\log p(\theta, \mathcal{D}) - \log q(\theta)). \tag{1.9}$$

In each BBVI iteration, a variational parameter $\nu \in \boldsymbol{\nu}$ is updated in the direction that the objective function $\mathcal{L}(\nu)$ increases as follows:

$$\nu^{(t+1)} \leftarrow \nu^{(t)} + \eta_t \nabla_\nu \mathcal{L}(\boldsymbol{\nu}), \tag{1.10}$$

where $\nu^{(t)}$ is the variational parameter at t-th iteration and η_t's, $t = 1, 2, \ldots$, are learning rates which satisfy Robbin-Monro conditions (Robbins and Monro, 1951):

$$\sum_{t=1}^{\infty} \eta_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty.$$

A closed form of the gradient $\nabla_\nu \mathcal{L}(\boldsymbol{\nu})$ is not always available. Ranganath et al. (2014) proposed an unbiased Monte Carlo estimate. Observe that:

$$\begin{aligned}
\nabla_\nu \mathcal{L}(\boldsymbol{\nu}) &= \nabla_\nu \int q(\theta)(\log p(\theta, \mathcal{D}) - \log q(\theta))\, d\theta \\
&= \int \nabla_\nu q(\theta)(\log p(\theta, \mathcal{D}) - \log q(\theta))\, d\theta - \int q(\theta) \nabla_\nu \log q(\theta) d\theta \\
&= \int \nabla_\nu q(\theta)(\log p(\theta, \mathcal{D}) - \log q(\theta))\, d\theta \\
&= \int q(\theta) \nabla_\nu \log q(\theta)(\log p(\theta, \mathcal{D}) - \log q(\theta))\, d\theta \\
&= \mathbb{E}_q(\nabla_\nu \log q(\theta)(\log p(\theta, \mathcal{D}) - \log q(\theta)))
\end{aligned} \tag{1.11}$$

The third equality is the fact that the expectation of a score function is zero. From the last expectation form in (1.11), we can induce a Monte Carlo estimate as follow:

$$\widehat{\nabla_\nu \mathcal{L}(\boldsymbol{\nu})} = \frac{1}{S}\sum_{s=1}^{S} \nabla_\nu \log q(\theta_s) \left(\log p\left(\theta_s, \mathcal{D}\right) - \log q(\theta_s)\right), \tag{1.12}$$

where θ_s is a draw from the current $q(\theta; \boldsymbol{\nu})$. Also, we define an empirical ELBO as the Monte Carlo estimate:

$$\widehat{\mathcal{L}(\boldsymbol{\nu})} = \frac{1}{S}\sum_{s=1}^{S} \left(\log p\left(\theta_s, \mathcal{D}\right) - \log q(\theta_s)\right), \tag{1.13}$$

Replacing $\nabla_\nu \mathcal{L}(\boldsymbol{\nu})$ in (1.10) with the unbiased estimate in (1.12), even though the closed forms of ELBO and its gradients are not exactly computable, we can update variational parameters until the empirical ELBO (1.13) converges.

1.3 Adaptive learning rates

To optimize the objective function (1.9), appropriate learning rates η_t in (1.10) can vary widely between variational parameters. Instead of a single learning rate, one can use an adaptive learning rate method. Kingma and Ba (2014) proposed an algorithm for first-order gradient-based optimization based on adaptive estimates of lower-order moments named Adam. Adam computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients. Algorithm 1.1 outlines the Adam algorithm.

Adam algorithm updates exponential moving averages of the gradient (m_t) and the squared gradient (u_t) where $\alpha_1, \alpha_2 \in [0, 1)$ control the exponential decay rates of these moving averages. Kingma and Ba (2014) suggested good default settings $\alpha_0 = 0.001$, $\alpha_1 = 0.9$, $\alpha_0 = 0.999$, and $\epsilon = 10^{-8}$. We use Adam learning rates in Chapter 3 for our variable selection algorithm.

Algorithm 1.1 Adam learning rates.

Initialize: α_0: Stepsize
Initialize: $\alpha_1, \alpha_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Initialize: The hyper parameters $m_0 \leftarrow 0$ and $u_0 \leftarrow 0$

while ELBO increases **do**
  Draw $\theta^{(s)} \sim q\left(\theta; \boldsymbol{\nu}^{(t)}\right)$, $s = 1, \ldots, S$;
  Get $\widehat{\nabla_\nu \mathcal{L}(\boldsymbol{\nu})}$ based on the S sample points;
  $m_t \leftarrow \alpha_1 m_{t-1} + (1 - \alpha_1) \cdot \left(\widehat{\nabla_\nu \mathcal{L}(\boldsymbol{\nu})}\right)$;
  $u_t \leftarrow \alpha_2 u_{t-1} + (1 - \alpha_2) \cdot \left(\widehat{\nabla_\nu \mathcal{L}(\boldsymbol{\nu})}\right)^2$;
  $\hat{m}_t \leftarrow m_t / (1 - \alpha_1^t)$, where α_1^t denotes α_1 to the power of t;
  $\hat{u}_t \leftarrow u_t / (1 - \alpha_2^t)$, where α_2^t denotes α_2 to the power of t;
  Update the parameter of interest: $\nu^{(t)} \leftarrow \nu^{(t-1)} + \alpha_0 \cdot \hat{m}_t / \left(\sqrt{\hat{u}} + \epsilon\right)$;
end while

Output: Optimal variational parameters $\boldsymbol{\nu}^*$

1.4 Variable selection

Statistical methodologies are well-established if the set of variables to consider is fixed and small. However, in high-dimensional setup in which the number of covariates (p) are much larger than the number of observations (n), classical models do not perform well, which leads statisticians to develop various variable selection methods. Variable selection in statistics means selecting among many variables which to include in a statistical model. From a total list of variables, significant variables are selected by removing irrelevant or redundant variables. In this section, we introduce a Bayesian variable selection method based on a spike and slab prior which was first suggested by Mitchell and Beauchamp (1988).

In a linear regression model, consider a sparse vector of the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^{p_n}$:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.14}$$

where $p > n$, $\boldsymbol{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $\boldsymbol{e} \in \mathbb{R}^n$. We assume that only a few number of β_i is nonzero and define an activation set of nonzero coefficients:

$$\mathcal{A} := \{\beta_i : \beta_i \neq 0\}. \tag{1.15}$$

To select explanatory variables corresponding to the nonzero coefficients, it is desirable that a tall and narrow function around zero is assigned to β_i's in $\mathcal{A}^c$. We call the tall and narrow function a "Spike" distribution. Whereas, for β_i's in $\mathcal{A}$, we use a flatter and diffused function called a "Slab" distribution. Plus, given that it is not known a priori which covariates should be included in a model, we introduce a latent binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top \in \{-1, 1\}^{p_n}$. If $\gamma_i = -1$, a spike distribution is used as a prior of β_i. If $\gamma_i = 1$, for a prior of β_i, a slab distribution is used as follows:

$$p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) = \prod_{i=1}^{p} p(\beta_i \mid \gamma_i),$$
$$\beta_i \mid \gamma_i \sim \frac{1-\gamma_i}{2} f_1(\beta_i) + \frac{1+\gamma_i}{2} f_2(\beta_i),$$

where f_1 and f_2 denote a spike distribution and a slab distribution respectively. One simple choice of a prior distribution of $\boldsymbol{\gamma}$ is independent Bernoulli:

$$p(\boldsymbol{\gamma}) = \prod_{i=1}^{p} \phi_i^{(1+\gamma_i)/2} (1-\phi_i)^{(1-\gamma_i)/2}, \tag{1.16}$$

where ϕ_i is the probability that $\gamma_i = 1$.

Various versions of spike and slab priors have emerged in the past literature. Mitchell and Beauchamp (1988) used a mixture of a point mass at zero (spike) and a diffuse uniform distribution (slab). With a point mass spike distribution, many previous studies used a Gaussian slab distribution which include but not limited to Li and Zhang (2010), Andersen et al. (2014), Xi et al. (2016) and Andersen et al. (2017). Another group of previous researches which includes but not limited to Johnstone and Silverman (2004), Castillo and Roquain (2020), and Ray and Szabó (2021) used a Laplace (double exponential) slab with a point mass spike. Two continuous distributions have been also considered as spike and slab distributions. George and McCulloch (1993) used a mixture Gaussian prior as follows:

$$\beta_i \mid \gamma_i \sim \frac{1-\gamma_i}{2} N(0, \tau_i^2) + \frac{1+\gamma_i}{2} N(0, c_i^2 \tau_i^2),$$

where c_i is set to be large. Park et al. (2022) used similar mixture Gaussian priors with applications to educational data. In a series of studies Ročková and George (2016), Ročková

(2018), and Ročková and George (2018), they have developed spike and slab LASSO priors using two Laplace distributions as follows:

$$p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) = \prod_{i=1}^{p} \left(\frac{1-\gamma_i}{2} \right) \frac{\lambda_0}{2} \exp\left(-\lambda_0 |\beta_i|\right) + \left(\frac{1+\gamma_i}{2} \right) \frac{\lambda_1}{2} \exp\left(-\lambda_1 |\beta_i|\right),$$

where λ_0 is a large scale parameter and λ_1 is a smaller scale parameter. Under the spike and slab LASSO priors, Gan et al. (2019) proposed an approach for precision matrix estimation called BAGUS, short for "Bayesian regularization for Graphical models with Unequal Shrinkage".

We can simply use the independent Bernoulli prior distributions on $\boldsymbol{\gamma}$ in (1.16) assuming that there is no inter-dependence between covariates or between regression coefficients. Dependent data, however, are now routinely analyzed. Some previous researches have used combination of an Ising model and a spike and slab prior to facilitate the catch of dependence. Li and Zhang (2010) considered Ising prior on $\boldsymbol{\gamma}$ and suggested a MCMC method with known structure among the covariates. Li et al. (2015) proposed a joint Ising and DiriChlet process for grouping and selecting significant voxels in functional magnetic resonance imaging (fMRI) data. In Chapter 3, using a spike and slab prior with Ising model on $\boldsymbol{\gamma}$, we will describe a VB algorithm for simultaneously selecting significant explanatory variables and estimating the regression coefficients when there exists structural dependence among the regression coefficients.

1.5 Posterior consistency

Establishing posterior consistency with contraction rates of a statistical method has been a fundamental research topic for a Bayesian to provide theoretical justification. In an estimation problem, the basic idea of posterior consistency is that the posterior distribution is concentrated around the true parameter. Many previous studies have constructed a theoretical basis for Bayesian methodology by dealing with posterior consistency. For example, in a high-dimensional linear regression setup, the posterior consistency of the regression coefficients with popular shrinkage priors under mild conditions was proved. More specifically, in

the high-dimensional linear model, $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$, posterior distribution satisfies:

$$\pi\left(\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| > \epsilon \mid \boldsymbol{y}\right) \to 0,$$

where ϵ is any positive constant and $\boldsymbol{\beta}_0$ are the true regression coefficients (See Armagan et al. (2013) for detailed statements and proof). As other examples, Sriram et al. (2013) provided a justification to use of (misspecified) asymmetric Laplace density for the response in Bayesian quantile regression by proving posterior consistency. More recently, Ghosh et al. (2018) considered a VAR model with two priors for the coefficient matrix and showed posterior consistency. Cao et al. (2019), for a covariance estimation and selection problem, established strong graph selection consistency and posterior convergence rates for estimation using Gaussian directed acyclic graph model. For VB, there are a few theoretical results. Wang and Blei (2019) established frequentist consistency and asymptotic normality of VB methods by proving a variational Bernstein–von Mises theorem. Bhattacharya and Maiti (2021) established the mean-field variational posterior consistency for a feed-forward artificial neural network model.

In our two-parameter Ising model estimation problem, we say the posterior is consistent if

$$\pi\left(\theta \in \mathcal{N}_n^c \mid \boldsymbol{X}^{(n)}\right) \to 0 \text{ in } \mathbb{P}_0^{(n)} \text{ probability},$$

where $\mathcal{N}_n$ is a shrinking neighborhood of the true parameter (β_0, B_0), $\mathbb{P}_0^{(n)}$ is the distribution induced by the true likelihood (1.1) with (β_0, B_0), and $\boldsymbol{X}^{(n)}$ is a sample vector from the Ising model with (β_0, B_0). Analogous to the true posterior, we say the variational posterior is consistent if

$$q^*\left(\theta \in \mathcal{N}_n^c\right) \to 0 \text{ in } \mathbb{P}_0^{(n)} \text{ probability}.$$

In Chapter 3, we derive variational posterior consistency with contraction rates obtained under the pseudo-likelihood and Gaussian mean-field variational family.

CHAPTER 2

A VARIATIONAL BAYES ALGORITHM AND POSTERIOR CONSISTENCY FOR TWO-PARAMETER ISING MODEL ESTIMATION

In this chapter, we describe a VB algorithm for Ising model estimation under pseudo-likelihood along with numerical studies for assessing performances.

2.1 Introduction

Estimation of Ising model parameters has received considerable attention in statistics and computer science literature. The existing literature can be broadly divided into two groups. Some literature assume that i.i.d. (independently and identically distributed) copies of data are available for inference; see Anandkumar et al. (2012), Bresler (2015), Lokhov et al. (2018), Ravikumar et al. (2010), and Xue et al. (2012). Another category of literature assumes that only one sample is observable; see Bhattacharya et al. (2018), Chatterjee et al. (2007), Comets (1992), Comets and Gidas (1991), Ghosal et al. (2020), Gidas (1988), and Guyon and Künsch (1992). In this dissertation, using variational Bayes (VB), we provide a new Bayesian methodology for model parameter estimation when one observes the data only once. Under the assumption of only one observation, Comets and Gidas (1991) showed that the MLE of $\beta > 0$ for Curie-Weiss model is consistent if $B \neq 0$ is known, and vice versa. They also proved that the joint MLE does not exist when neither β nor B is given. In this regard, Ghosal et al. (2020) addressed joint estimation of (β, B) using pseudo-likelihood and showed that the pseudo-likelihood estimator is consistent under some conditions on coupling matrix A_n. We also assume only one observation of $\boldsymbol{x}$ and provide a variational Bayes algorithm for model parameter estimation with its posterior consistency.

One of the main challenges in the Bayesian estimation of Ising models lies in the intractable nature of the normalizing constant in the likelihood. Following the works of Ghosal et al. (2020), Bhattacharya et al. (2018) and Okabayashi et al. (2011), we replace the true

likelihood of the Ising model by a pseudo-likelihood and we establish that the posterior based on the pseudo-likelihood is consistent for a suitable choice of the prior distribution. Further, we use variational Bayes (VB) approach which has recently become a popular and computationally powerful alternative to MCMC. In order to approximate the unknown posterior distribution using VB, we propose a Gaussian mean field family and general bivariate normal family with transformation of the parameters to $(\log \beta, B)$. For implementation of VB, we consider a black box variational inference (BBVI), Ranganath et al. (2014). In BBVI, we need to be able to evaluate the likelihood to compute the gradient estimates, but the existence of an unknown normalizing constant in likelihood of Ising model prevents us using BBVI directly. So, as mentioned above, we use pseudo-likelihood as in Ghosal et al. (2020). Replacing the true likelihood of Ising model with pseudo-likelihood, we are able to compute all the quantities needed for implementing BBVI. Our VB algorithm based on optimization is computationally more powerful than the sampling based MCMC methods (Møller et al., 2006). Also, use of PyTorch's automatic differentiation enables us to further reduce computational costs.

2.2 VB algorithm

Let $\theta = (\beta, B)$ be the parameter set in a two-parameter Ising model. To develop variational Bayes algorithm, we consider the following independent prior distribution $p(\theta) = p_\beta(\beta)p_B(B)$, with $p_\beta(\beta)$ as a log-normal prior for β and $p_B(B)$ as a normal prior for B as follows:

$$p_\beta(\beta) = \frac{1}{\beta\sqrt{2\pi}} e^{-\frac{(\log \beta)^2}{2}}, \quad p_B(B) = \frac{1}{\sqrt{2\pi}} e^{-\frac{B^2}{2}}. \tag{2.1}$$

The assumption of log-normal prior on β is to ensure the positivity of β. Let $L(\theta)$ be the pseudo-likelihood function given by (1.2), then the above prior structure leads to the following posterior distribution

$$\Pi(\mathcal{A} \mid X^{(n)}) = \frac{\int_{\mathcal{A}} \pi(\theta, X^{(n)}) d\theta}{m\left(X^{(n)}\right)} = \frac{\int_{\mathcal{A}} L(\theta)p(\theta)d\theta}{\int L(\theta)p(\theta)d\theta}, \tag{2.2}$$

for any set $\mathcal{A} \subseteq \Theta$ where Θ denotes the parameter space of θ. Note, $\pi(\theta, X^{(n)})$ is the joint density of θ and the data $X^{(n)}$ and $m\left(X^{(n)}\right) = \int L(\theta)p(\theta)d\theta$ is the marginal density of $X^{(n)}$ which is free from the parameter set θ. Next, we provide a variational approximation to the posterior distribution (2.2) considering two choices of the variational family in order to obtain approximated posterior distribution (variational posterior). One candidate of our variational family, for the virtue of simplicity, is a mean-field (MF) Gaussian family as follows:

$$\mathcal{Q}^{\mathbf{MF}} = \left\{ q(\theta) \mid \ q(\theta) = q_\beta(\beta)q_B(B), \log\beta \sim N(\mu_1, \sigma_1^2), B \sim N(\mu_2, \sigma_2^2) \right\}. \tag{2.3}$$

The above variational family is the same as a lognormal distribution on β and normal distribution on B. Also, we point out that β and B are independent in $\mathcal{Q}^{\mathbf{MF}}$ and each $q(\theta) \in \mathcal{Q}^{\mathbf{MF}}$ is governed by its own parameter set, $\nu^{\mathbf{MF}} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^\top$. $\nu^{\mathbf{MF}}$ denotes the set of variational parameters which will be updated to find the optimal variational distribution closest to the true posterior (2.2).

In addition, we suggest a bivariate normal (BN) family to exploit the interdependence among the parameters (β, B) as follows:

$$\mathcal{Q}^{\mathbf{BN}} = \left\{ q(\theta) \mid \ q(\theta) = q(\beta, B), (\log\beta, B) \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\}, \tag{2.4}$$

where $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$.

$\mathcal{Q}^{\mathbf{MF}}$ can also be represented as (independent) bivariate normal family. The variational parameters of BN family are $\nu^{\mathbf{BN}} = (\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12})^\top$. Once a variational family is selected, one can find the variational posterior by maxiimizing the ELBO between a variational distribution $q \in \mathcal{Q}$ and the true posterior (2.2). Recall the updates in (1.10) and the Monte Carlo estimates of the gradients in (1.12):

$$\nu^{(t+1)} \leftarrow \nu^{(t)} + \eta_t \widehat{\nabla_\nu \mathcal{L}(\boldsymbol{\nu})},$$

$$\widehat{\nabla_\nu \mathcal{L}(\boldsymbol{\nu})} = \frac{1}{S}\sum_{s=1}^{S} \nabla_\nu \log q(\theta_s) \left(\log p\left(\theta_s, \mathcal{D}\right) - \log q(\theta_s)\right).$$

Starting with initial values $\boldsymbol{\nu}^{(0)}$, we update the variational parameters in the direction of increasing ELBO using BBVI. The summary of BBVI algorithm is shown in Algorithm 2.1.

Algorithm 2.1 Black box variational inference (BBVI)

Initialize: $p(\theta)$, $q(\theta; \boldsymbol{\nu}^{(0)})$ and learning rate sequence η_t.

1: **while** ELBO increases **do**
2: Draw $\theta^{(s)} \sim q\left(\theta; \boldsymbol{\nu}^{(t)}\right)$, $s = 1, \ldots, S$;
3: Get $\widehat{\nabla_\nu \mathcal{L}(\boldsymbol{\nu})}$ based on the S sample points;
4: Update $\nu^{(t+1)} \leftarrow \nu^{(t)} + \eta_t \widehat{\nabla_\nu \mathcal{L}(\boldsymbol{\nu})}$;
5: **end while**

Output: Optimal variational parameters $\boldsymbol{\nu}^*$

In the next subsection, we discuss more details of implementing BBVI with the mean-field family (2.3).

2.2.1 VB algorithm with MF family

In the mean-field family, we have four variational parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2)$. For each variational parameter ν, we should compute $\nabla_\nu \log q(\theta)$ to evaluate the Monte Carlo gradients. Although one can simply use PyTorch's automatic differentiation without manual calculation, it is worth manually calculating the gradients to fully understand a BBVI algorithm. First, for μ_i, $i = 1, 2$, the gradients are:

$$\nabla_{\mu_1} \log q(\theta; \boldsymbol{\nu}) = \frac{\log \beta - \mu_1}{\sigma_1^2},$$
$$\nabla_{\mu_2} \log q(\theta; \boldsymbol{\nu}) = \frac{B - \mu_2}{\sigma_2^2}.$$

For σ_i, $i = 1, 2$, we need to be more careful because σ_i must be always positive. During the updates, it may occur that σ_i takes a negative value. In order to preclude this issue, we consider a reparametrization $\sigma_i = \log\left(1 + e^{\sigma_i'}\right)$ and update the quantity σ_i', as a free

parameter, instead of updating σ_i. Then,

$$\nabla_{\sigma_1'} \log q(\theta;\boldsymbol{\nu}) = \frac{e^{\sigma_1'}}{1+e^{\sigma_1'}} \left(\frac{(\log\beta - \mu_1)^2}{\sigma_1^3} - \frac{1}{\sigma_1} \right),$$
$$\nabla_{\sigma_2'} \log q(\theta;\boldsymbol{\nu}) = \frac{e^{\sigma_2'}}{1+e^{\sigma_2'}} \left(\frac{(B - \mu_2)^2}{\sigma_2^3} - \frac{1}{\sigma_2} \right).$$

In the next subsection, we discuss more details with the bivariate normal family (2.4).

2.2.2 VB algorithm with BN family

In addition to the positivity condition on some variational parameters, we should control postive definiteness of the covariance matirx $\boldsymbol{\Sigma}$ in (2.4). To guarantee the positive definiteness, we use Cholesky decomposition such that $\Sigma = LL^T$, where $L = \left(\begin{smallmatrix} l_{11} & 0 \\ l_{12} & l_{22} \end{smallmatrix}\right)$. We update the elements of the lower triangular matrix (l_{11}, l_{12}, l_{22}) in place of directly updating $(\sigma_{11}, \sigma_{12}, \sigma_{22})$ to avoid the cases of negative definite $\boldsymbol{\Sigma}$. Computation of $\nabla_\nu \log q(\theta;\boldsymbol{\nu})$ with BN family is as follows:

$$\nabla_{\mu_1} \log q(\theta;\boldsymbol{\nu}) = \frac{1}{l_{22}^2} \left(\frac{(l_{22}^2 + l_{12}^2)(\log\beta - \mu_1)}{l_{11}} - \frac{l_{12}(B-\mu_2)}{l_{11}} \right),$$
$$\nabla_{\mu_2} \log q(\theta;\boldsymbol{\nu}) = \frac{1}{l_{22}^2} \left(B - \mu_2 - \frac{l_{12}(\log\beta - \mu_1)}{l_{11}} \right),$$
$$\nabla_{l_{11}'} \log q(\theta;\boldsymbol{\nu})$$
$$= \frac{e^{l_{11}'}}{1+e^{l_{11}'}} \left(-\frac{1}{l_{11}} + \frac{1}{l_{22}^2} \left(\frac{(l_{22}^2 + l_{12}^2)(\log\beta - \mu_1)^2}{l_{11}^3} - \frac{l_{12}(\log\beta - \mu_1)(B-\mu_2)}{l_{11}^2} \right) \right),$$
$$\nabla_{l_{22}'} \log q(\theta;\boldsymbol{\nu})$$
$$= \frac{e^{l_{22}'}}{l_{22}(1+e^{l_{22}'})} \left(\frac{1}{l_{22}^2} \left(\frac{(l_{22}^2 + l_{12}^2)(\log\beta - \mu_1)^2}{l_{11}^2} - \frac{2l_{12}(\log\beta - \mu_1)(B-\mu_2)}{l_{11}} \right) \right.$$
$$\left. + \left(\frac{(B-\mu_2))}{l_{22}} \right)^2 - \left(\frac{\log\beta - \mu_1}{l_{11}} \right)^2 - 1 \right),$$
$$\nabla_{l_{12}} \log q(\theta;\boldsymbol{\nu}) = \frac{1}{l_{22}^2} \left(\frac{(\log\beta - \mu_1)(B-\mu_2)}{l_{11}} - \frac{l_{12}(\log\beta - \mu_1)^2}{l_{11}^2} \right),$$

where $l_{ii} = \log\left(1 + e^{l_{ii}'}\right)$ for $i = 1, 2$.

2.3 PMLE and MCMC

We compare our VB algorithm with two other methods, a PMLE method (Ghosal et al., 2020) and a MCMC based method (Møller et al., 2006). In thid section, we briefly introduce the two competitors.

PMLE: Let $h(\beta, B)$ denote the pseudo-likelihood in (1.2). Ghosal et al. (2020) used grid search to find pseudo maximum likelihood estimate (PMLE) for Ising parameters which simultaneously satisfies $\frac{\partial}{\partial \beta} \log h(\beta, B) = 0$ and $\frac{\partial}{\partial B} \log h(\beta, B) = 0$.

$$\frac{\partial}{\partial \beta} \log h(\beta, B) = \sum_{i=1}^{n} m_i(\boldsymbol{x}) \left(x_i - \tanh\left(\beta m_i(\boldsymbol{x}) + B\right)\right) = 0,$$
$$\frac{\partial}{\partial B} \log h(\beta, B) = \sum_{i=1}^{n} \left(x_i - \tanh\left(\beta m_i(\boldsymbol{x}) + B\right)\right) = 0.$$

We create a grid such that the search space for β contains all points from 0.1 to 2 in increments of 0.01 and the search space for B increases from -1 to 1 by 0.01.

MCMC: Let $p(\theta)$ be a prior distribution of θ and $p(\boldsymbol{x} \mid \theta)$ be the true likelihood in (1.1). The closed form of the true posterior (1.4) is not available because the integral in the denominator is not analytically tractable. Instead, one can use a sampling method. Consider a Metropolis-Hastings ratio given by:

$$MH(\theta' \mid \theta) = \frac{p(\theta') f_{\theta'}(\boldsymbol{x}) u(\theta \mid \theta')}{p(\theta) f_{\theta}(\boldsymbol{x}) u(\theta' \mid \theta)} \times \frac{Z_n(\theta)}{Z_n(\theta')}, \tag{2.5}$$

where $u(\theta' \mid \theta)$ is the proposal density and $f_\theta(\boldsymbol{x})$ is an unnormalized density of Ising model, i.e, $p(\theta \mid \boldsymbol{x}) = Z_n(\theta)^{-1} f_\theta(\boldsymbol{x})$. Obtaining the Metropolis-Hastings ratio, however, is still limited because we cannot compute the normalizing constant in $p(\boldsymbol{x} \mid \theta)$ even with a moderate amount of n. In order to remove the ratio of normalizing constants in (2.5), Møller et al. (2006) proposed an alternative approach using an auxiliary variable $\boldsymbol{z}$ with conditional distribution $g(\boldsymbol{z} \mid \theta, \boldsymbol{x})$. Targeting $\pi(\theta, \boldsymbol{z} \mid \boldsymbol{x}) \propto p(\theta) g(\boldsymbol{z} \mid \theta, \boldsymbol{x}) \frac{1}{Z_n(\theta)} f_\theta(\boldsymbol{x})$ instead of $\pi(\theta \mid \boldsymbol{x})$ and taking the proposal density for $\boldsymbol{z}'$ to be an Ising likelihood depending on θ', that is, $\boldsymbol{z}' \sim \frac{1}{Z_n(\theta')} f_{\theta'}(\boldsymbol{z}')$, we can cancel the normalizing constants in the Metropolis-Hastings ratio

(2.5) as follows:

$$MH(\theta', \boldsymbol{z}' \mid \theta, \boldsymbol{z}) = \frac{g(\boldsymbol{z}' \mid \theta', \boldsymbol{x}) p(\theta') f_{\theta'}(\boldsymbol{x}) f_{\theta}(\boldsymbol{z}) u(\theta \mid \theta')}{g(\boldsymbol{z} \mid \theta, \boldsymbol{x}) p(\theta) f_{\theta}(\boldsymbol{x}) f_{\theta'}(\boldsymbol{z}') u(\theta' \mid \theta)}. \tag{2.6}$$

As suggested by Møller et al. (2006), the conditional density $g(\boldsymbol{z} \mid \theta, \boldsymbol{x})$ is approximated by $f_{\tilde{\theta}}(\boldsymbol{z})/Z_n(\tilde{\theta})$ which does not depend on θ, where $\tilde{\theta}$ is PMLE. Also, using independent log-normal and normal distributions so that $u(\theta \mid \theta')/u(\theta' \mid \theta) = 1$, the ratio (2.6) further reduces to:

$$MH(\theta', \boldsymbol{z}' \mid \theta, \boldsymbol{z}) = \mathbb{1}(\theta \in \Theta) \frac{f_{\tilde{\theta}}(\boldsymbol{z}') f_{\theta'}(\boldsymbol{x}) f_{\theta}(\boldsymbol{z})}{f_{\tilde{\theta}}(\boldsymbol{z}) f_{\theta}(\boldsymbol{x}) f_{\theta'}(\boldsymbol{z}')}. \tag{2.7}$$

We accept the proposal $(\theta', \boldsymbol{z}')$ as a new state with probability $\min\{1, MH(\theta', \boldsymbol{z}' \mid \theta, \boldsymbol{z})\}$. Although Møller et al. (2006)'s approach intelligently controls the normalizing constant, the algorithm is still computationally expensive because it involves sampling from an Ising likelihood at each Metropolis-Hastings iteration.

2.4 Numerical experiments

2.4.1 Generating observed data

For numerical experiments, we need a coupling matrix A_n and an observed vector $\boldsymbol{x}_{observed}$ from (1.1). First, for generating a random d-regular graph and its scaled adjacency matrix, we use a python package `NetworkX`. Using the scaled adjacency matrix as our coupling matrix A_n, we facilitate Metropolis-Hastings algorithm to generate an observed vector $\boldsymbol{x}_{observed}$ with true parameters (β_0, B_0) as follows:

0. Define $H(\boldsymbol{x}) = \frac{\beta_0}{2} \boldsymbol{x}^\top A_n \boldsymbol{x} + B_0 \sum_{i=1}^n x_i$ and start with a random binary vector $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.

1. Randomly choose a spin x_i, $i \in \{1, \ldots, n\}$.

2. Flip the chosen spin, i.e. $x_i = -x_i$, and calculate $\Delta H = H(\boldsymbol{x}_{new}) - H(\boldsymbol{x}_{old})$ due to this flip.

3. The probability that we accept $\boldsymbol{x}_{new}$ is:

$$\mathbb{P}(\text{accept } \boldsymbol{x}_{new}) = \begin{cases} 1, & \text{if } \Delta H > 0, \\ \exp\left(\Delta H\right), & \text{otherwise.} \end{cases}$$

4. If rejected, put the spin back, i.e. $x_i = -x_i$.

5. Go to 1 until the maximum number of iterations (L) is reached.

6. After $L = 1,000,000$ iterations, the last result is a sample $\boldsymbol{x}_{observed}$ we use.

One can read Izenman (2021) for more details of sampling from Ising model.

2.4.2 Performance Comparison

We compare the performance of the parameter estimation methods for two-parameter Ising model (1.1) under various combinations of (d, n) and (β_0, B_0). Using the given coupling matrix A_n for each scenario, we repeat the following steps R times:

- Generate an observed vector $\boldsymbol{x}$ from (1.1) with true parameters (β_0, B_0).
- Using the proposed BBVI algorithm with MF family or BN family, obtain the optimal variational parameters $\boldsymbol{\nu}^*$.
- We get the estimates $\hat{\theta} = (\log \hat{\beta}, \hat{B})^\top = (\mu_1^*, \mu_2^*)^\top$.

We use $S = 20$ or $S = 200$ as the Monte Carlo sample size in (1.12). Figure 2.1 describes ELBO convergences for the two different sample sizes with MF family and BN family. The figure indicates that the ELBO converges well with a moderate choice of S. Further, for more stable convergence, one might choose higher S and BN family over MF family.

We use Mean squared error (MSE) as the measurement for assessing the performances with $R = 50$ pairs of estimates $(\hat{\beta}_1, \hat{B}_1), \ldots, (\hat{\beta}_R, \hat{B}_R)$:

$$MSE = \frac{1}{R} \sum_{r=1}^{R} \left(\left(\hat{\beta}_r - \beta_0 \right)^2 + \left(\hat{B}_r - B_0 \right)^2 \right). \tag{2.8}$$

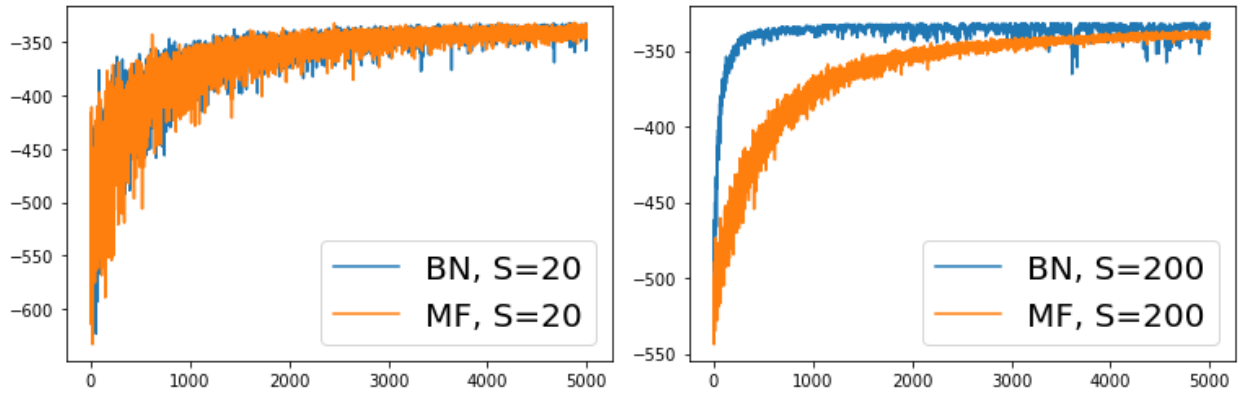

Figure 2.1: Left: ELBO convergence with two variational families (BN and MF) and $S = 20$. Right: ELBO convergence with two variational families (BN and MF) and $S = 200$. Blue lines represent BN family and orange lines represent MF family in each plot.

For each pair of (β_0, B_0) we take $d = 10, 50$. The two numbers in each cell of the tables, Table 2.1, 2.2, 2.3, and 2.4, represent MSE values or convergence time when $n = 100$ and $n = 500$ respectively. First, we consider a small value of $\beta_0 = 0.2$ with $B_0 = \pm 0.2, \pm 0.5$. In these cases, as shown in Table 2.1 and 2.2, PMLE is the fastest but less accurate especially for $d = 50$. MCMC achieves smaller MSEs but it has the highest runtimes. Our VB methods notably reduce the runtimes without compromising accuracy.

Second, for higher interaction parameter $\beta_0 = 0.7$ with $B_0 = \pm 0.2, \pm 0.5$, our VB algorithms are more accurate than the others (See Table 2.3 and 2.4). The numerical studies validate the superiority of our proposed VB methods. For more practical applications, we used our algorithm to regenerate an image in the next subsection.

2.4.3 Image reconstruction

Ising model can be used for constructing an image in computer vision field. In particular, the Bayesian procedure facilitate the reconstruction easily by using the posterior predictive distribution Halim (2007). Consider an image in which each pixel represents either -1(white) or 1(black). For choice of coupling matrix A_n, we use twelve-nearest neighbor structure and construct corresponding scaled adjacency matrix Hurn et al. (2003). Then, we generate such images following the steps in the subsection 2.4.1 with a true parameter pair (β_0, B_0) and use

Table 2.1: Mean squared errors and computation times for each pair of (β_0, B_0) when $n = 100$ (left numbers) and $n = 500$ (right numbers) given the degree of underlying graph (d).

Degree of graph (d)	Method	Monte Carlo samples (S)	$(0.2, 0.2)$	$(0.2, -0.2)$	Convergence time (sec)
10	PMLE[1]	-	0.119 / 0.049	0.091 / 0.022	3.2 / 3.6
	MCMC[2]	-	0.098 / 0.194	0.097 / 0.146	165.2 / 675.8
	MF[3] family	20	0.059 / 0.018	0.057 / 0.014	6.3 / 10.3
		200	0.037 / 0.027	0.032 / 0.013	10.7 / 15.9
	BN[4] family	20	0.064 / 0.021	0.061 / 0.018	7.0 / 12.3
		200	0.041 / 0.026	0.035 / 0.009	12.3 / 17.8
50	PMLE	-	0.369 / 0.105	0.299 / 0.170	3.1 / 3.6
	MCMC	-	0.165 / 0.084	0.144 / 0.110	168.1 / 678.0
	MF family	20	0.072 / 0.027	0.073 / 0.030	6.4 / 10.1
		200	0.070 / 0.023	0.070 / 0.027	10.8 / 15.9
	BN family	20	0.081 / 0.045	0.082 / 0.049	7.1 / 12.1
		200	0.073 / 0.058	0.067 / 0.067	12.2 / 17.5

[1]PMLE, pseudo maximum likelihood estimate (Ghosal et al., 2020); [2]MCMC, markov chain monte carlo (Møller et al., 2006); [3]MF, mean-field; [4]BN, bivariate normal

it as our given data $\boldsymbol{x}_{observed}$. With the generated image $\boldsymbol{x}_{observed}$ and coupling matrix A_n, we obtain $(\hat{\beta}, \hat{B})$ after implementing the parameter estimation procedure based on BN family. The estimates $(\hat{\beta}, \hat{B})$ are used for data regeneration following the steps in the subsection 2.4.1 again. In Figure 2.2, we plot two original images in the left column. The first original image was generated with $\beta_0 = 1.2, B_0 = 0.2$ (left top) and we use $\beta_0 = 1.2, B_0 = -0.2$ for the second one (left bottom). Also, in the right column, there are two corresponding images regenerated with $(\hat{\beta} = 1.071, \hat{B} = 0.357)$ (right top) and $(\hat{\beta} = 0.982, \hat{B} = -0.326)$ (right bottom) respectively. It seems that, using two-parameter Ising model and our VB method, we can reconstruct the overall tendency of black and white images fairly well. For more precise pixel-by-pixel reconstruction, one can utilize multiple threshold parameters, $\boldsymbol{B} = (B_1, \ldots, B_n)^\top$ (See the section 2.6).

Table 2.2: Mean squared errors and computation times for each pair of (β_0, B_0) when $n = 100$ (left numbers) and $n = 500$ (right numbers) given the degree of underlying graph (d).

Degree of graph (d)	Method	Monte Carlo samples (S)	$(0.2, 0.5)$	$(0.2, -0.5)$	Convergence time (sec)
10	PMLE	-	0.270 / 0.053	0.126 / 0.068	3.1 / 3.5
	MCMC	-	0.228 / 0.275	0.174 / 0.255	164.9 / 673.5
	MF family	20	0.075 / 0.017	0.075 / 0.017	6.5 / 10.2
		200	0.071 / 0.015	0.070 / 0.014	10.8 / 16.1
	BN family	20	0.090 / 0.039	0.091 / 0.038	7.1 / 12.0
		200	0.088 / 0.033	0.087 / 0.032	12.3 / 17.1
50	PMLE	-	0.792 / 0.197	0.653 / 0.251	3.2 / 3.5
	MCMC	-	0.407 / 0.153	0.380 / 0.169	166.1 / 675.1
	MF family	20	0.088 / 0.023	0.089 / 0.024	6.3 / 10.2
		200	0.082 / 0.019	0.080 / 0.021	10.4 / 16.4
	BN family	20	0.090 / 0.065	0.112 / 0.072	7.2 / 12.1
		200	0.110 / 0.081	0.102 / 0.096	12.2 / 17.2

2.5 Real data analysis

In two-parameter Ising model, higher value of β implies stronger interactions between connected nodes and the threshold parameter B controls the model size (number of 1s), where the model size is greater for $B > 0$, smaller for $B < 0$. In this section, applying our methods to a real data set, we obtain the estimated interaction parameter and the threshold parameter $(\hat{\beta}, \hat{B})$.

2.5.1 Data description

Stanford Network Analysis Project (SNAP) provides a Facebook network data set (Leskovec and Krevl, 2014) available at `http://snap.stanford.edu/data/ego-Facebook.html`. The Facebook network consists of 4,039 nodes and 88,234 edges. Each node represents a Facebook user and there is an edge between two nodes if corresponding users are friends. The data set also contains user features such as birthday, school, gender and location. The features are fully anonymized. For instance, while the original data may include a feature "location

Table 2.3: Mean squared errors and computation times for each pair of (β_0, B_0) when $n = 100$ (left numbers) and $n = 500$ (right numbers) given the degree of underlying graph (d).

Degree of graph (d)	Method	Monte Carlo samples (S)	$(0.7, 0.2)$	$(0.7, -0.2)$	Convergence time (sec)
10	PMLE	-	0.439 / 0.074	0.393 / 0.083	3.1 / 3.6
	MCMC	-	0.114 / 0.080	0.111 / 0.064	170.3 / 678.1
	MF family	20	0.109 / 0.135	0.095 / 0.144	6.3 / 10.0
		200	0.116 / 0.140	0.100 / 0.148	10.5 / 16.0
	BN family	20	0.085 / 0.074	0.080 / 0.081	6.9 / 11.9
		200	0.087 / 0.080	0.080 / 0.088	12.3 / 17.0
50	PMLE	-	0.718 / 0.333	0.730 / 0.386	3.2 / 3.8
	MCMC	-	0.170 / 0.107	0.172 / 0.120	171.0 / 673.6
	MF family	20	0.093 / 0.185	0.093 / 0.178	6.2 / 10.2
		200	0.101 / 0.191	0.103 / 0.184	10.4 / 16.1
	BN family	20	0.068 / 0.107	0.071 / 0.086	7.0 / 12.2
		200	0.068 / 0.117	0.075 / 0.110	12.5 / 17.3

= Michigan", the anonymized data would simply contain"location = anonymized location A". Thus, using the anonymized data, we can determine whether two users stay in the same location, but we do not know where.

Among the $4,039$ users, we select only users who disclose gender information to create a sub-graph such that there are $3,948$ nodes and $84,716$ edges in the sub-graph. Later we used the gender information for binary observations. Each node has different number of neighbors (degree). The maximum degree of the sub-graph is $1,024$, and the minimum is 1, with an average degree 42.92 (Figure 2.3 shows the nodes and edges in the reduced network).

2.5.2 Parameter estimation

We utilize the selected users ($n = 3,948$) as a real data set to apply our VB algorithm with the features school and gender as observed binary vectors. For the school feature, we encode 1 if a user (node) belongs to an anonymized school A, otherwise -1. For the gender feature, we encode a group by 1 and the other group by -1. The model sizes are 114 and $2,417$ for school and gender respectively. Note that, no matter which feature is used, the

Table 2.4: Mean squared errors and computation times for each pair of (β_0, B_0) when $n = 100$ (left numbers) and $n = 500$ (right numbers) given the degree of underlying graph (d).

Degree of graph (d)	Method	Monte Carlo samples (S)	$(0.7, 0.5)$	$(0.7, -0.5)$	Convergence time (sec)
10	PMLE	-	0.603 / 0.228	0.788 / 0.233	3.2 / 3.7
	MCMC	-	0.143 / 0.113	0.178 / 0.121	170.0 / 678.5
	MF family	20	0.083 / 0.184	0.078 / 0.193	6.3 / 10.1
		200	0.103 / 0.188	0.096 / 0.199	10.4 / 16.2
	BN family	20	0.049 / 0.074	0.047 / 0.079	7.8 / 12.3
		200	0.059 / 0.077	0.056 / 0.083	12.5 / 17.5
50	PMLE	-	0.893 / 0.638	0.804 / 0.781	3.1 / 3.8
	MCMC	-	0.144 / 0.154	0.126 / 0.255	171.9 / 677.1
	MF family	20	0.080 / 0.219	0.071 / 0.209	6.2 / 10.0
		200	0.095 / 0.225	0.091 / 0.219	10.6 / 16.0
	BN family	20	0.044 / 0.080	0.040 / 0.074	7.7 / 12.1
		200	0.051 / 0.085	0.048 / 0.079	12.4 / 17.3

connectivity among nodes does not change. One can expect that the interaction parameter β is higher when we use the school feature because people from the same school are more likely to be Facebook friends with each other. Also, we expect the threshold parameter B will be negative for school and positive for gender because of the model sizes. Table 2.5 summarizes the estimated parameters with standard errors (SE) (for VB and MCMC) and runtimes.

Table 2.5: The estimated parameters with standard errors (SE) in parentheses and time costs for the features gender and school.

Feature	Method	Monte Carlo samples (S)	$\hat{\beta}$ (SE)	$\hat{B}$ (SE)	Convergence time (sec)
Gender	PMLE	-	0.250(-)	0.180(-)	5.9
	MCMC	-	0.112(0.066)	0.210(0.031)	27109.2
	MF family	200	0.132(0.070)	0.189(0.027)	173.3
	BN family	200	0.106(0.089)	0.248(0.033)	271.2
School	PMLE	-	0.260(-)	-1.560(-)	6.0
	MCMC	-	0.203(0.087)	-1.610(0.093)	27059.3
	MF family	200	0.252(0.061)	-1.445(0.049)	173.4
	BN family	200	0.299(0.092)	-1.602(0.103)	265.3

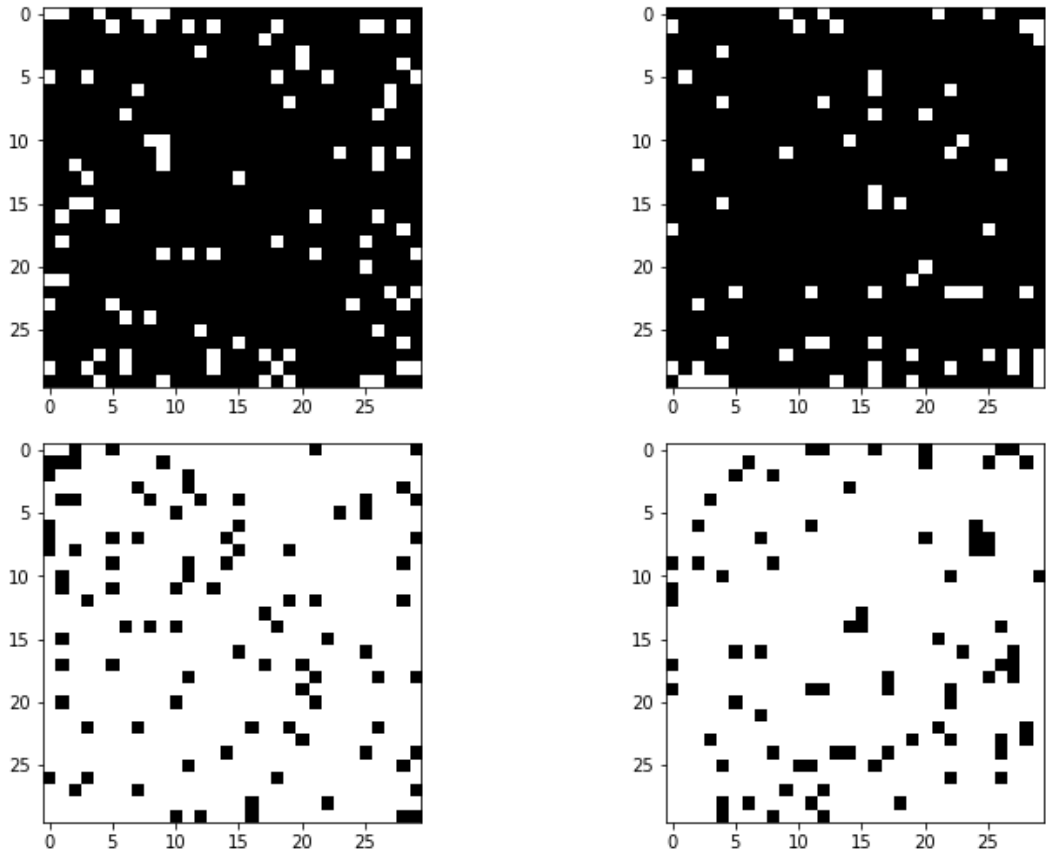

Figure 2.2: Original images (left) and the estimated images (right).

SEs of MCMC in Table 2.5 are calculated based on 10,000 draws after the burn-in period of 10,000 iterations. For our VB methods, to calculate SEs, we sample the same amount of β and B (10,000 draws) from the optimal variational distributions and calculate sample standard deviations. Figure 2.4 indicate density plots of the draws from BN family, MF family, and MCMC for the features gender and school.

Computational gain is very clear from the figures in Table 2.5. While estimated parameters are comparable for all the methods, the MCMC implementation takes about fifty times more time compared to VB to achieve similar level of accuracy. The PMLE approach does not produce SE and thus limited for statistical inference.

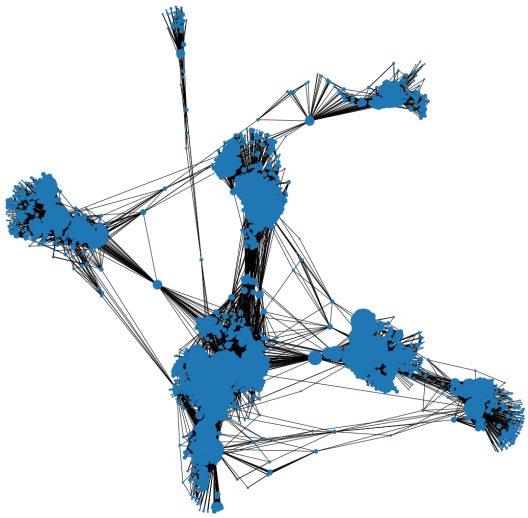

Figure 2.3: Visualization of Facebook network data where the size of circle represents the degree of the node.

2.6 Extension to multi threshold parameters

Beyond the two-parameter Ising model which is a main material of this paper, we can extend the parameter estimation procedure using VB to more general Ising model. Allowing multi-threshold parameters, that is, $\boldsymbol{B} = (B_1, \ldots, B_n)$, the likelihood is:

$$\mathbb{P}^{(n)}_{\beta,\boldsymbol{B}}(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{Z_n(\beta, \boldsymbol{B})} \exp\left(\frac{\beta}{2}\boldsymbol{x}^\top A_n \boldsymbol{x} + \sum_{i=1}^{n} B_i x_i\right). \tag{2.9}$$

Note, there are $n+1$ unknown parameters in the likelihood (2.9) and the corresponding pseudo-likelihood is:

$$\begin{aligned}
&\prod_{i=1}^{n} \mathbb{P}^{(n)}_{\beta,\boldsymbol{B}}\left(X_i = x_i \mid X_j, j \neq i\right) \\
&= 2^{-n} \exp\left(\sum_{i=1}^{n} \left(\beta x_i m_i(\boldsymbol{x}) + B_i x_i - \log\cosh(\beta m_i(\boldsymbol{x}) + B_i)\right)\right).
\end{aligned} \tag{2.10}$$

We introduce a VB algorithm for estimating the parameters $\theta := (\log\beta, B_1, \ldots, B_n)$ given A_n and $\boldsymbol{x}$. As a natural extension, consider the following multivariate Gaussian variational

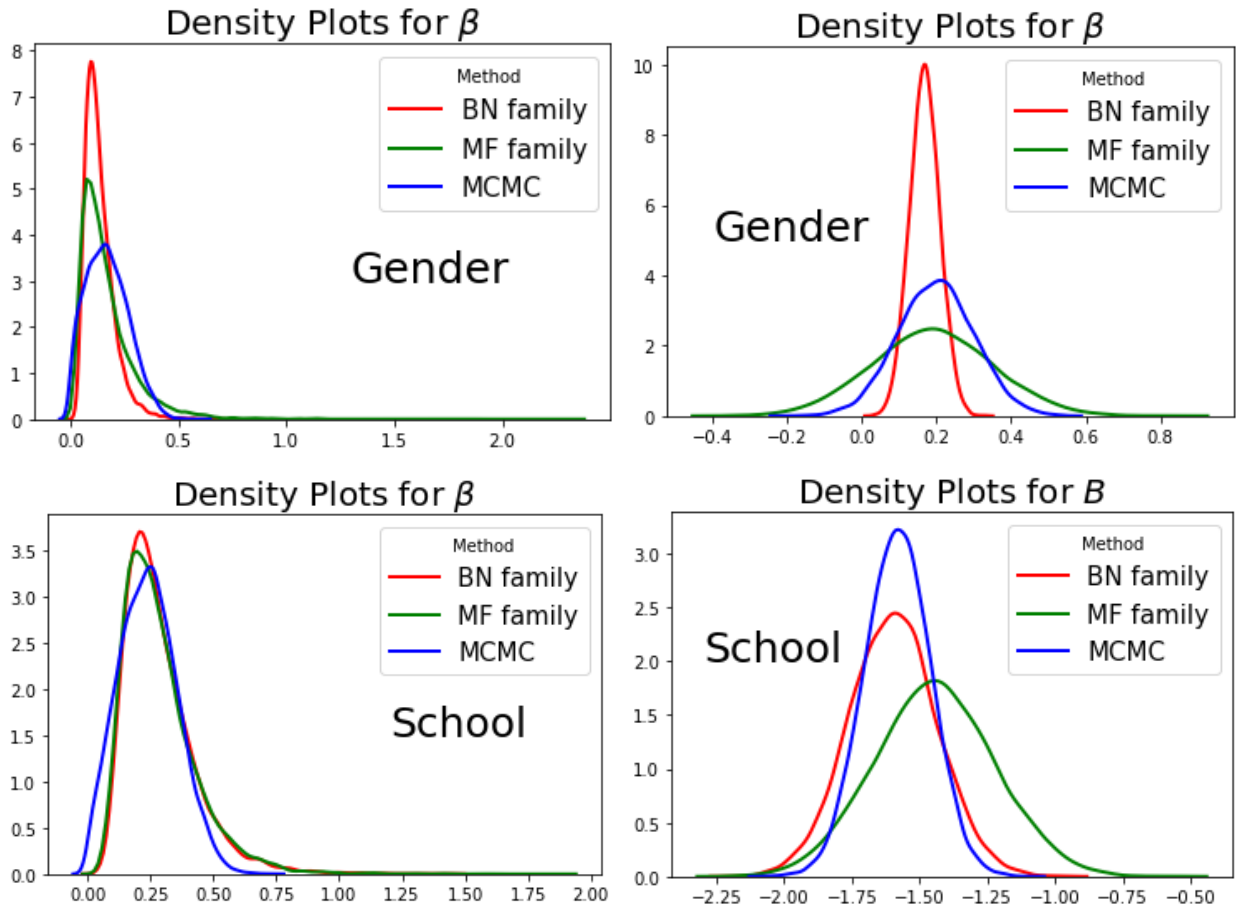

Figure 2.4: Density plots for the estimated parameters (left: β, right: B) from VB with BN family (red), VB with MF family (green), and MCMC (blue) for the features gender and school.

family:

$$\mathcal{Q}^{MG} = \left\{ q(\theta) \mid \ q(\theta) = q(\beta, \boldsymbol{B}), (\log \beta, \boldsymbol{B}) \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\}, \tag{2.11}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{n+1}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{(n+1)\times(n+1)}$. Without any assumption on $\boldsymbol{\Sigma}$, updating all the variational parameters and finding variational posterior require quite demanding computaional costs because the total number of updated parameters is $(n+1)+(n+1)(n+2)/2$. If we assume $\boldsymbol{\Sigma}$ is a diagonal matrix, that is, the variational family (2.11) is mean-field, the number of parameters to be updated reduces to $2(n+1)$. Under the pseudo-likelihood (2.10) and multivariate Gaussian family (2.11) with a diagonal covariance matrix $\boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_n^2)$, we can easily compute the gradients (1.12) for updating the variational parameters and develop

a VB algorithm for multi-threshod parameter Ising model.

2.7 Posterior consistency

The main theoretical contribution of this work lies in establishing the consistency of the variational posterior for the Ising model with the true likelihood replaced by the pseudo-likelihood. In this direction, we first establish the rates at which the true posterior based on the pseudo-likelihood concentrates around the ε_n- shrinking neighborhoods of the true parameters. With a suitable bound on the Kulback-Leibler distance between the true posterior (under pseudo-likelihood) and the variational posterior, we next establish the rate of contraction for the variational posterior and demonstrate that the variational posterior also concentrates around ε_n-shrinking neighborhoods of the true parameter. These results have been derived under three set of assumptions on the coupling matrix A_n. Indeed, we demonstrate that the variational posterior consistency holds for the same set of assumptions on A_n as those needed for the convergence of the maximum likelihood estimates based on the pseudo-likelihood. One of the main caveats in establishing the posterior contraction rates under the pseudo-likelihood structure is in ensuring that the concentration of the variational posterior occurs in $\mathbb{P}_0^{(n)}$ probability where $\mathbb{P}_0^{(n)}$ is the distribution induced by the true likelihood and not the pseudo-likelihood. Indeed, we could show that in $\mathbb{P}_0^{(n)}$ probability, the contraction of variational posterior happens at the rate $1 - 1/M_n$ in contrast to the faster rate $1 - \exp(-Cn\varepsilon_n^2), C > 0$ for the true posterior. As a final theoretical contribution, we establish that the variational Bayes estimator convergences to the true parameters at the rate $1/\varepsilon_n$ where ε_n can be chosen $n^{-\delta}$, $0 < \delta < 1/2$ provided the A_n matrix satisfies certain regularity assumptions.

2.7.1 Sketch of Proof

In this subsection, we states main theorem and provide a sketch of proof for consistency of the variational posterior (1.7). In this direction, we establish the variational posterior

contraction rates to evaluate how well the posterior distribution of β and B under the variational approximation concentrates around the true values β_0 and B_0. Towards the proof, we make the following assumptions:

Assumption 1 (Bounded row sums of A_n)**.** *The row sums of A_n are bounded above*

$$\max_{i\in[n]} \sum_{j=1}^{n} A_n(i,j) \leq \gamma,$$

for a constant γ independent of n.

Assumption 1 is the same as (1.2) in Ghosal et al. (2020). As a consequence of Assumption 1, it can be shown $|m_i(\boldsymbol{x})| \leq \gamma$, $i = 1 \ldots, n$.

Assumption 2 (Mean field assumption on A_n)**.** *Let $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ such that*

$$(i)\ \sum_{i=1}^{n}\sum_{j=1}^{n} A_n(i,j) = O(n\epsilon_n^2), \qquad (ii)\ \sum_{i=1}^{n}\sum_{j=1}^{n} A_n(i,j)^2 = o(n\epsilon_n^2).$$

Assumption 2-(i) is the same as condition (1.4) in Ghosal et al. (2020) on A_n for $\epsilon_n = 1$. Assumption 2-(ii) is the same as (1.6) in Ghosal et al. (2020) with $\epsilon_n = 1$. For more details on the mean field assumption, we refer to Definition 1.3 in Basak and Mukherjee (2017).

Assumption 3 (Bounded variance of A_n)**.** *Let $\bar{A}_n = (1/n)\sum_{i=1}^{n}\sum_{j=1}^{n} A_n(i,j)$,*

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{n} A_n(i,j) - \bar{A}_n)^2 > 0.$$

Finally, the Assumption 3 corresponds to (1.7) in Ghosal et al. (2020). The validity of Assumption 3 ensures that $T_n(\boldsymbol{x}) = (1/n)\sum_{i=1}^{n}(m_i(\boldsymbol{x}) - \bar{m}(\boldsymbol{x}))^2$ is bounded below and above in probability, an essential requirement towards the proof of contraction rates of the variational posterior.

Let $\theta = (\beta, B)$ be the model parameter and $\theta_0 = (\beta_0, B_0)$ be the true parameter from which the data are generated. Let $L(\theta)$ and $L(\theta_0)$ denote the pseudo-likelihood as in (1.2) under the model parameters and true parameters respectively. Further, let L_0 denote the true probability mass function from which $X^{(n)}$ is generated. Thus, L_0 is as in (1.1) with

$\theta = \theta_0$. We shall use the notations $\mathbb{E}_0^{(n)}$ and $\mathbb{P}_0^{(n)}$ to denote expectation and probability mass function with respect to L_0.

We next present the main theorem which establishes the contraction rate for the variational posterior. Following the proof, we next establish the contraction rate of the variational Bayes estimator as a corollary. We shall use the term with dominating probability to imply that under $\mathbb{P}_0^{(n)}$, the probability of the event goes to 1 as $n \to \infty$.

Theorem 1 (Posterior Contraction)**.** *Let $\mathcal{U}_{\varepsilon_n} = \{\theta : \|\theta - \theta_0\|_2 \leq \varepsilon_n\}$ be neighborhood of the true parameters. Suppose ϵ_n satisfies Assumption 2, then in $\mathbb{P}_0^{(n)}$ probability*

$$Q^*(\mathcal{U}_{\varepsilon_n}^c) \to 0, \;\; n \to \infty,$$

where $\varepsilon_n = \epsilon_n \sqrt{M_n \log n}$ for any slowly increasing sequence $M_n \to \infty$ satisfying $\varepsilon_n \to 0$.

The above result establishes that the posterior distribution of β and B concentrates around the true value β_0 and B_0 at a rate slight larger than ϵ_n. The proof of the above theorem rests on following lemmas, whose proofs have been deferred to the Section 2.10, 2.11, and 2.12.

Lemma 1. *There exists a constant $C_0 > 0$, such that for any $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$,*

$$\mathbb{P}_0^{(n)} \left(\log \int_{\mathcal{U}_{\epsilon_n}^c} \frac{L(\theta)}{L(\theta_0)} p(\theta) d\theta \leq -C_0 n \epsilon_n^2 \right) \to 1, \;\; n \to \infty.$$

Lemma 2. *Let ϵ_n be the sequence satisfying the Assumption 2, then for any $C > 0$,*

$$\mathbb{P}_0^{(n)} \left(\left| \log \int \frac{L(\theta)}{L(\theta_0)} p(\theta) d\theta \right| \leq C n \epsilon_n^2 \log n \right) \to 1.$$

Lemma 3. *Let ϵ_n be the sequence satisfying Assumption 2, then for some $Q \in \mathcal{Q}^{\boldsymbol{MF}}$ and any $C > 0$,*

$$\mathbb{P}_0^{(n)} \left(\int \log \frac{L(\theta_0)}{L(\theta)} q(\theta) d\theta \leq C n \epsilon_n^2 \log n \right) \to 1.$$

Lemma 1 and Lemma 2 taken together suffice to establish the posterior consistency of the true posterior based on the pseudo-likelihood $L(\theta)$ as in (2.2). Lemma 3 on the other hand

is the additional condition which needs to hold to ensure the consistency of the variational posterior. We next state an important result which relates the variational posterior to the true posterior.

Formula for KL divergence: By Corollary 4.15 in Boucheron et al. (2013),

$$KL(P_1, P_2) = \sup_f \left[\int f dP_1 - \log \int e^f dP_2 \right].$$

Using the above formula in the context of variational distributions, we get

$$\int f dQ^* \leq KL(Q^*, \Pi(| \ X^{(n)})) + \log \int e^f d\Pi(| \ X^{(n)}). \tag{2.12}$$

The above relation serves as an important tool towards the proof of Theorem 1. Next, we provide a brief sketch of the proof. Further details on the proof have been deferred to Section 2.13.

Sketch of proof of Theorem 1: Let $f = (C_0/2)n\varepsilon_n^2 1[\theta \in \mathcal{U}_{\varepsilon_n}^c]$, then

$$\begin{aligned}
&(C_0/2)n\varepsilon_n^2 Q^*(\mathcal{U}_{\varepsilon_n}^c) \leq KL(Q^*, \Pi(| \ X^{(n)})) + \log(e^{(C_0/2)n\varepsilon_n^2}\Pi(\mathcal{U}_{\varepsilon_n}^c \mid X^{(n)}) + \Pi(\mathcal{U}_{\varepsilon_n} \mid X^{(n)})) \\
&\implies Q^*(\mathcal{U}_{\varepsilon_n}^c) \leq \frac{2}{C_0 n\varepsilon_n^2} KL(Q^*, \Pi(| \ X^{(n)})) + \frac{2}{C_0 n\varepsilon_n^2} \log(1 + e^{(C_0/2)n\varepsilon_n^2}\Pi(\mathcal{U}_{\varepsilon_n}^c \mid X^{(n)})).
\end{aligned}$$

By Lemma 2 and 3, it can be established with dominating probability for any $C > 0$, as $n \to \infty$

$$KL\left(Q^*, \Pi(| \ X^{(n)})\right) \leq Cn\epsilon_n^2 \log n.$$

By Lemma 1 and 2, it can be established with dominating probability, as $n \to \infty$

$$\Pi(\mathcal{U}_{\varepsilon_n}^c \mid X^{(n)}) \leq e^{-C_1 n\varepsilon_n^2}, \tag{2.13}$$

for any $C_1 > C_0/2$. Therefore, with dominating probability

$$\begin{aligned}
Q^*(\mathcal{U}_{\varepsilon_n}^c) &\leq \frac{2C}{C_0 M_n} + \frac{2}{C_0 n\varepsilon_n^2} \log\left(1 + e^{-(C_1 - C_0/2)n\varepsilon_n^2}\right) \\
&\sim \frac{2C}{C_0 M_n} + \frac{e^{-(C_1 - C_0/2)n\varepsilon_n^2}}{C_0 n\varepsilon_n^2} \to 0.
\end{aligned} \tag{2.14}$$

This completes the proof.

Note that (2.13) gives the statement for the contraction of the true posterior. Similarly the contraction rate for the variational posterior follows as a consequence of (2.14). An important difference to note is that $Q^*(\mathcal{U}^c_{\varepsilon_n})$ goes to 0 at the rate $1/M_n$ in contrast to the faster rate $e^{-C_1 n\varepsilon_n^2}$ for the true posterior.

Note, Theorem 1 gives the contraction rate of the variational posterior. However, the convergence of the of variational Bayes estimator to the true values of β_0 and B_0 is not immediate. The following corollary gives the convergence rate for the variational Bayes estimate as long as Assumptions 1, 2 and 3 hold.

Corollary 1 (Variational Bayes Estimator Convergence)**.** *Let ε_n be as in Theorem 1, then in $\mathbb{P}_0^{(n)}$ probability,*

$$\frac{1}{\varepsilon_n}\mathbb{E}_{Q^*}(\|\theta-\theta_0\|_2) \to 0, \;\; as\, n \to \infty.$$

Next, we provide a brief sketch of the proof. Further details of the proof have been deferred to Section 2.14.

Sketch of proof of Corollary 1: Let $f = (C_2/2)n\varepsilon_n\|\theta-\theta_0\|_2$, then

$$\begin{aligned}(C_2/2)n\varepsilon_n &\int \|\theta-\theta_0\|_2 dQ^*(\theta) \\ &\leq KL(Q^*, \Pi(|\, X^{(n)})) + \log\left(\int e^{C_2 n\varepsilon_n\|\theta-\theta_0\|_2/2} d\Pi(\theta \mid X^{(n)})\right).\end{aligned}$$

By Lemma 2 and 3, it can be established with dominating probability, for any $C > 0$

$$KL(Q^*, \Pi(|\, X^{(n)})) \leq Cn\epsilon_n^2 \log n.$$

By Lemma 1, and 2, it can be established with dominating probability, for some $C_2 > 0$

$$\int e^{(C_2/2)n\varepsilon_n\|\theta-\theta_0\|_2} d\Pi(\theta \mid X^{(n)}) \leq \frac{1}{(C_2/2)n\varepsilon_n^2} e^{Cn\epsilon_n^2 \log n}. \tag{2.15}$$

Therefore, with dominating probability

$$\int \|\theta-\theta_0\|_2 dQ^*(\theta) \leq \frac{2C\varepsilon_n}{C_2 M_n} - \frac{2\log(C_2/2)}{C_2 n\varepsilon_n} - \frac{2\varepsilon_n \log(n\varepsilon_n^2)}{C_2 n\varepsilon_n^2} + \frac{2C\varepsilon_n}{C_2 M_n} \leq \varepsilon_n o(1).$$

This completes the proof.

(2.15) follows as a consequence of convergence of the true posterior. An important thing to note that if ε_n can be made arbitrarily close to $n^{-\delta}$ for $0 < \delta < 1/2$, it guarantees close to $\sqrt{n}$ convergence.

2.8 Preliminary notations and Lemmas

Let $\theta = (\beta, B)$. Define $W_n := (W_{1n}, W_{2n})$ where

$$\begin{aligned} W_{1n}(\theta \mid \boldsymbol{x}) &= \sum_{i=1}^{n} m_i(\boldsymbol{x}) \left(x_i - \tanh(\beta m_i(\boldsymbol{x}) + B)\right), \\ W_{2n}(\theta \mid \boldsymbol{x}) &= \sum_{i=1}^{n} \left(x_i - \tanh(\beta m_i(\boldsymbol{x}) + B)\right). \end{aligned} \tag{2.16}$$

Also, define

$$H_n(\theta \mid \boldsymbol{x}) := \begin{bmatrix} \sum_{i=1}^{n} m_i(\boldsymbol{x})^2 S_i(\theta \mid \boldsymbol{x}) & \sum_{i=1}^{n} m_i(\boldsymbol{x}) S_i(\theta \mid \boldsymbol{x}) \\ \sum_{i=1}^{n} m_i(\boldsymbol{x}) S_i(\theta \mid \boldsymbol{x}) & \sum_{i=1}^{n} S_i(\theta \mid \boldsymbol{x}) \end{bmatrix}, \tag{2.17}$$

$$R_{1n}(\theta \mid \boldsymbol{x}) := \begin{bmatrix} \sum_{i=1}^{n} m_i(\boldsymbol{x})^3 \left(h_i(\theta \mid \boldsymbol{x}) - h_i^3(\theta \mid \boldsymbol{x})\right) & \sum_{i=1}^{n} m_i(\boldsymbol{x})^2 \left(h_i(\theta \mid \boldsymbol{x}) - h_i^3(\theta \mid \boldsymbol{x})\right) \\ \sum_{i=1}^{n} m_i(\boldsymbol{x})^2 \left(h_i(\theta \mid \boldsymbol{x}) - h_i^3(\theta \mid \boldsymbol{x})\right) & \sum_{i=1}^{n} m_i(\boldsymbol{x}) \left(h_i(\theta \mid \boldsymbol{x}) - h_i^3(\theta \mid \boldsymbol{x})\right) \end{bmatrix}, \tag{2.18}$$

and

$$R_{2n}(\theta \mid \boldsymbol{x}) := \begin{bmatrix} \sum_{i=1}^{n} m_i(\boldsymbol{x})^2 \left(h_i(\theta \mid \boldsymbol{x}) - h_i^3(\theta \mid \boldsymbol{x})\right) & \sum_{i=1}^{n} m_i(\boldsymbol{x}) \left(h_i(\theta \mid \boldsymbol{x}) - h_i^3(\theta \mid \boldsymbol{x})\right) \\ \sum_{i=1}^{n} m_i(\boldsymbol{x}) \left(h_i(\theta \mid \boldsymbol{x}) - h_i^3(\theta \mid \boldsymbol{x})\right) & \sum_{i=1}^{n} \left(h_i(\theta \mid \boldsymbol{x}) - h_i^3(\theta \mid \boldsymbol{x})\right) \end{bmatrix}, \tag{2.19}$$

where $S_i(\theta \mid \boldsymbol{x}) = \operatorname{sech}^2(\beta m_i(\boldsymbol{x}) + B)$ and $h_i(\theta \mid \boldsymbol{x}) = \tanh(\beta m_i(\boldsymbol{x}) + B)$.

Lemma 4. *Let W_{1n} and W_{2n} be as in* (2.16)*, then*

$$\frac{1}{n}\mathbb{E}_0^{(n)} \left(W_{1n}(\theta_0 \mid \boldsymbol{x})\right)^2 < \infty \qquad \frac{1}{n}\mathbb{E}_0^{(n)} \left(W_{2n}(\theta_0 \mid \boldsymbol{x})\right)^2 < \infty$$

Proof. See the lemma 2.1 in Ghosal et al. (2020). □

Lemma 5. *Let p_1 and p_2 be any two density functions. Then,*

$$\mathbb{E}_{P_1}\left(\left|\log\frac{p_1}{p_2}\right|\right) \leq KL(P_1, P_2) + \frac{2}{e}$$

Proof. See the lemma 4 in Lee (2000). □

Lemma 6. *Let $T_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \left(m_i(\boldsymbol{x}) - \frac{1}{n}\sum_{i=1}^n m_i(\boldsymbol{x})\right)^2$. Suppose Assumptions 1, 2 and 3 hold, then*

$$T_n(\boldsymbol{x}) = O_p(1), \quad 1/T_n(\boldsymbol{x}) = O_p(1)$$

Proof. See the theorem 1.4 in Ghosal et al. (2020). □

2.9 Taylor expansion for log-likelihood

Lemma 7. *Consider the term $(\theta-\theta_0)^\top H_n(\theta_0 \mid \boldsymbol{x})(\theta-\theta_0)$ where H_n is the same as in (2.17). Then, for some $C_1, C_2 > 0$, we have*

$$\mathbb{P}_0^{(n)}\left(C_1 n\|\theta-\theta_0\|_2^2 \leq (\theta-\theta_0)^\top H_n(\theta_0 \mid \boldsymbol{x})(\theta-\theta_0) \leq C_2 n\|\theta-\theta_0\|_2^2\right) \to 1, n\to\infty$$

Proof. For some $M_1, M_2 > 0$, let $\mathcal{A}_{1n} = \{\boldsymbol{x} : T_n(\boldsymbol{x}) \leq M_1\}$, $\mathcal{A}_{2n} = \{\boldsymbol{x} : T_n(\boldsymbol{x}) \geq M_2\}$, and $\mathcal{A}_n = \mathcal{A}_{1n} \cap \mathcal{A}_{2n}$, then $\mathbb{P}_0^{(n)}(\mathcal{A}_n) \to 1$.

This is because by Lemma 6, there exists M_1 and M_2 such that

$$\mathbb{P}_0^{(n)}(T_n > M_1) = \mathbb{P}_0^{(n)}(1/T_n < 1/M_1) \to 0$$
$$\mathbb{P}_0^{(n)}(T_n < M_2) = \mathbb{P}_0^{(n)}(1/T_n > 1/M_2) \to 0$$

The remaining part of the proof works with only $\boldsymbol{x} \in \mathcal{A}_n$. Let $e_1^H \geq e_2^H$ be the eigenvalues of $H_n(\theta_0 \mid \boldsymbol{x})$. The trace of $H_n(\theta_0 \mid \boldsymbol{x})$, is

$$tr\left(H_n(\theta_0 \mid \boldsymbol{x})\right) = e_1^H + e_2^H = \sum_{i=1}^n \operatorname{sech}^2\left(\beta_0 m_i(\boldsymbol{x}) + B_0\right)\left(m_i^2(\boldsymbol{x}) + 1\right) \leq n(1+\gamma^2)$$

where we used $|m_i(\boldsymbol{x})| \leq \gamma$ based on Assumption 1. Note (2.7) in Ghosal et al. (2020) gives a lower bound of e_2^H:

$$e_2^H \geq \frac{\operatorname{sech}^4(\beta_0\gamma + |B_0|)}{1+\gamma^2} n T_n(\boldsymbol{x}), \tag{2.20}$$

where $T_n(\boldsymbol{x})$ is as in the Lemma 6. By spectral decomposition of $H_n(\theta_0 \mid \boldsymbol{x})$,

$$\begin{aligned}(\theta-\theta_0)^\top H_n(\theta_0 \mid \boldsymbol{x})(\theta-\theta_0) &\leq e_1^H \left\{(\beta-\beta_0)^2+(B-B_0)^2\right\} \\ &\leq \left(n(1+\gamma^2)-e_2^H\right)\left\{(\beta-\beta_0)^2+(B-B_0)^2\right\} \\ &\leq n\left((1+\gamma^2)-\frac{\operatorname{sech}^4(\beta_0\gamma+|B_0|)}{1+\gamma^2}T_n(\boldsymbol{x})\right)\left\{(\beta-\beta_0)^2+(B-B_0)^2\right\}\end{aligned} \tag{2.21}$$

Also,

$$\begin{aligned}(\theta-\theta_0)^\top H_n(\theta_0 \mid \boldsymbol{x})(\theta-\theta_0) &\geq e_2^H \left\{(\beta-\beta_0)^2+(B-B_0)^2\right\} \\ &\geq \frac{\operatorname{sech}^4(\beta_0\gamma+|B_0|)}{1+\gamma^2}nT_n(\boldsymbol{x})\left\{(\beta-\beta_0)^2+(B-B_0)^2\right\}\end{aligned} \tag{2.22}$$

Since $M_2 \leq T_n(\boldsymbol{x}) \leq M_1$ for every $\boldsymbol{x} \in \mathcal{A}_n$, the proof follows. □

Lemma 8. *For R_{1n} and R_{2n} as in* (2.18) *and* (2.19) *respectively, let*

$$3R_n(\tilde{\theta}, \theta-\theta_0 \mid \boldsymbol{x}) \tag{2.23}$$

$$= (\beta-\beta_0)(\theta-\theta_0)^\top R_{1n}(\tilde{\theta} \mid \boldsymbol{x})(\theta-\theta_0) + (B-B_0)(\theta-\theta_0)^\top R_{2n}(\tilde{\theta} \mid \boldsymbol{x})(\theta-\theta_0) \tag{2.24}$$

where $R_n = R_n(\tilde{\theta}, \theta-\theta_0 \mid \boldsymbol{x})$ and $\tilde{\theta} = \theta_0 + c(\theta-\theta_0)$ $0<c<1$. Then, as $n \to \infty$ for some $C_1, C_2 > 0$ we have

$$\mathbb{P}_0^{(n)}(M_1 n\Delta^* \leq R_n \leq M_2 n\Delta^*) \to 1,$$

where $\Delta^* = ((\beta-\beta_0)\gamma + (B-B_0))\|\theta_0-\theta\|_2^2$.

Proof. For some $M_1, M_2 > 0$, let $\mathcal{A}_{1n} = \{\boldsymbol{x} : T_n(\boldsymbol{x}) \leq M_1\}$, $\mathcal{A}_{2n} = \{\boldsymbol{x} : T_n(\boldsymbol{x}) \geq M_2\}$.
Let $\mathcal{A}_n = \mathcal{A}_{1n} \cap \mathcal{A}_{2n}$, then $\mathbb{P}_0^{(n)}(\mathcal{A}_n) \to 1$.
This is because by Lemma 6, there exists $M_1, M_2 > 0$ such that

$$\mathbb{P}_0^{(n)}(T_n > M_1) = \mathbb{P}_0^{(n)}(1/T_n < 1/M_1) \to 0$$

$$\mathbb{P}_0^{(n)}(T_n < M_2) = \mathbb{P}_0^{(n)}(1/T_n > 1/M_2) \to 0$$

The remaining part of the proof works with only $\boldsymbol{x} \in \mathcal{A}_n$.

The determinant of $R_{1n}(\tilde{\theta} \mid \boldsymbol{x})$ is:

$$\begin{aligned} &det(R_{1n}(\tilde{\theta} \mid \boldsymbol{x})) \\ &= \frac{1}{2}\sum_{i,j=1}^{n} m_i(\boldsymbol{x})m_j(\boldsymbol{x})\left(h_i(\boldsymbol{x}) - h_i^3(\boldsymbol{x})\right)\left(h_j(\boldsymbol{x}) - h_j^3(\boldsymbol{x})\right)\left(m_i(\boldsymbol{x}) - m_j(\boldsymbol{x})\right)^2 \end{aligned}$$

where $h_i(\boldsymbol{x}) = \tanh(\tilde{\beta}m_i(\boldsymbol{x}) + \tilde{B})$. Since $\tanh(\cdot) - \tanh^3(\cdot)$ has maximum value 0.38 at $\sqrt{3}/3$ and $m_i(\boldsymbol{x}) \leq \gamma$ by Assumption 1. Therefore,

$$|det(R_{1n}(\tilde{\theta} \mid \boldsymbol{x}))| \leq \frac{1}{2}\gamma^2(0.38)^2\sum_{i=1}^{n}\sum_{j=1}^{n}\left(m_i(\boldsymbol{x}) - m_j(\boldsymbol{x})\right)^2 = \gamma^2(0.38)^2n^2T_n(\boldsymbol{x}).$$

The trace of $R_{1n}(\tilde{\theta} \mid \boldsymbol{x})$ is:

$$tr(R_{1n}(\tilde{\theta} \mid \boldsymbol{x})) = \sum_{i=1}^{n} m_i(\boldsymbol{x})(h_i(\boldsymbol{x}) - h_i^3(\boldsymbol{x}))\left(m_i^2(\boldsymbol{x}) + 1\right) \leq n0.38\gamma(1+\gamma^2).$$

Let $e_1^{R_{1n}} \geq e_2^{R_{1n}}$ be eigenvalues of $R_{1n}(\tilde{\theta} \mid \boldsymbol{x})$.

$$e_2^{R_{1n}} \geq \frac{e_1^{R_{1n}}e_2^{R_{1n}}}{e_1^{R_{1n}} + e_2^{R_{1n}}} = \frac{det(R_{1n}(\tilde{\theta} \mid \boldsymbol{x}))}{tr(R_{1n}(\tilde{\theta} \mid \boldsymbol{x}))} \qquad \geq -\frac{\gamma^2(0.38)^2n^2T_n(\boldsymbol{x})}{n0.38\gamma(1+\gamma^2)} = -\frac{0.38\gamma}{1+\gamma^2}nT_n(\boldsymbol{x}).$$

Therefore,

$$(\theta - \theta_0)^\top R_{1n}(\tilde{\theta} \mid \boldsymbol{x})(\theta - \theta_0) \geq e_2^{R_{1n}}||\theta - \theta_0||_2^2 \geq -\frac{0.38\gamma}{1+\gamma^2}nT_n(\boldsymbol{x})||\theta - \theta_0||_2^2 \tag{2.25}$$

and

$$\begin{aligned} (\theta - \theta_0)^\top R_{1n}(\tilde{\theta} \mid \boldsymbol{x})(\theta - \theta_0) &\leq e_1^{R_{1n}}||\theta - \theta_0||_2^2 = (tr(R_{1n}(\tilde{\theta} \mid \boldsymbol{x})) - e_2^{R_{1n}}))||\theta - \theta_0||_2^2 \\ &\leq 0.38\gamma n\left((1+\gamma^2) + \frac{T_n(\boldsymbol{x})}{1+\gamma^2}\right)||\theta - \theta_0||_2^2. \end{aligned} \tag{2.26}$$

With the same argument, we can get:

$$(\theta - \theta_0)^\top R_{2n}(\tilde{\theta} \mid \boldsymbol{x})(\theta - \theta_0) \geq -\frac{0.38}{1+\gamma^2}nT_n(\boldsymbol{x})||\theta - \theta_0||_2^2, \tag{2.27}$$

$$(\theta - \theta_0)^t R_{2n}(\tilde{\theta} \mid \boldsymbol{x})(\theta - \theta_0) \leq 0.38n\left((1+\gamma^2) + \frac{T_n(\boldsymbol{x})}{1+\gamma^2}\right)||\theta - \theta_0||_2^2 \tag{2.28}$$

Using (2.25), (2.26), (2.27) and (2.28) and noting $M_2 \leq T_n(\boldsymbol{x}) \leq M_1$ for every $\boldsymbol{x} \in \mathcal{A}_n$, the proof follows. □

Lemma 9. *Let $q(\theta) \in \mathcal{Q}^{MF}$ with $\mu_1 = \log \beta_0$, $\mu_2 = B_0$, and $\sigma_1^2 = \sigma_2^2 = 1/n$, then*

$$\frac{1}{n\epsilon_n^2 \log n} KL(Q, P) \to 0, \;\; n \to \infty$$

Proof. Using the same notation in (2.3), the KL divergence is:

$$\begin{aligned} KL(Q, P) &= \mathbb{E}_{q_\beta(\beta) q_B(B)} \left(\log q_\beta(\beta) + \log q_B(B) - \log p_\beta(\beta) - \log p_B(B) \right) \\ &= KL\left(Q_\beta, P_\beta\right) + KL\left(Q_B, P_B\right) \\ &= \frac{1}{2}\left((\log \beta_0)^2 + \frac{1}{n} + B_0^2 + \frac{1}{n} - 2 \right) + \log n = o(n\epsilon_n^2 \log n), \quad \text{since } n\epsilon_n^2 \to \infty \end{aligned}$$

□

2.10 Technical details of Lemma 1

Proof of Lemma 1. Let $\mathcal{V}_{\epsilon_n} = \{|\beta - \beta_0| < \epsilon_n, |B - B_0| < \epsilon_n\}$. Then $\mathcal{V}_{\epsilon_n} \subseteq \mathcal{U}_{\sqrt{2}\epsilon_n}$ which implies $\mathcal{U}^c_{\sqrt{2}\epsilon_n} \subseteq \mathcal{V}^c_{\epsilon_n}$ which further implies

$$\log \int_{\mathcal{U}^c_{\sqrt{2}\epsilon_n}} \frac{L(\theta)}{L(\theta_0)} p(\theta) d\theta \leq \log \int_{\mathcal{V}^c_{\epsilon_n}} \frac{L(\theta)}{L(\theta_0)} p(\theta) d\theta \tag{2.29}$$

We shall now establish for some $C_0 > 0$

$$\mathbb{P}_0^{(n)} \left(\log \int_{\mathcal{V}^c_{\epsilon_n}} \frac{L(\theta)}{L(\theta_0)} p(\theta) d\theta \leq -C_0 n\epsilon_n^2 \right) \to 1, \;\; n \to \infty,$$

which in lieu of (2.29) completes the proof.

Define $\mathcal{A}_{1n} = \{\boldsymbol{x} : W_{1n}(\theta_0 \mid \boldsymbol{x})^2 + W_{1n}(\theta_0 \mid \boldsymbol{x})^2 \leq n^{2/3}\}$ and $\mathcal{A}_{2n} = \{\boldsymbol{x} : T_n(\boldsymbol{x}) \geq M\}$ for some $M > 0$. Define $\mathcal{A}_n = \mathcal{A}_{1n} \cap \mathcal{A}_{2n}$.

Here $\mathbb{P}_0^{(n)}(\mathcal{A}_n) \to 1$. This because by Markov's inequality and Lemma 4,

$$\begin{aligned} &\mathbb{P}_0^{(n)} \left(W_{1n}(\theta_0 \mid \boldsymbol{x})^2 + W_{2n}(\theta_0 \mid \boldsymbol{x})^2 > n^{2/3}\varepsilon \right) \\ &\leq \frac{1}{n^{4/3}} \mathbb{E}_0^{(n)} \left(W_{1n}(\theta_0 \mid \boldsymbol{x})^2 + W_{2n}(\theta_0 \mid \boldsymbol{x})^2 \right) \to 0. \end{aligned}$$

and by Lemma 6 $\mathbb{P}_0^{(n)}(T_n < M) = \mathbb{P}_0^{(n)}(1/T_n > 1/M) \to 0$.

We shall show for $\boldsymbol{x} \in \mathcal{A}_n$, $L(\theta)/L(\theta_0) \leq e^{-C_0 n\epsilon_n^2}$, $\forall\ \theta \in \mathcal{V}_{\epsilon_n}^c$ which implies $\forall\ \boldsymbol{x} \in \mathcal{A}_n$,

$$\log \int_{\mathcal{V}_{\epsilon_n}^c} (L(\theta)/L(\theta_0))p(\theta)d\theta = \log \int_{\mathcal{V}_{\epsilon_n}^c} (L(\theta)/L(\theta_0))p(\theta)d\theta$$
$$\leq \log \left(e^{-C_0 n\epsilon_n^2} \int_{\mathcal{V}_{\epsilon_n}^c} p(\theta)d\theta \right) \leq -C_0 n\epsilon_n^2, \tag{2.30}$$

since $p(\mathcal{V}_{\epsilon_n^c}) \leq 1$. This completes the proof since $\mathbb{P}_0^{(n)}(\mathcal{A}_n) \to 1$ as $n \to \infty$.

Next, note that $\mathcal{V}_{\epsilon_n}^c$ is given by the union of the following terms

$$V_{1n} = \{(\beta, B) : \beta - \beta_0 \geq \epsilon_n, B \geq B_0\}, V_{2n} = \{(\beta, B) : \beta - \beta_0 \geq \epsilon_n, B < B_0\}$$
$$V_{3n} = \{(\beta, B) : \beta - \beta_0 < -\epsilon_n, B \geq B_0\}, V_{4n} = \{(\beta, B) : \beta - \beta_0 < -\epsilon_n, B < B_0\}$$
$$V_{5n} = \{(\beta, B) : \beta \geq \beta_0, B - B_0 \geq \epsilon_n\}, V_{6n} = \{(\beta, B) : \beta < \beta_0, B - B_0 \geq \epsilon_n\}$$
$$V_{7n} = \{(\beta, B) : \beta \geq \beta_0, B - B_0 < -\epsilon_n\}, V_{8n} = \{(\beta, B) : \beta < \beta_0, B - B_0 < -\epsilon_n\}$$

We shall now show for $\boldsymbol{x} \in \mathcal{A}_n$ and $\theta \in V_{1n}$, $L(\theta)/L(\theta_0) \leq e^{-C_0 n\epsilon_n^2}$. The proof of other parts follow similarly.

(a) Let $\theta = (\beta, B)$ and $\theta_0' = (\beta_0 + \epsilon, B_0)$, where $\beta \geq \beta_0 + \epsilon$ and $B \geq B_0$. Also, define

$$\theta_t = \theta_0' + t(\theta - \theta_0') \text{ where } 0 < t < 1.$$

Consider a function g:

$$g(t) = f(\theta_t) = \log L(\theta_t) - \log L(\theta_0') - \Delta_n(\theta_0')^\top(\theta_t - \theta_0'),$$

where $\Delta_n(\theta) = (\nabla_\beta \log L(\theta), \nabla_B \log L(\theta))^\top$. Note that $g(t)$ is a function of t. We want to show $g(t) \leq g(0)$ provided $t > 0$. We shall instead show $g'(t) \leq 0$. By Taylor expansion,

$$g'(t) = g'(0) + g''(\tilde{t})t.$$

for some $\tilde{t} \in [0, t]$. Here, $g'(0) = 0$ and $g''(\tilde{t}) = -(\theta - \theta_0')^\top H_n(\theta_{\tilde{t}} \mid \boldsymbol{x})(\theta - \theta_0') \leq 0$ where H_n as in (2.17) is a positive definite matrix (by (2.22) in 7 and $T_n(\boldsymbol{x}) \geq 0$). Since $g(t)$ is decreasing for $0 < t < 1$, thus

$$g(1) \leq g(0) \implies f(\theta) \leq f(\theta_0')$$

(b) Similarly, let $\theta = (\beta, B)$ $\theta_0'' = (\beta_0, B_0 + \epsilon)$, where $\beta \geq \beta_0$ and $B \geq B_0 + \epsilon$. Define

$$\theta_t = \theta_0' + t(\theta - \theta_0'') \text{ where } 0 < t < 1.$$

$$h(t) = f(\theta_t) = \log L(\theta_t) - \log L(\theta_0'') - \Delta_n(\theta_0'')^\top(\theta_t - \theta_0'').$$

With similar argument in (a), we conclude that $h(1) \leq h(0) \implies f(\theta) \leq f(\theta_0'')$. Therefore,

$$\begin{aligned}
&\sup_{\theta \in V_{1n}} (\log L(\theta) - \log L(\theta_0)) \\
&\leq \sup_{\{\beta - \beta_0 \in [\epsilon_n, \epsilon], B \geq B_0\}} (\log L(\theta) - \log L(\theta_0)) + \sup_{\{\beta > \beta_0 + \epsilon, B \geq B_0\}} (\log L(\theta) - \log L(\theta_0)) \\
&\leq \sup_{\{\beta - \beta_0 \in [\epsilon_n, \epsilon], B \geq B_0\}} (\log L(\theta) - \log L(\theta_0)) + (\log L(\theta_0') - \log L(\theta_0)) \\
&\leq \sup_{\{\beta - \beta_0 \in [\epsilon_n, \epsilon], B - B_0 \in [0, \epsilon]\}} (\log L(\theta) - \log L(\theta_0)) + \sup_{\{\beta - \beta_0 \in [\epsilon_n, \epsilon], B > B_0 + \epsilon\}} (\log L(\theta) - \log L(\theta_0)) \\
&+ \log L(\theta_0') - \log L(\theta_0)) \\
&\leq \sup_{\{\beta - \beta_0 \in [\epsilon_n, \epsilon], B - B_0 \in [0, \epsilon]\}} (\log L(\theta) - \log L(\theta_0)) + \sup_{\{\beta \geq \beta_0, B > B_0 + \epsilon\}} (\log L(\theta) - \log L(\theta_0)) \\
&+ \log L(\theta_0') - \log L(\theta_0)) \\
&\leq \sup_{\{\beta - \beta_0 \in [\epsilon_n, \epsilon], B - B_0 \in [0, \epsilon]\}} (\log L(\theta) - \log L(\theta_0)) + \log L(\theta_0'') - \log L(\theta_0) \\
&+ \log L(\theta_0') - \log L(\theta_0) \\
&\leq \sup_{\{\beta - \beta_0 \in [\epsilon_n, \epsilon], B - B_0 \in [0, \epsilon]\}} 3(\log L(\theta) - \log L(\theta_0)) \leq -C_0 n \epsilon_n^2 \qquad (2.31)
\end{aligned}$$

where the second inequality follows from (a) and fifth inequality follows from (b) above. Finally for the last inequality, consider Taylor expansion for $\log L(\theta)$ upto the second order

$$\log L(\theta) - \log L(\theta_0) = W_n(\theta_0 \mid \boldsymbol{x})^\top(\theta - \theta_0) - \frac{1}{2}(\theta - \theta_0)^\top H_n(\tilde{\theta} \mid \boldsymbol{x})(\theta - \theta_0)$$

where $\tilde{\theta} = \theta_0 + c(\theta - \theta_0)$, $0 < c < 1$ and W_n and H_n are as defined in (2.16) and (2.17) respectively.

By Cauchy Schwarz inequality,

$$\begin{aligned}
|W_n(\theta_0 \mid \boldsymbol{x})^\top(\theta - \theta_0))| &\leq \left(\left(W_{1n}(\theta_0 \mid \boldsymbol{x})^2 + W_{2n}(\theta_0 \mid \boldsymbol{x})^2\right) \|\theta - \theta_0\|_2^2 + 1\right) \\
&\leq n^{2/3}\|\theta - \theta_0\|_2^2 + 1
\end{aligned}$$

for every $\boldsymbol{x} \in \mathcal{A}_n$. Further

$$\begin{aligned}\log L(\theta) - \log L(\theta_0) &\leq n^{2/3}\|\theta - \theta_0\|_2^2 + 1 - \frac{1}{2}(\theta - \theta_0)^\top H_n(\tilde{\theta} \mid \boldsymbol{x})(\theta - \theta_0) \\ &\leq n^{2/3}\|\theta - \theta_0\|_2^2 + 1 - \frac{\operatorname{sech}^4(\tilde{\beta}\gamma + |\tilde{B}|)}{1+\gamma^2} n T_n(\boldsymbol{x})\|\theta - \theta_0\|_2^2 \\ &\leq \left(n^{2/3} + \frac{1}{||\theta - \theta_0||_2^2} - \frac{\operatorname{sech}^4(\tilde{\beta}\gamma + |\tilde{B}|)}{1+\gamma^2}\frac{n}{M} \right) \|\theta - \theta_0\|_2^2\end{aligned}$$

where the second inequality is a consequence of the lower bound (2.20) and the third inequality holds since $\boldsymbol{x} \in \mathcal{A}_n$. Taking sup over the set $\{\beta - \beta_0 \in [\epsilon_n, \epsilon], B - B_0 \in [0, \epsilon]\}$ on both sides,

$$\begin{aligned}&\sup_{\{\beta-\beta_0\in[\epsilon_n,\epsilon],B-B_0\in[0,\epsilon]\}} (\log L(\theta) - \log L(\theta_0)) \\ &\leq \sup_{\{\beta-\beta_0\in[\epsilon_n,\epsilon],B-B_0\in[0,\epsilon]\}} \left(n^{2/3} + \frac{1}{\epsilon_n^2} - \frac{\operatorname{sech}^4((\beta_0+\epsilon)\gamma + (B_0+\epsilon))}{1+\gamma^2}\frac{n}{M} \right) \|\theta - \theta_0\|_2^2 \\ &\leq -C_0 n\epsilon_n^2 \qquad (2.32)\end{aligned}$$

for some $C_0 > 0$ as $n \to \infty$ since $n^{2/3}$ and $1/\epsilon_n^2 = o(n)$. This completes the proof. □

2.11 Technical details of Lemma 2

Lemma 10. *Let L_0 and $L(\theta_0)$ represent the true likelihood* (1.1) *and the pseudo-likelihood* (1.2) *with the true parameters θ_0, respectively. Then,*

$$\frac{1}{n\epsilon_n^2}\mathbb{E}_0^{(n)}(\log L_0 - \log L(\theta_0)) \to 0, \;\; n \to \infty.$$

Proof.

$$L_0 = \frac{e^{f_{\theta_0}(\boldsymbol{x})}}{\sum_{\boldsymbol{x}\in\{-1,1\}^n} e^{f_{\theta_0}(\boldsymbol{x})}} = \frac{e^{f_{\theta_0}(\boldsymbol{x})}}{Z_n(\theta_0)}$$

where $f_{\theta_0}(\boldsymbol{x}) = (\beta_0/2)\boldsymbol{x}^\top A_n \boldsymbol{x} + B_0 \boldsymbol{x}^\top \mathbf{1}$. Define $b(\boldsymbol{x};\theta) = (b_1(\boldsymbol{x};\theta), \cdots, b_n(\boldsymbol{x};\theta))$ where

$$b_i(\boldsymbol{x};\theta) = E(X_i \mid X_j, j \neq i) = \tanh(\beta m_i(\boldsymbol{x}) + B).$$

Then $L(\theta) = e^{g(\boldsymbol{x}, b(\boldsymbol{x};\theta))}$ where the function g for $\boldsymbol{v}, \boldsymbol{w} \in [-1,1]^n$ is defined as

$$g(\boldsymbol{v}, \boldsymbol{w}) = \sum_{i=1}^n \frac{1+v_i}{2}\log\frac{1+w_i}{2} + \frac{1-v_i}{2}\log\frac{1-w_i}{2}$$

Also, define $I(\boldsymbol{v}) = g(\boldsymbol{v}, \boldsymbol{v})$. Now, observe that

$$\begin{aligned}
\mathbb{E}_0^{(n)}(\log L_0 - \log L(\theta_0)) &= \mathbb{E}_0^{(n)}(f_{\theta_0}(\boldsymbol{x}) - g(\boldsymbol{x}, b(\boldsymbol{x};\theta_0)) - \log Z_n(\theta_0)) \\
&= \mathbb{E}_0^{(n)}(f_{\theta_0}(\boldsymbol{x}) - f_{\theta_0}(b(\boldsymbol{x};\theta_0)) + \mathbb{E}_0^{(n)}(f_{\theta_0}(b(\boldsymbol{x};\theta_0)) - I(b(\boldsymbol{x};\theta_0))) \\
&+ \mathbb{E}_0^{(n)}(I(b(\boldsymbol{x};\theta_0)) - g(x, b(\boldsymbol{x};\theta_0))) - \log Z_n(\theta_0) \\
&\leq (\mathbb{E}_0^{(n)}(f_{\theta_0}(\boldsymbol{x}) - f_{\theta_0}(b(\boldsymbol{x};\theta_0)))^2)^{1/2} + (\mathbb{E}_0^{(n)}(I(b(\boldsymbol{x};\theta_0)) - g(\boldsymbol{x}, b(\boldsymbol{x};\theta_0)))^2)^{1/2} \\
&+ \mathbb{E}_0^{(n)}(f_{\theta_0}(b(\boldsymbol{x};\theta_0)) - I(b(\boldsymbol{x};\theta_0))) - \log Z_n(\theta_0), \qquad (2.33)
\end{aligned}$$

where the last step is due to Hölder's inequality.

Under Assumption 2, mimicking the proof of Lemmas 3.2 and 3.3 in Basak and Mukherjee (2017) with n replaced by $n\epsilon_n^2$, we get

$$(\mathbb{E}_0^{(n)}(f_{\theta_0}(\boldsymbol{x}) - f_{\theta_0}(b(\boldsymbol{x};\theta_0)))^2)^{1/2} = o(n\epsilon_n^2) \qquad (2.34)$$

$$(\mathbb{E}_0^{(n)}(I(b(\boldsymbol{x};\theta_0)) - g(\boldsymbol{x}, b(\boldsymbol{x};\theta_0)))^2)^{1/2} = o(n\epsilon_n^2) \qquad (2.35)$$

Also for $r_n = \sup_{\boldsymbol{v}\in[-1,1]^n}(f_{\theta_0}(\boldsymbol{v}) - I(\boldsymbol{v}))$, we have

$$\mathbb{E}_0^{(n)}(f_{\theta_0}(b(\boldsymbol{x},\theta_0)) - I(b(\boldsymbol{x},\theta_0))) \leq r_n$$

By Theorem 1.6 in Chatterjee and Dembo (2016) with the fact $\partial^2 f_{\theta_0}/\partial x_i^2 = 0$, $i = 1, \cdots, n$, we have $-\log Z_n(\theta_0) \leq -r_n$. Therefore,

$$\mathbb{E}_0^{(n)}(f_{\theta_0}(b(\boldsymbol{x};\theta_0)) - I(b(\boldsymbol{x};\theta_0)) - \log Z_n(\theta_0)) \leq 0. \qquad (2.36)$$

Using (2.34), (2.35) and (2.36) in (2.33) completes the proof. □

Lemma 11. *Note that $L(\theta)$ is not a valid density function. So, we consider $\tilde{L}(\theta) = L(\theta)/J_n(\theta)$ where $J_n(\theta) = \sum_{\boldsymbol{x}\in\{-1,1\}^n} L(\theta)$ such that $\sum_{\boldsymbol{x}\in\{-1,1\}^n} \tilde{L}(\theta) = 1$. Then for every θ,*

$$J_n(\theta) \leq \beta\epsilon_n\sqrt{n(1+\gamma^2)/2} + o(n\epsilon_n^2)\left(\log 3\sqrt{2} - \log\epsilon_n\right).$$

Proof. Let $N_n(\epsilon_n) := \{i \in [n] : |\lambda_i(A_n)| > \epsilon_n/\sqrt{2}\}$ and with the mean field condition in the Assumption 2, it is easy to note that

$$\frac{|N_n(\epsilon_n)|}{n} \leq \frac{2}{n\epsilon_n^2} \sum_{i\in[n]} \lambda_i \left(A_n\right)^2 = \frac{2}{n\epsilon_n^2} \sum_{i,j=1}^{n} A_n(i,j)^2 \to 0, \;\; n \to \infty \tag{2.37}$$

Set $k_n = |N_n(\epsilon_n)|$ and let $D_{n,0}(\epsilon_n)$ be a $\epsilon_n\sqrt{n/2}$ net of the set $\{\boldsymbol{f} \in \mathbb{R}^{k_n} : \sum f_i^2 \leq n\}$ of size at most $(3\sqrt{2}/\epsilon_n)^{k_n}$. The existence of such a net is standard (see for example Lemma 2.6 in Milman and Schechtman (1986)).

Let $\{\boldsymbol{p}_1, \cdots, \boldsymbol{p}_n\}$ be the eigen vectors of A_n. Then setting

$$D_{n,1}(\epsilon_n) := \left\{ \sum_{i\in N_n(\epsilon_n)} c_i\lambda_i(A_n)\boldsymbol{p}_i, \boldsymbol{c} \in D_{n,0}(\epsilon_n) \right\}$$

We claim $D_{n,1}(\epsilon_n)$ is $\epsilon_n\sqrt{n(1+\gamma^2)/2}$ of the set $\{A_n\boldsymbol{x} : \boldsymbol{x} \in \{-1,1\}^n\}$. Indeed any $\boldsymbol{x} \in \{-1,1\}^n$ can be written as $\sum_{i=1}^n f_i\boldsymbol{p}_i$ where $\sum_{i=1}^n f_i^2 = \sum x_i^2 = n$. In particular, it means $\sum_{i\in N_n(\epsilon_n)} f_i \leq n$, which implies there exists a $\boldsymbol{c} \in D_{n,0}(\epsilon_n)$ such that $||\boldsymbol{c} - \boldsymbol{f}|| \leq \epsilon_n\sqrt{n/2}$.

Let $\sum_{i\in N_n(\epsilon_n)} c_i\lambda_i(A_n)\boldsymbol{p}_i \in D_{n,1}(\epsilon_n)$, then

$$\begin{aligned} ||A_n\boldsymbol{x} - \sum_{i\in N_n(\epsilon_n)} c_i\lambda_i(A_n)\boldsymbol{p}_i||_2^2 &= \sum_{i\in N_n(\epsilon_n)} (c_i - f_i)^2\lambda_i(A_n)^2 + \sum_{i\notin N_n(\epsilon_n)} \lambda_i(A_n)^2 f_i^2 \\ &\leq \frac{\gamma^2 n\epsilon_n^2}{2} + \frac{n\epsilon_n^2}{2} \end{aligned}$$

where the last inequality is a consequence of $\max_{i\in[n]} |\lambda_i(A_n)| \leq \max_{i\in[n]} \sum_{j=1}^n |A_n(i,j)| \leq \gamma$ and the definition of the set $N_n(\epsilon_n)$.

In particular for any $\boldsymbol{x} \in \{-1,1\}^n$, there exists at least one $\boldsymbol{p} \in D_{n,1}(\epsilon_n)$ such that $||\boldsymbol{p} - m(\boldsymbol{x})|| \leq \epsilon_n\sqrt{n(1+\gamma^2)/2}$. For any $\boldsymbol{p} \in D_{n,1}(\epsilon_n)$, let

$$\mathcal{P}(\boldsymbol{p}) := \left\{ \boldsymbol{x} \in \{-1,1\}^n : ||\boldsymbol{p} - m(\boldsymbol{x})|| \leq \epsilon_n\sqrt{n(1+\gamma^2)/2} \right\}.$$

Therefore,

$$\sum_{\boldsymbol{x}\in\{-1,1\}^n} e^{g(\boldsymbol{x},b(\boldsymbol{x};\theta))} = \sum_{\boldsymbol{p}\in D_{n,1}(\epsilon_n)} \sum_{\boldsymbol{x}\in\mathcal{P}(\boldsymbol{p})} e^{g(\boldsymbol{x},b(\boldsymbol{x};\theta))}$$

Setting $\boldsymbol{u}(\boldsymbol{p}) := \tanh(\beta \boldsymbol{p} + B)$ if $||\boldsymbol{p} - m(\boldsymbol{x})|| \leq \epsilon_n \sqrt{n(1+\gamma^2)/2}$, then we have

$$|g(\boldsymbol{x}, b(\boldsymbol{x};\theta)) - g(\boldsymbol{x}, \boldsymbol{u}(\boldsymbol{p}))| \leq 2\beta \sum_{i=1}^{n} |m_i(\boldsymbol{x}) - p_i| \leq 2\beta\epsilon_n \sqrt{n(1+\gamma^2)/2}$$

Finally,

$$\begin{aligned}\sum_{\boldsymbol{x}\in\{-1,1\}^n} e^{g(\boldsymbol{x},b(\boldsymbol{x};\theta))} &\leq e^{2\beta\epsilon_n\sqrt{n(1+\gamma^2)/2}} \sum_{\boldsymbol{p}\in D_{n,1}(\epsilon_n)} \sum_{\boldsymbol{x}\in\mathcal{P}(\boldsymbol{p})} e^{g(\boldsymbol{x},u(\boldsymbol{p}))} \\ &\leq e^{2\beta\epsilon_n\sqrt{n(1+\gamma^2)/2}} \sum_{\boldsymbol{p}\in D_{n,1}(\epsilon_n)} \sum_{\boldsymbol{x}\in\{-1,1\}^n} e^{g(\boldsymbol{x},u(\boldsymbol{p}))} = e^{2\beta\epsilon_n\sqrt{n(1+\gamma^2)/2}} |D_{n,1}(\epsilon_n)|\end{aligned}$$

where the last equality follows since $\sum_{\boldsymbol{x}\in\{-1,1\}^n} e^{g(x,u)} = 1$ for any $u \in [-1,1]^n$. Therefore,

$$\log J_n(\theta) \leq \beta\epsilon_n \sqrt{n(1+\gamma^2)/2} + \log |D_{n,1}(\epsilon_n)|$$

Since $|D_{n,1}(\epsilon_n)| = |D_{n,0}(\epsilon_n)|$, therefore

$$\log |D_{n,1}(\epsilon_n)| \leq |N_n(\epsilon_n)|(\log 3\sqrt{2} - \log \epsilon_n)$$

The proof follows since $|N_n(\epsilon_n)| = o(n\epsilon_n^2)$. □

Lemma 12. *Define* $\mathcal{V}_{\epsilon_n} := \{\theta : |\beta - \beta_0| < \epsilon_n, |B - B_0| < \epsilon_n\}$. *Then,*

$$\mathcal{V}_{\epsilon_n} \subseteq \mathcal{K}_{\epsilon_n}, \text{ for } n \text{ sufficiently large}$$

where $\mathcal{K}_{\epsilon_n} := \{\theta : \mathbb{E}_0^{(n)}(\log(L(\theta_0)/L(\theta))) < 3n\epsilon_n^2\}$.

Proof. For any $\theta \in \mathcal{V}_{\epsilon_n}$, using the decomposition in (2.43), we get

$$\mathbb{E}_0^{(n)} (\log L(\theta_0) - \log L(\theta)) = \mathbb{E}_0^{(n)} \left(-\textcircled{1} - \textcircled{2} + \textcircled{3} - \textcircled{4} \right) \leq 3n\epsilon_n^2$$

where the last inequality is justified next.

For some $M > 0$ using Lemma 4, we get

$$\begin{aligned}-\mathbb{E}_0^{(n)}(\textcircled{1}) &= (\beta_0 - \beta)\mathbb{E}_0^{(n)}(W_{1n}(\theta_0|\boldsymbol{x}) \leq \sqrt{n}|\beta_0 - \beta| \left(\frac{1}{n} \mathbb{E}_0^{(n)} \left(W_{1n}(\theta_0|\boldsymbol{x}) \right)^2 \right)^{1/2} \\ &\leq M\sqrt{n}\epsilon_n\end{aligned}$$

$$-\mathbb{E}_0^{(n)}(\textcircled{2}) = (B_0 - B)\mathbb{E}_0^{(n)}(W_{2n}(\theta_0|\boldsymbol{x}) \leq \sqrt{n}|B_0 - B| \left(\frac{1}{n}\mathbb{E}_0^{(n)}\left(W_{2n}(\theta_0|\boldsymbol{x})\right)^2\right)^{1/2}$$
$$\leq M\sqrt{n}\epsilon_n$$

By relation (2.21), we get

$$\mathbb{E}_0^{(n)}(\textcircled{3}) \leq n||\theta - \theta_0||_2^2 \left((1+\gamma^2) - \frac{\operatorname{sech}^4(\beta_0\gamma + |B_0|)}{1+\gamma^2}\mathbb{E}_0^{(n)}\left(T_n(\boldsymbol{x})\right)\right) \leq 2(1+\gamma^2)n\epsilon_n^2$$

By relation (2.44), we get

$$-\mathbb{E}_0^{(n)}(\textcircled{4}) \leq \frac{0.38n}{3(1+\gamma^2)}\mathbb{E}_0^{(n)}\left(T_n(\boldsymbol{x})\right)||\theta - \theta_0||_2^2(|\beta - \beta_0|\gamma + |B - B_0|) \leq \frac{0.38\gamma^2(1+\gamma)}{3(1+\gamma^2)}n\epsilon_n^3$$

□

Lemma 13. *With prior distribution $p(\theta)$ as in (2.1), we have*

$$\int_{\mathcal{V}_{\epsilon_n}} p(\theta)d\theta \geq C\epsilon_n^2, \quad \text{for some } C > 0$$

Proof. By mean value theorem with $\beta^\star \in [\beta_0 - \epsilon_n, \beta_0 + \epsilon_n]$ and $B^\star \in [B_0 - \epsilon_n, B_0 + \epsilon_n]$,

$$\begin{aligned}
\int_{V_{\epsilon_n}} p(\theta)d\theta &= \int_{\beta_0-\epsilon}^{\beta_0+\epsilon} \frac{1}{\beta\sqrt{2\pi}} e^{-\frac{(\log\beta)^2}{2}} d\beta \int_{B_0-\epsilon}^{B_0+\epsilon} \frac{1}{\sqrt{2\pi}} e^{-\frac{B^2}{2}} dB \\
&= \frac{2\epsilon_n}{\beta^\star\sqrt{2\pi}} e^{-\frac{(\log\beta^\star)^2}{2}} \frac{2\epsilon_n}{\sqrt{2\pi}} e^{-\frac{(B^\star)^2}{2}} \\
&= \exp\left(-(\log\pi - \log 2 - 2\log\epsilon_n) - \frac{1}{2}\left(2\log\beta^\star + (\log\beta^\star)^2 + (B^\star)^2\right)\right) \\
&\geq \exp\left(-(\log\pi - \log 2 - 2\log\epsilon_n) - \frac{1}{2}\left(2u_1 + \tilde{u}_1 + u_2\right)\right) \\
&\geq Ce^{2\log\epsilon_n} = C\epsilon_n^2
\end{aligned}$$

where the above result follow since $\epsilon_n \to 0$ implies $u_1 \leq \max(\log(\beta_0+1), \log(\beta_0+1))$, $\tilde{u}_1 \leq \max((\log(\beta_0-1))^2, (\log(\beta_0+1))^2)$, and $u_2 = \max((B_0-1)^2, (B_0+1)^2)$. □

Proof of Lemma 2. Let $L^* = \int L(\theta)p(\theta)d\theta$, $J_n^* = \sum_{\boldsymbol{x}\in\{-1,1\}^n} L^*$. Then,

$$J_n^* = \sum_{\boldsymbol{x}\in\{-1,1\}^n} \int L(\theta)p(\theta)d\theta.$$

Since $L(\theta)p(\theta) > 0$, Tonelli's theorem allows for interchange of the order of summation and integral. Using Lemma 11 and $-\log \epsilon_n = O(\log n)$, we get

$$\begin{aligned} J_n^* &= \int \sum_{\boldsymbol{x}\in\{-1,1\}^n} L(\theta)p(\theta)d\theta = \int J_n(\theta)p(\theta)d\theta \\ &= \epsilon_n\sqrt{n(1+\gamma^2)/2}E_P(\beta) + o(n\epsilon_n^2)(\log 3\sqrt{2} - \log\epsilon_n) \\ &= \epsilon_n\sqrt{ne(1+\gamma^2)/2} + o(n\epsilon_n^2)(\log 3\sqrt{2} - \log\epsilon_n) = o(n\epsilon_n^2\log n) \end{aligned} \tag{2.38}$$

Also, by Lemma 11 and $-\log \epsilon_n = O(\log n)$,

$$\log J_n(\theta_0) = \beta_0\epsilon_n\sqrt{n(1+\gamma^2)/2} + o(n\epsilon_n^2)(\log 3\sqrt{2} - \log\epsilon_n) = o(n\epsilon_n^2\log n) \tag{2.39}$$

$$P_0^n\left(\left|\log\int (L(\theta)/L(\theta_0))p(\theta)d\theta\right| > Cn\epsilon_n^2\log n\right) \tag{2.40}$$

$$\begin{aligned} &\leq \frac{1}{Cn\epsilon_n^2\log n}\mathbb{E}_0^{(n)}\left(\left|\log\int (L(\theta)/L(\theta_0))p(\theta)d\theta\right|\right) \\ &= \frac{1}{Cn\epsilon_n^2\log n}\mathbb{E}_0^{(n)}\left(|\log\left(L^*/L(\theta_0)\right)|\right) \\ &\leq \frac{1}{Cn\epsilon_n^2\log n}\left(KL(L_0,\tilde{L}^*) + KL(L_0,\tilde{L}(\theta_0)) + \left|\log\frac{J_n^*}{J_n(\theta_0)}\right| + \frac{4}{e}\right) \\ &\leq \frac{2}{Cn\epsilon_n^2\log n}\Big(\mathbb{E}_0^{(n)}\left(\log L_0 - \log L(\theta_0)\right) + \mathbb{E}_0^{(n)}(\log L(\theta_0) - \log L^*) \\ &+ 2(\log J_n^* + \log J_n(\theta_0)) + \frac{4}{e}\Big) \end{aligned} \tag{2.41}$$

where the second last step follows from Lemma 5.

Then, using the set $\mathcal{K}_{\epsilon_n}$ in Lemma 12, we get

$$
\begin{aligned}
&\mathbb{E}_0^{(n)}(\log L(\theta_0) - \log L^*)) \\
&= \mathbb{E}_0^{(n)}(\log L(\theta_0) - \log \int L(\theta)p(\theta)d\theta) \\
&\leq \mathbb{E}_0^{(n)}\left(\log(L(\theta_0)/\int_{\mathcal{K}_{\epsilon_n^2}} L(\theta)p(\theta)d\theta)\right) \\
&\leq \mathbb{E}_0^{(n)}\left(\log L(\theta_0) - \log\left(\frac{p(\mathcal{K}_{\epsilon_n^2})}{p(\mathcal{K}_{\epsilon_n^2})}\int_{\mathcal{K}_{\epsilon_n^2}} L(\theta)p(\theta)d\theta))\right)\right) \\
&\leq \mathbb{E}_0^{(n)}(\log L(\theta_0)) - \log(p(\mathcal{K}_{\epsilon_n^2}) + \mathbb{E}_0^{(n)}\left(E_{p|\mathcal{K}_{\epsilon_n^2}}(-\log L(\theta))\right) \\
&\leq -\log(p(\mathcal{K}_{\epsilon_n^2})) + \mathbb{E}_0^{(n)}\left(\log L(\theta_0) - \int_{\mathcal{K}_{\epsilon_n^2}} \log L(\theta)p|\mathcal{K}_{\epsilon_n^2}(\theta)d\theta\right) \text{ Jensen's Inequality} \\
&= -\log(p(\mathcal{K}_{\epsilon_n^2})) + \int_{\mathcal{K}_{\epsilon_n^2}} \mathbb{E}_0^{(n)}(\log(L(\theta_0) - L(\theta))p|\mathcal{K}_{\epsilon_n^2}(\theta)d\theta \\
&\leq -2\log(C'\epsilon_n^2) + 3n\epsilon_n^2 = o(n\epsilon_n^2 \log n) \qquad (2.42)
\end{aligned}
$$

where the last line follows from Lemma 12 and Lemma 13. The final order is because $-\log \epsilon_n = O(\log n)$ and $n\epsilon_n^2 \to \infty$ and $\log n \to \infty$.

The proof follows by using relations (2.42), (2.39) and (2.38) in (2.40). □

2.12 Technical details of Lemma 3

Lemma 14. *Let $q(\theta) \in \mathcal{Q}^{MF}$ with $\mu_1 = \log \beta_0$, $\mu_2 = B_0$, and $\sigma_1^2 = \sigma_2^2 = 1/n$, then*

$$\frac{1}{n\epsilon_n^2}\int \mathbb{E}_0^{(n)}(\log L(\theta_0) - \log L(\theta))q(\theta)d\theta \lesssim 0, \;\; n \to \infty$$

Proof. Using the Taylor expansion of $\log L(\theta)$ around $\theta = \theta_0$, we get

$$
\begin{aligned}
\log L(\theta_0) - \log L(\theta) = \log L(\theta_0) - \log L(\theta_0) - \underbrace{(\beta - \beta_0)W_{1n}(\theta_0|\boldsymbol{x})}_{①} - \underbrace{W_{2n}(\theta_0|\boldsymbol{x})(B - B_0)}_{②} \\
+ \underbrace{\frac{1}{2}(\theta - \theta_0)^\top H_n(\theta_0|\boldsymbol{x})(\theta - \theta_0)}_{③} - \underbrace{R_n(\tilde{\theta}, \theta - \theta_0|\boldsymbol{x})}_{④} \qquad (2.43)
\end{aligned}
$$

W_{1n}, W_{2n} is as in (2.16), H_n is as in (2.17) and $R_n(\tilde{\theta}, \theta-\theta_0|\boldsymbol{x})$ is defined in (2.23). Therefore,

$$\begin{aligned}
&\int \mathbb{E}_0^{(n)} \left(\log L(\theta_0) - \log L(\theta)\right) q(\theta) d\theta \\
&= -\int \mathbb{E}_0^{(n)} \left(\textcircled{1}\right) q(\theta) d\theta - \int \mathbb{E}_0^{(n)} \left(\textcircled{2}\right) q(\theta) d\theta \\
&+ \int \mathbb{E}_0^{(n)} \left(\textcircled{3}\right) q(\theta) d\theta - \int \mathbb{E}_0^{(n)} \left(\textcircled{4}\right) q(\theta) d\theta,
\end{aligned}$$

$$\begin{aligned}
-\frac{1}{n\epsilon_n^2} \int \mathbb{E}_0^{(n)} \left(\textcircled{1}\right) q(\theta) d\theta &= \frac{1}{n\epsilon_n^2} \int (\beta_0 - \beta) \mathbb{E}_0^{(n)} \left(W_{1n}(\theta_0|\boldsymbol{x})\right) q(\theta) d\theta \\
&\leq \frac{1}{n\epsilon_n^2} \int |\beta_0 - \beta| \sqrt{n} \left(\frac{1}{n} \mathbb{E}_0^{(n)} \left(W_{1n}(\theta_0|\boldsymbol{x})\right)^2\right)^{1/2} q(\beta) d\beta, \\
&\leq \frac{M\sqrt{n}}{n\epsilon_n^2} \int |\beta - \beta_0| q_\beta(\beta) d\beta \\
&\leq \frac{M\sqrt{n}}{n\epsilon_n^2} \left(\int (\beta - \beta_0)^2 q_\beta(\beta) d\beta\right)^{1/2} \\
&= \frac{M\sqrt{n}}{n\epsilon_n^2} \left(e^{2\log\beta_0} (e^{2/n} - 2e^{1/2n} + 1)\right)^{1/2} \sim \frac{M e^{\log\beta_0}}{n\epsilon_n^2} \to 0
\end{aligned}$$

where the second inequality above above line holds by Hölder inequality and third inequality holds by Lemma 4, for some constant M. Finally the last convergence to 0 is $n\epsilon_n^2 \to \infty$. Similarly,

$$\begin{aligned}
-\frac{1}{n\epsilon_n^2} \int \mathbb{E}_0^{(n)} \left(\textcircled{2}\right) q(\theta) d\theta &= \frac{1}{n\epsilon_n^2} \int (B_0 - B) \mathbb{E}_0^{(n)} \left(W_{2n}(\theta_0|\boldsymbol{x})\right) q(\theta) d\theta \\
&\leq \frac{\sqrt{n}}{n\epsilon_n^2} \left(\frac{1}{n} \mathbb{E}_0^{(n)} \left(W_{2n}(\theta_0|\boldsymbol{x})\right)^2\right)^{1/2} \int |B - B_0| q_B(B) dB \\
&\frac{M\sqrt{n}}{n\epsilon_n^2} \left(\int (B - B_0)^2 q(B) dB\right)^{1/2} \sim \frac{M}{n\epsilon_n^2} \to 0
\end{aligned}$$

Using the upper bound (2.21) and $n\epsilon_n^2 \to \infty$, we get

$$\begin{aligned}
&\frac{1}{n\epsilon_n^2} \int \mathbb{E}_0^{(n)} \left(\textcircled{3}\right) q(\theta) d\theta \\
&\leq \frac{1}{2\epsilon_n^2} \left((1+\gamma^2) - \frac{\operatorname{sech}^4(\beta_0\gamma + |B_0|)}{1+\gamma^2} \mathbb{E}_0^{(n)} \left(T_n(\boldsymbol{x})\right)\right) \int \left\{(\beta - \beta_0)^2 + (B - B_0)^2\right\} q(\theta) d\theta \\
&\leq \frac{(1+\gamma^2)}{2\epsilon_n^2} \left(e^{2\log\beta_0} \left(e^{2/n} - 2e^{1/2n} + 1\right) + 1/n\right) \\
&\sim \frac{(e^{2\log\beta_0} + 1)(1+\gamma^2)}{2n\epsilon_n^2} \to 0
\end{aligned}$$

where the second inequality holds since $T_n(\boldsymbol{x}) = (1/n)\sum_{i=1}^{n}\left(m_i(\boldsymbol{x}) - (1/n)m_i(\boldsymbol{x})\right)^2 \geq 0$. For the remainder term, using relations (2.25) and (2.27) in relation (2.23), we get

$$-\mathbb{E}_0^{(n)}\left(\textcircled{4}\right) \leq \frac{0.38n}{3(1+\gamma^2)}\mathbb{E}_0^{(n)}\left(T_n(\boldsymbol{x})\right)\left\{(\beta-\beta_0)^2+(B-B_0)^2\right\}\left\{(\beta-\beta_0)\gamma+(B-B_0)\right\} \tag{2.44}$$

Further,

$$\begin{aligned} & n\int\left\{(\beta-\beta_0)^2+(B-B_0)^2\right\}\left\{(\beta-\beta_0)\gamma+(B-B_0)\right\}q(\theta)d\theta \\ & = \underbrace{n\int\left\{(e^{\log\beta}-e^{\log\beta_0})^2+(B-B_0)^2\right\}\left\{(e^{\log\beta}-e^{\log\beta_0})\gamma+(B-B_0)\right\}q(\theta)d\theta}_{\textcircled{5}} \end{aligned}$$

$$\begin{aligned} \textcircled{5} &= n\int\left(e^{3\log\beta}-3e^{2\log\beta+\log\beta_0}+3e^{\log\beta+2\log\beta_0}-e^{3\log\beta_0}\right)q(\beta)d\beta && (2.45)\\ &+ n\int\left(B^3-3B_0B^2+3B_0^2B-B_0^3\right)q(B)dB && (2.46)\\ &+ n\int\left(e^{2\log\beta}-2e^{\log\beta+\log\beta_0}+e^{2\log\beta_0}\right)(B-B_0)\,q(\beta)q(B)d\beta dB && (2.47)\\ &+ n\gamma\int(B-B_0)^2\left(e^{\log\beta}-e^{\log\beta_0}\right)q(\beta)q(B)d\beta dB && (2.48) \end{aligned}$$

$$(2.45) = ne^{3\log\beta_0}\left(e^{9/2n}-3e^{2/n}+3e^{1/2n}-e^0\right)\sim 0$$

$$(2.46) = n\left(B_0^3+3B_0/n-3B_0(B_0^2+1/n)+3B_0^3-B_0^3\right)=0,$$

$$(2.47) = n\left(B_0e^{2\log\beta_0}\left(e^{2/n}-2e^{1/2n}+e^0\right)-B_0e^{2\log\beta_0}\left(e^{2/n}-2e^{1/2n}+e^0\right)\right)=0,$$

$$(2.48) = ne^{\log\beta_0}\left(\frac{\gamma}{n}e^{1/2n}-\frac{\gamma}{n}e^0\right)=\gamma e^{\log\beta_0}(e^{1/2n}-e^0)\sim 0$$

$$-\frac{1}{n\epsilon_n^2}\int\mathbb{E}_0^{(n)}\left(\textcircled{4}\right)q(\theta)d\theta \lesssim 0$$

since $T_n(\boldsymbol{x}) \leq \gamma^2$ (since by Assumption 1, $m_i(\boldsymbol{x}) \leq \gamma$) and $n\epsilon_n^2 \to \infty$. □

Poof of Lemma 3. With the q as in Lemma 14, using Markov's inequality,

$$
\begin{aligned}
&P_0^n\left(\int q(\theta)\log(L(\theta_0)/L(\theta))d\theta > Cn\epsilon_n^2\log n\right) \qquad (2.49)\\
&\le \frac{1}{Cn\epsilon_n^2\log n}\mathbb{E}_0^{(n)}\left|\int q(\theta)\log(L(\theta_0)/L(\theta))d\theta\right|\\
&\le \frac{1}{Cn\epsilon_n^2\log n}\mathbb{E}_0^{(n)}\left(\int q(\theta)\left|\log(L(\theta_0)/L(\theta))\right|d\theta\right)\\
&\le \frac{1}{Cn\epsilon_n^2\log n}\int q(\theta)\mathbb{E}_0^{(n)}\left(\left|\log(L(\theta_0)/L(\theta))\right|\right)d\theta \qquad \text{Fubini's theorem}\\
&= \frac{1}{Cn\epsilon_n^2\log n}\int q(\theta)\mathbb{E}_0^{(n)}\left(\left|\log\left(\frac{L_0}{L(\theta)}\frac{L(\theta_0)}{L_0}\right)\right|\right)d\theta\\
&= \frac{1}{Cn\epsilon_n^2\log n}\int q(\theta)\mathbb{E}_0^{(n)}\left(\left|\log\left(\frac{L_0}{\tilde{L}(\theta)}\frac{\tilde{L}(\theta_0)}{L_0}\frac{J_n(\theta_0)}{J_n(\theta)}\right)\right|\right)d\theta\\
&= \frac{1}{Cn\epsilon_n^2\log n}\int q(\theta)\mathbb{E}_0^{(n)}\left(\left|\log\left(\frac{L_0}{\tilde{L}(\theta)}\right)+\log\left(\frac{\tilde{L}(\theta_0)}{L_0}\right)+\log\left(\frac{J_n(\theta_0)}{J_n(\theta)}\right)\right|\right)d\theta\\
&\le \frac{1}{Cn\epsilon_n^2\log n}\int q(\theta)\mathbb{E}_0^{(n)}\left(\left|\log\left(\frac{L_0}{\tilde{L}(\theta)}\right)\right|+\left|\log\left(\frac{\tilde{L}(\theta_0)}{L_0}\right)\right|+\left|\log\left(\frac{J_n(\theta_0)}{J_n(\theta)}\right)\right|\right)d\theta\\
&\le \frac{1}{Cn\epsilon_n^2\log n}\int q(\theta)\left(\mathrm{KL}\left(L_0,\tilde{L}(\theta)\right)+\mathrm{KL}\left(L_0,\tilde{L}(\theta_0)\right)+\left|\log\left(\frac{J_n(\theta_0)}{J_n(\theta)}\right)\right|+\frac{4}{e}\right)d\theta \qquad (2.50)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&P_0^n\left(\int q(\theta)\log(L(\theta_0)/L(\theta))d\theta > Cn\epsilon_n^2\log n\right)\\
&= \frac{1}{Cn\epsilon_n^2\log n}\left(2\mathbb{E}_0^{(n)}(\log L_0-\log L(\theta_0))+\int q(\theta)\mathbb{E}_0^{(n)}(\log L(\theta_0)-\log L(\theta))d\theta\right)\\
&+\frac{1}{Cn\epsilon_n^2\log n}\left(2\log J_n(\theta_0)+2\int q(\theta)\log J_n(\theta)d\theta+\frac{4}{e}\right)\to 0 \qquad (2.51)
\end{aligned}
$$

where the inequality in second last step is due to Lemma 5. The last convergence to 0 is explained next. By Lemma 10 and Lemma 14 respectively, we get

$$
\begin{aligned}
\mathbb{E}_0^{(n)}(\log L_0-\log L(\theta_0)) &= o(n\epsilon_n^2)\\
\int q(\theta)\mathbb{E}_0^{(n)}(\log L(\theta_0)-\log L(\theta))d\theta &\le o(n\epsilon_n^2)
\end{aligned}
$$

By (2.39), $\log J_n(\theta_0) = o(n\epsilon_n^2\log n)$ and by Lemma 11 and $-\log\epsilon_n = O(\log n)$

$$
\int q(\theta)\log J_n(\theta)d\theta \le \epsilon_n\sqrt{n(1+\gamma^2)/2}\int \beta q(\beta)d\beta + o(n\epsilon_n^2)(\log 3\sqrt{2}-\log\epsilon_n) = o(n\epsilon_n^2)
$$

where we use $E_Q(\beta) = \exp(\log\beta_0 + 1/n)\to\beta_0$ □

2.13 Proof of Theorem 1

In this section, with dominating probability term is used to imply that under $\mathbb{P}_0^{(n)}$, the probability of the event goes to 1 as $n \to \infty$.

$$\begin{aligned} KL(Q, \Pi(|X^{(n)})) &= \int q(\theta) \log q(\theta) d\theta - \int q(\theta) \log \pi(\theta|X^{(n)}) d\theta \\ &= \int q(\theta) \log q(\theta) d\theta - \int q(\theta) \log \frac{L(\theta)p(\theta)}{\int L(\theta)p(\theta)d\theta} d\theta \\ &= KL(Q, P) - \int \log(L(\theta)/L(\theta_0)) q(\theta) d\theta + \log \int (L(\theta)/L(\theta_0)) p(\theta) d\theta \\ &= KL(Q, P) + \int \log(L(\theta_0)/L(\theta)) q(\theta) d\theta + \log \int (L(\theta)/L(\theta_0)) p(\theta) d\theta \end{aligned} \tag{2.52}$$

By Lemma 9, $KL(Q, P) = o(n\epsilon_n^2 \log n) \leq (C/3) n\epsilon_n^2 \log n$. By Lemma 3, with dominating probability

$$\int \log(L(\theta_0)/L(\theta)) q(\theta) d\theta \leq (C/3) n\epsilon_n^2 \log n$$

for any $C > 0$. By Lemma 2, with dominating probability

$$\log \int (L(\theta)/L(\theta_0)) p(\theta) d\theta \leq (C/3) n\epsilon_n^2 \log n$$

Therefore, with dominating probability, for any $C > 0$,

$$KL(Q, \Pi(|X^{(n)})) \leq Cn\epsilon_n^2$$

Further,

$$\Pi(\mathcal{U}_{\varepsilon_n}^c | X^{(n)}) = \frac{\int_{\mathcal{U}_{\varepsilon_n}^c} L(\theta)p(\theta)d\theta}{\int L(\theta)p(\theta)d\theta} = \frac{\int_{\mathcal{U}_{\varepsilon_n}^c} (L(\theta)/L(\theta_0))p(\theta)d\theta}{\int (L(\theta)/L(\theta_0))p(\theta)d\theta}$$

By Lemma 1, with dominating probability, for any $C > 0$, as $n \to \infty$

$$\int_{\mathcal{U}_{\varepsilon_n}^c} (L(\theta)/L(\theta_0)) p(\theta) d\theta \leq \exp(-C_0 n\varepsilon_n^2)$$

By Lemma 2, with dominating probability

$$\int (L(\theta)/L(\theta_0)) p(\theta) d\theta \geq \exp(-Cn\epsilon_n^2 \log n)$$

Therefore, with dominating probability

$$\Pi(\mathcal{U}^c_{\varepsilon_n}|X^{(n)}) \leq \exp(-C_0 n\varepsilon_n^2(1 - C/M_n)) \leq \exp(-C_1 n\varepsilon_n^2)$$

for any $C_0 > C_1/2$. This is because for n sufficiently large $1 - C/M_n > 1/2$.

This completes the proof.

2.14 Proof of Corollary 1

By Lemma 1, with dominating probability, there exists $C_0(r) > 0$ such that as $n \to \infty$,

$$\int_{\mathcal{U}^c_{r\varepsilon_n}} (L(\theta)/L(\theta_0))p(\theta)d\theta \leq \exp(-C_0(r)r^2 n\varepsilon_n^2)$$

Let us assume,

$$C_0(r) \geq C_2/r \quad \text{for all } r > 0 \text{ for some constant } C_2 > 0 \tag{2.53}$$

Numerical evidence for validity of this assumption been provided in the section 2.14.1. However, the explicit theoretical derivation is technically involved and has been avoided in this thesis. By Lemma 2, with dominating probability

$$\int (L(\theta)/L(\theta_0))p(\theta)d\theta \geq \exp(-Cn\epsilon_n^2 \log n)$$

Note, that

$$\Pi(\mathcal{U}^c_{r\varepsilon_n}|X^{(n)}) = \frac{\int_{\mathcal{U}^c_{r\varepsilon_n}} L(\theta)p(\theta)d\theta}{\int L(\theta)p(\theta)d\theta} = \frac{\int_{\mathcal{U}^c_{r\varepsilon_n}} (L(\theta)/L(\theta_0))p(\theta)d\theta}{\int (L(\theta)/L(\theta_0))p(\theta)d\theta}$$

Therefore, with dominating probability

$$\Pi(\mathcal{U}^c_{r\varepsilon_n}|X^{(n)}) \leq \exp(-C_2 rn\varepsilon_n^2)\exp(Cn\epsilon_n^2 \log n)$$

Following steps of proof of proposition 11 on page 2,111 in Van Der Vaart and Van Zanten (2011),

$$\int e^{(C_2/2)n\varepsilon_n||\theta-\theta_0||_2} d\Pi(\mathcal{U}^c_{\varepsilon_n}|X^{(n)}) = \int_0^\infty e^{(C_2/2)n\varepsilon_n^2 r}\Pi(||\theta - \theta_0||_2 \geq r\varepsilon_n|X^{(n)})dr$$

Therefore,

$$\begin{aligned}
&\int e^{(C_2/2)n\varepsilon_n||\theta-\theta_0||_2} d\Pi(\mathcal{U}_{\varepsilon_n}^c|X^{(n)}) \\
&= \exp(Cn\epsilon_n^2 \log n) \int_0^\infty \exp((C_2/2)rn\varepsilon_n^2) \exp(-C_2 rn\varepsilon_n^2) dr \\
&= \exp(Cn\epsilon_n^2 \log n) \int_0^\infty \exp(-(C_2/2)rn\varepsilon_n^2) dr = \frac{2}{C_2 n\varepsilon_n^2} \exp(Cn\epsilon_n^2 \log n)
\end{aligned}$$

This completes the proof.

2.14.1 Proof of Relation (2.53)

Lemma 15. *With $\mathcal{U}_{r\epsilon_n}$ same as in Lemma 1, we have for some $C_2 > 0$,*

$$\mathbb{P}_0^{(n)} \left(\sup_{\theta \in \mathcal{U}_{r\epsilon_n}^c} \int \frac{L(\theta)}{L(\theta_0)} p(\theta) d\theta \leq -C_2 rn\epsilon_n^2 \right) \to 1, n \to \infty \tag{2.54}$$

Proof. Following the proof of Lemma 1, we relate $\mathcal{U}_{r\epsilon_n}$ to $\mathcal{V}_{r\epsilon_n} = \{|\beta - \beta_0| < r\epsilon_n, |B - B_0| < r\epsilon_n\}$. Note, $\mathcal{V}_{r\epsilon_n}$ can be split into sets $V_{1n}, \cdots, V_{8n}$ as in the proof of Lemma 1. We study the behavior of

$$\sup_{\theta \in V_{1n}} (\log L(\theta) - \log L(\theta_0))$$

where $V_{1n} = \{\beta \geq \beta_0 + r\epsilon_n, B \geq B_0\}$. The proof for other cases of $V_{2n}, \cdots, V_{8n}$ follows similarly. The proof of (2.54) will then follow as a consequence of (2.30).

By a simple modification of the steps for the relation (2.31) used in the proof of Lemma 1, it can be shown that

$$\sup_{\theta \in V_{1n}} (\log L(\theta) - \log L(\theta_0)) \leq O(\log L((\beta_0 + r\epsilon_n, B_0)) - \log L(\beta_0, B_0))$$

Next, we numerically demonstrate that

$$\underbrace{\log L((\beta_0 + r\epsilon_n, B_0)) - \log L(\beta_0, B_0)}_{\text{LHS}} \leq \underbrace{-Crn\epsilon_n^2}_{\text{RHS}}, \ n \to \infty$$

where $C = 0.01$. Note, for computing LHS in the above relation, $\boldsymbol{x}$ is generated using relation (1.1) with $\beta = \beta_0$ and $B = B_0$. For varying values of $r \in (0, 1000)$, we consider two ratios

$$\rho_1 = \text{Proportion of } r \text{ values where LHS} > \text{RHS}$$

$$\rho_2 = \text{Proportion of } r \text{ values where LHS} > \text{RHS provided LHS} < 0$$

The reason for considering the two proportion ρ_1 and ρ_2 is because RHS is a negative quantity and the upper bound is justified only when the LHS is als a negative quantity.

The tables below give the values of ρ_1 and ρ_2 for varying values of n, d and ϵ_n. It is evident that both ρ_1 and ρ_2 approach zero as sample size increases. The fact that ρ_1 approaches 0 as $n \to \infty$ is also immediate from the proof of Lemma 1 which shows that LHS becomes negative as $n \to \infty$,see relation (2.32).

			ρ_1 for $d = 5$ (ρ_2 for $d = 5$)		
ϵ_n	β	B	$n = 200$	$n = 1000$	$n = 5000$
$n^{-0.4}$	0.2	0.2	0.0(0.0)	0.0(0.0)	0.0(0.0)
		-0.2	0.002(0.002)	0.0(0.0)	0.0(0.0)
		0.5	0.0(0.0)	0.0(0.0)	0.0(0.0)
		-0.5	0.092(0.0)	0.0(0.0)	0.092(0.0)
	0.5	0.2	0.092(0.0)	0.0(0.0)	0.0(0.0)
		-0.2	0.0(0.0)	0.0(0.0)	0.0(0.0)
		0.5	0.0(0.0)	0.0(0.0)	0.0(0.0)
		-0.5	0.0(0.0)	0.0(0.0)	0.0(0.0)
$n^{-0.2}$	0.2	0.2	0.091(0.0)	0.040(0.009)	0.0(0.0)
		-0.2	0.0(0.0)	0.014(0.008)	0.0(0.0)
		0.5	0.0(0.0)	0.056(0.008)	0.0(0.0)
		-0.5	0.0(0.0)	0.0(0.0)	0.0(0.0)
	0.5	0.2	0.060(0.009)	0.0(0.0)	0.0(0.0)
		-0.2	0.036(0.009)	0.0(0.0	0.0(0.0)
		0.5	0.066(0.009)	0.018(0.006)	0.014(0.007)
		-0.5	0.091(0.010)	0.021(0.007)	0.016(0.007)

□

			ρ_1 for $d = 30$ (ρ_2 for $d = 30$)		
ϵ_n	β	B	$n = 200$	$n = 1000$	$n = 5000$
$n^{-0.4}$	0.2	0.2	0.0(0.0)	0.062(0.0)	0.0(0.0)
		-0.2	0.026(0.017)	0.092(0.0)	0.0(0.0)
		0.5	0.056(0.008)	0.023(0.009)	0.0(0.0)
		-0.5	0.0(0.0)	0.025(0.008)	0.0(0.0)
	0.5	0.2	0.091(0.0)	0.081(0.014)	0.091(0.0)
		-0.2	0.067(0.014)	0.0(0.0)	0.093(0.0)
		0.5	0.0(0.0)	0.0(0.0)	0.0(0.0)
		-0.5	0.0(0.0)	0.0(0.0)	0.0(0.0)
$n^{-0.2}$	0.2	0.2	0.028(0.028)	0.0(0.0)	0.0(0.0)
		-0.2	0.091(0.0)	0.023(0.023)	0.0(0.0)
		0.5	0.0(0.0)	0.021(0.008)	0.0(0.0)
		-0.5	0.0(0.0)	0.0(0.0)	0.0(0.0)
	0.5	0.2	0.039(0.015)	0.0(0.0)	0.004(0.004)
		-0.2	0.044(0.013)	0.042(0.012)	0.0(0.0)
		0.5	0.0(0.0)	0.046(0.009)	0.0(0.0)
		-0.5	0.015(0.007)	0.0(0.0)	0.0(0.0)

CHAPTER 3

BAYESIAN VARIABLE SELECTION IN A STRUCTURED REGRESSION MODEL

In the history of statistics, estimation of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ in a regression model, $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$, has been one of the most popular research topics, where, $X \in \mathbb{R}^{n \times p}$ is a given design matrix, $\boldsymbol{y} \in \mathbb{R}^n$ is a response vector, and $\boldsymbol{e} \sim MVN(0, I)$ is an error term. We consider a high-dimensional problem $n < p$, with sparse regression coefficients in which some β_i's are non-zero and others are exactly zero. In this sparse high-dimensional setting, many previous literatures have considered Bayesian approaches to variable selection (Ray and Szabó, 2021; Ročková and George, 2018; Martin et al., 2017; Ročková and George, 2014; Carbonetto and Stephens, 2012). In addition to the sparsity on $\boldsymbol{\beta}$, we consider a structurally dependent feature space. Dependent feature vector commonly occurs in genetics, neuroimaging, and image analysis. For example, suppose that an image is given in which each pixel has a value. There are signal areas in the image such that the pixels in the areas have non-zero values and the others pixels are all zeros. The shape of the signal areas varies such as rectangles, circles, or letters (See Figure 3.1). The image can be represented by a matrix and corresponding $\boldsymbol{\beta}$ is the vectorization of the matrix. In such structured regression models,

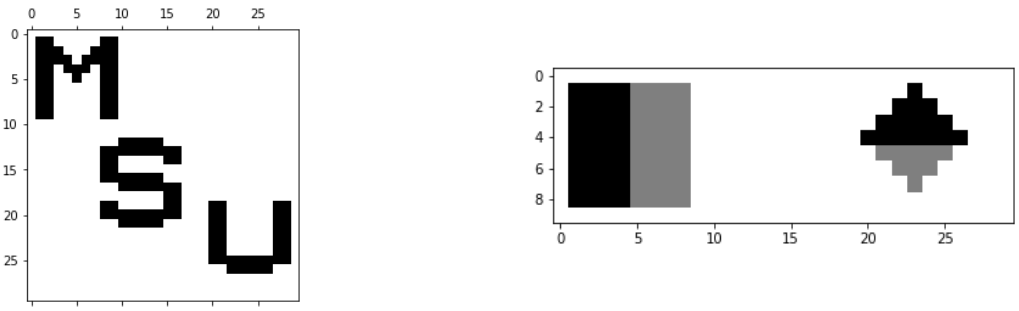

Figure 3.1: Example of the structured regression coefficients. White pixels denote the corresponding β_i's are zeros and darker pixels denote the corresponding β_i's have larger values.

problems may arise with the classic variable selection methods. In order to use the structural information and select signal features, Li and Zhang (2010) employed latent binary variables

$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^\top \in \{0,1\}^p$ with Ising prior on it and a spike and slab prior on $\boldsymbol{\beta}$ for approximated the posterior selection probabilities by adapting Gibbs sampling which involves a matrix inversion. Due to the computational cost of the matrix inversion, Li and Zhang (2010)'s method are not scalable when p or the degree of the underlying graph is large. Chang et al. (2018) proposed a Bayesian shrinkage approach via EM algorithm which is scalable to high dimensional settings. We suggest a variational Bayes method (Jordan et al., 1999) to simultaneously select signal coefficients and estimate the magnitudes of the signals where a predetermined coupling matrix is considered to incorporate the network structure of features. The coupling matrix in our method does not necessarily represent the true connection between features, which allows us to utilize a common structure of a coupling matrix such as *k-nearest* neighbor in image analysis.

3.1 Model and methodology

Throughout this chapter, we assume that p_n is the number of covariates which depends on the number of observations n. In a linear regression model, let $\boldsymbol{\beta} \in \mathbb{R}^{p_n}$ and $\sigma^2 > 0$ denote the regression coefficients and the residual variance respectively. Since we consider high-dimensional setup, we assume that $p_n > n$ and $p_n \to \infty$ as $n \to \infty$. Then, the linear regression model is written as:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}, \tag{3.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is an observed response vector, $X \in \mathbb{R}^{n \times p_n}$ is a given design matrix, and $\boldsymbol{e} = (e_1, \ldots, e_n)^\top \sim MVN\,(0,\ \sigma^2 I_n)$. Since estimating σ^2 is not of our interest, with out loss of generality, we assume $\sigma^2 = 1$. Our Bayesian variable selection method is easily extendable to the case of unknown σ^2 with an appropriate prior distribution. Now, there are p_n unknown parameters in the regression model. In addition to the high-dimensional regression setup, we further assume structured covariates for which it is desirable to capture the intrinsic dependence among covariates. Introducing binary latent variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p_n})^\top \in \{-1,1\}^{p_n}$, we perform Bayesian variable selection using a

spike and slab method as follows:

$$
\begin{aligned}
p(\boldsymbol{\gamma}) &= \frac{1}{2^{p_n}} \exp\left(\sum_{i=1}^{p_n} \left(b_0 \gamma_i m_i(\boldsymbol{\gamma}) + a_0 \gamma_i - \log\cosh(b_0 m_i(\boldsymbol{\gamma}) + a_0) \right) \right), \\
p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) &= \prod_{i=1}^{p_n} p(\beta_i \mid \gamma_i), \\
p(\beta_i \mid \gamma_i) &\sim \left(\frac{1-\gamma_i}{2} \right) \mathbb{1}(\beta_i = 0) + \left(\frac{1+\gamma_i}{2} \right) \mathcal{N}\left(0, \tau^2\right),
\end{aligned}
\tag{3.2}
$$

where $m(\boldsymbol{\gamma}) := (m_1(\boldsymbol{\gamma}), \ldots, m_{p_n}(\boldsymbol{\gamma}))^\top = J_n \boldsymbol{\gamma}$, $J_n \in \mathbb{R}^{p_n \times p_n}$ is a coupling matrix, and $\mathcal{N}(0, \tau^2)$ denotes a Gaussian distribution with probability density function as follows:

$$
p(\beta_i \mid \gamma_i = 1) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{\beta_i^2}{2\tau^2} \right).
$$

One can notice that $p(\boldsymbol{\gamma})$ is a pseudo-likelihood of Ising model. To approximate the unknown posterior distributions, we define a variational family as follows:

$$
\begin{aligned}
q(\boldsymbol{\gamma}) &= \frac{1}{2^{p_n}} \exp\left(\sum_{i=1}^{p_n} \left(b \gamma_i m_i(\boldsymbol{\gamma}) + a_i \gamma_i - \log\cosh(b m_i(\boldsymbol{\gamma}) + a_i) \right) \right), \\
q(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) &= \prod_{i=1}^{p_n} q(\beta_i \mid \gamma_i), \\
q(\beta_i \mid \gamma_i) &\sim \left(\frac{1-\gamma_i}{2} \right) \mathbb{1}(\beta_i = 0) + \left(\frac{1+\gamma_i}{2} \right) \mathcal{N}\left(\mu_i, \sigma_i^2\right).
\end{aligned}
\tag{3.3}
$$

Note that we have $3p_n + 1$ variational parameters needed to be updated. We point out that using the multiple threshold parameters $\boldsymbol{a} = (a_1, \ldots, a_{p_n})^\top$, the posterior inclusion probability, that is, the probability that $\gamma_i = 1$ given $\boldsymbol{y}$, can be different over i, which is desirable for variable selection.

3.1.1 ELBO optimization

Let $\boldsymbol{\nu} = (\boldsymbol{a}^\top, b, \boldsymbol{\mu}^\top, \boldsymbol{\sigma}^\top)$ denote the set of variational parameters, where $\boldsymbol{a} = (a_1, \ldots, a_{p_n})^\top$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{p_n})^\top$, and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{p_n})^\top$. Then, the negative ELBO is:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\nu}) &= \mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[\log q(\boldsymbol{\beta},\boldsymbol{\gamma}) - \log p(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{y})\right] \\ &= \underbrace{\mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[-\log p\left(\boldsymbol{y} \mid \boldsymbol{\beta}\right)\right]}_{①} + \underbrace{\mathbb{E}_{q(\boldsymbol{\gamma})}\left[\log q(\boldsymbol{\gamma}) - \log p(\boldsymbol{\gamma})\right]}_{②} \\ &\quad + \underbrace{\mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[\log q(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) - \log p(\boldsymbol{\beta} \mid \boldsymbol{\gamma})\right]}_{③}.\end{aligned} \tag{3.4}$$

The first term is:

$$\begin{aligned}① &= \mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[\frac{1}{2}\left(\boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta} - 2\boldsymbol{y}^\top X\boldsymbol{\beta} + \boldsymbol{y}^\top\boldsymbol{y}\right)\right] + C \\ &= \mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[\frac{1}{2}\left(\sum_{k=1}^{n}\left(\sum_{i=1}^{p_n} X_{ki}\beta_i\right)^2 - 2\sum_{i=1}^{p_n} X_{\cdot i}^\top\boldsymbol{y}\beta_i\right) + \frac{1}{2}\left(\boldsymbol{y}^\top\boldsymbol{y}\right)\right] + C \\ &= \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[\sum_{k=1}^{n}\sum_{i=1}^{p_n} X_{ki}^2\beta_i^2 - 2\sum_{i=1}^{p_n} X_{\cdot i}^\top\boldsymbol{y}\beta_i + \sum_{k=1}^{n}\sum_{i\neq l} X_{ki}X_{kl}\beta_i\beta_l\right] + C,\end{aligned}$$

where $X_{\cdot i}$ is i-th column of X. We compute the closed form of the above expectation:

$$\begin{aligned}&\mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[\sum_{k=1}^{n}\sum_{i=1}^{p_n} X_{ki}^2\beta_i^2 - 2\sum_{i=1}^{p_n} X_{\cdot i}^\top\boldsymbol{y}\beta_i + \sum_{k=1}^{n}\sum_{i\neq l} X_{ki}X_{kl}\beta_i\beta_l\right] \\ &= \sum_{k=1}^{n}\sum_{i=1}^{p_n} X_{ki}^2\phi_i\mathbb{E}_{\beta_i\sim\mathcal{N}(\mu_i,\sigma_i^2)}[\beta_i^2] - 2\sum_{i=1}^{p_n} X_{\cdot i}^\top\boldsymbol{y}\phi_i\mathbb{E}_{\beta_i\sim\mathcal{N}(\mu_i,\sigma_i^2)}[\beta_i] \\ &\quad + \sum_{k=1}^{n}\sum_{i\neq l} X_{ki}X_{kl}\phi_i\phi_l\mathbb{E}_{\beta_i\sim\mathcal{N}(\mu_i,\sigma_i^2)}[\beta_i]\mathbb{E}_{\beta_l\sim\mathcal{N}(\mu_l,\sigma_l^2)}[\beta_l] \\ &= \sum_{k=1}^{n}\sum_{i=1}^{p_n} X_{ki}^2\phi_i(\mu_i^2 + \sigma_i^2) - 2\sum_{i=1}^{p_n} X_{\cdot i}^\top\boldsymbol{y}\phi_i\mu_i + \sum_{k=1}^{n}\sum_{i\neq l} X_{ki}X_{kl}\phi_i\phi_l\mu_i\mu_l,\end{aligned}$$

where $\phi_i = e^{a_i+bm_i(\boldsymbol{\gamma})}/\left(e^{a_i+bm_i(\boldsymbol{\gamma})} + e^{-a_i-bm_i(\boldsymbol{\gamma})}\right)$ is the marginal probability $q(\gamma_i = 1)$. Similarly, let $\phi_{i0} = e^{a_0+b_0m_i(\boldsymbol{\gamma})}/\left(e^{a_0+b_0m_i(\boldsymbol{\gamma})} + e^{-a_0-b_0m_i(\boldsymbol{\gamma})}\right)$ denote the marginal probability induced from the prior $p(\boldsymbol{\gamma})$. The closed form of the second term in the negative ELBO (3.4)

is:

$$
\begin{aligned}
\textcircled{2} &= \mathbb{E}_{q(\boldsymbol{\gamma})}\left[\sum_{i=1}^{p_n}\left(\log\phi_i - \log\phi_{0i}\right)\right] \\
&= \sum_{i=1}^{p}\phi_i\left(\log\phi_i - \log\phi_{0i}\right) + (1-\phi_i)\left(\log(1-\phi_i) - \log(1-\phi_{0i})\right).
\end{aligned}
$$

The last one is:

$$
\begin{aligned}
\textcircled{3} &= \mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[\sum_{i=1}^{p_n}\left(\log q(\beta_i \mid \gamma_i) - \log p(\beta_i \mid \gamma_i)\right)\right] \\
&= \sum_{i=1}^{p_n}\phi_i KL\left(\mathcal{N}(\mu_i,\sigma_i^2),\mathcal{N}(0,\tau^2)\right) \\
&= \sum_{i=1}^{p_n}\frac{\phi_i}{2}\left(\frac{\mu_i^2+\sigma_i^2}{\tau^2} - 1 + 2\log\tau - 2\log\sigma_i\right).
\end{aligned}
$$

The last result of $-(\textcircled{1}+\textcircled{2}+\textcircled{3})$ provides the closed form of the ELBO. Let $\boldsymbol{\phi} := (\phi_1,\ldots,\phi_{p_n})^\top$ and let $\boldsymbol{\phi}^{0.5} := (\sqrt{\phi_1},\ldots,\sqrt{\phi_{p_n}})^\top$. Then, in a matrix notation, we write:

$$
\begin{aligned}
2\cdot\textcircled{1} &= \sum_{k=1}^{n}\sum_{i=1}^{p_n}X_{ki}^2\phi_i(\mu_i^2+\sigma_i^2) - 2\sum_{i=1}^{p}X_{\cdot i}^\top\boldsymbol{y}\phi_i\mu_i + \sum_{k=1}^{n}\sum_{i\neq l}X_{ki}X_{kl}\phi_i\phi_l\mu_i\mu_l + C \\
&= \left(\boldsymbol{\phi}^{0.5}\circ\boldsymbol{\mu}\right)^\top\left(X^\top X\circ I_{p_n}\right)\left(\boldsymbol{\phi}^{0.5}\circ\boldsymbol{\mu}\right) + \left(\boldsymbol{\phi}^{0.5}\circ\boldsymbol{\sigma}\right)^\top\left(X^\top X\circ I_{p_n}\right)\left(\boldsymbol{\phi}^{0.5}\circ\boldsymbol{\sigma}\right) \\
&\quad - 2\boldsymbol{y}^\top X(\boldsymbol{\phi}\circ\boldsymbol{\mu}) + (\boldsymbol{\phi}\circ\boldsymbol{\mu})^\top\left(X^\top X\circ(\mathbf{1}_{p_n\times p_n} - I_{p_n})\right)(\boldsymbol{\phi}\circ\boldsymbol{\mu}) + C, \qquad (3.5)
\end{aligned}
$$

where $\circ$ denote the element-wise product and $\mathbf{1}_{p_n\times p_n}\in\mathbb{R}^{p_n\times p_n}$ is a matrix whose elements are all ones. Define another vector $\boldsymbol{\phi}_0 := (\phi_{01},\ldots,\phi_{0p_n})^\top$ and let $\log(\cdot)$ mean element-wise logarithm when the input is a vector of a matrix. Then,

$$
\begin{aligned}
\textcircled{2} &= \sum_{i=1}^{p}\phi_i\left(\log\phi_i - \log\phi_{0i}\right) + (1-\phi_i)\left(\log(1-\phi_i) - \log(1-\phi_{0i})\right) \\
&= \boldsymbol{\phi}^\top\left(\log\boldsymbol{\phi} - \log\boldsymbol{\phi}_0\right) + (\mathbf{1}-\boldsymbol{\phi})^\top\left(\log(\mathbf{1}-\boldsymbol{\phi}) - \log(\mathbf{1}-\boldsymbol{\phi}_0)\right),
\end{aligned}
$$

and

$$
\begin{aligned}
\textcircled{3} &= \sum_{i=1}^{p_n}\frac{\phi_i}{2}\left(\frac{\mu_i^2+\sigma_i^2}{\tau^2} - 1 + 2\log\tau - 2\log\sigma_i\right) \\
&= 0.5\boldsymbol{\phi}^\top\left(\tau^{-2}\left(\boldsymbol{\mu}\circ\boldsymbol{\mu}+\boldsymbol{\sigma}\circ\boldsymbol{\sigma}\right) + (2\log\tau - 1)\cdot\mathbf{1} - 2\log\boldsymbol{\sigma}\right).
\end{aligned}
$$

With the closed form of the ELBO, the gradients of the ELBO with respect to variational parameters are also easily computable in closed forms. First, we compute the gradients $\nabla_{a_i}\phi_i$ and $\nabla_b\phi_i$:

$$\nabla_{a_i}\phi_i = \frac{2e^{2(a_i+bm_i(\boldsymbol{\gamma}))}}{\left(e^{2(a_i+bm_i(\boldsymbol{\gamma}))}+1\right)^2},$$
$$\nabla_b\phi_i = \frac{2m_i(\boldsymbol{\gamma})e^{2(a_i+bm_i(\boldsymbol{\gamma}))}}{\left(e^{2(a_i+bm_i(\boldsymbol{\gamma}))}+1\right)^2}.$$

Observe that $\nabla_{a_i}\phi_j = 0$ if $i \neq j$. We define $\nabla_{\boldsymbol{a}}\boldsymbol{\phi} := \left(\nabla_{a_1}\phi_1, \ldots, \nabla_{a_{p_n}}\phi_{p_n}\right)^\top$ and $\nabla_b\boldsymbol{\phi} := \left(\nabla_b\phi_1, \ldots, \nabla_b\phi_{p_n}\right)^\top$. Then, by chain rule, we can compute the closed forms of all the gradients needed.

Gradients with respect to $\boldsymbol{a}$:

$$\begin{aligned}2\cdot\nabla_{\boldsymbol{a}}\textcircled{1} = &\left(X^\top X\circ I_p\right)\left(\boldsymbol{\mu}\circ\boldsymbol{\mu}\circ\nabla_{\boldsymbol{a}}\boldsymbol{\phi}\right)+\left(X^\top X\circ I_p\right)\left(\boldsymbol{\sigma}\circ\boldsymbol{\sigma}\circ\nabla_{\boldsymbol{a}}\boldsymbol{\phi}\right)\\ &-2\left(X^\top\boldsymbol{y}\right)\circ\boldsymbol{\mu}\circ\nabla_{\boldsymbol{a}}\boldsymbol{\phi}+2\left(\left(X^\top X\circ\left(\mathbf{1}_{p_n\times p_n}-I_{p_n}\right)\right)\left(\boldsymbol{\phi}\circ\boldsymbol{\mu}\right)\right)\circ\left(\boldsymbol{\mu}\circ\nabla_{\boldsymbol{a}}\boldsymbol{\phi}\right),\end{aligned}$$

$$\nabla_{\boldsymbol{a}}\textcircled{2} = \nabla_{\boldsymbol{a}}\boldsymbol{\phi}\circ\left(\log\boldsymbol{\phi}-\log\boldsymbol{\phi}_0\right)-\nabla_{\boldsymbol{a}}\boldsymbol{\phi}\left(\log\left(\mathbf{1}-\boldsymbol{\phi}\right)-\log\left(\mathbf{1}-\boldsymbol{\phi}_0\right)\right),$$

$$\nabla_{\boldsymbol{a}}\textcircled{3} = 0.5\nabla_{\boldsymbol{a}}\boldsymbol{\phi}\circ\left(\tau^{-1}\left(\boldsymbol{\mu}\circ\boldsymbol{\mu}+\boldsymbol{\sigma}\circ\boldsymbol{\sigma}\right)+\left(2\log\tau-1\right)\cdot\mathbf{1}-2\log\boldsymbol{\sigma}\right).$$

Gradients with respect to b:

$$\begin{aligned}&0.5\nabla_b\textcircled{1}\\ &=\left(\nabla_b\boldsymbol{\phi}^{0.5}\circ\boldsymbol{\mu}\right)^\top\left(X^\top X\circ I_{p_n}\right)\left(\nabla_b\boldsymbol{\phi}^{0.5}\circ\boldsymbol{\mu}\right)+\left(\nabla_b\boldsymbol{\phi}^{0.5}\circ\boldsymbol{\sigma}\right)^\top\left(X^\top X\circ I_{p_n}\right)\left(\nabla_b\boldsymbol{\phi}^{0.5}\circ\boldsymbol{\sigma}\right)\\ &-2\boldsymbol{y}^\top X(\nabla_b\boldsymbol{\phi}\circ\boldsymbol{\mu})+\left(\nabla_b\boldsymbol{\phi}\circ\boldsymbol{\mu}\right)^\top\left(X^\top X\circ\left(\mathbf{1}_{p_n\times p_n}-I_{p_n}\right)\right)\left(\boldsymbol{\phi}\circ\boldsymbol{\mu}\right)\\ &+\left(\boldsymbol{\phi}\circ\boldsymbol{\mu}\right)^\top\left(X^\top X\circ\left(\mathbf{1}_{p_n\times p_n}-I_{p_n}\right)\right)\left(\nabla_b\boldsymbol{\phi}\circ\boldsymbol{\mu}\right),\end{aligned}$$

$$\nabla_b\textcircled{2} = \nabla_b\boldsymbol{\phi}^\top\left(\log\boldsymbol{\phi}-\log\boldsymbol{\phi}_0\right)-\nabla_b\boldsymbol{\phi}^\top\left(\log\left(\mathbf{1}-\boldsymbol{\phi}\right)-\log\left(\mathbf{1}-\boldsymbol{\phi}_0\right)\right),$$

$$\nabla_b\textcircled{3} = 0.5\nabla_b\boldsymbol{\phi}^\top\left(\tau^{-1}\left(\boldsymbol{\mu}\circ\boldsymbol{\mu}+\boldsymbol{\sigma}\circ\boldsymbol{\sigma}\right)+\left(2\log\tau-1\right)\cdot\mathbf{1}-2\log\boldsymbol{\sigma}\right).$$

Gradients with respect to $\boldsymbol{\mu}$:

$$0.5\nabla_{\boldsymbol{\mu}}\textcircled{1} = \left(X^\top X \circ I_p\right)(\boldsymbol{\mu} \circ \boldsymbol{\phi}) - 2(X^\top \boldsymbol{y}) \circ \boldsymbol{\phi} + 2\left(\left(X^\top X \circ (\mathbf{1}_{p_n \times p_n} - I_{p_n})\right)(\boldsymbol{\phi} \circ \boldsymbol{\mu})\right) \circ \boldsymbol{\phi},$$

$$\nabla_{\boldsymbol{\mu}}\textcircled{3} = \tau^{-2}\boldsymbol{\phi} \circ \boldsymbol{\mu}.$$

Gradients with respect to $\boldsymbol{\sigma}$:

$$0.5\nabla_{\boldsymbol{\sigma}}\textcircled{1} = \left(X^\top X \circ I_p\right)(\boldsymbol{\sigma} \circ \boldsymbol{\phi}),$$

$$\nabla_{\boldsymbol{\mu}}\textcircled{3} = \tau^{-2}\boldsymbol{\phi} \circ \boldsymbol{\mu}.$$

3.2 Implementation details

Starting with initial variational parameters $\boldsymbol{\nu}^{(0)}$, the first step for implementing the VB algorithm is to draw a $\boldsymbol{\gamma}$ from the current $q(\boldsymbol{\gamma}; b^{(t)}, \boldsymbol{a}^{(t)})$ using Gibbs sampling. Note that we do not need to draw samples from $q(\boldsymbol{\beta} \mid \boldsymbol{\gamma})$. Given a $\boldsymbol{\gamma}$, we can obtain the closed forms of ELBO and all the gradients $\widehat{\nabla_\nu \mathcal{L}(\boldsymbol{\nu})}$ described in the subsection 3.1.1. Then, we can update the current variational parameters as follows:

$$\nu^{(t+1)} \leftarrow \nu^{(t)} + \eta_t(\nu) \cdot \widehat{\nabla_\nu \mathcal{L}(\boldsymbol{\nu})}.$$

Note that, we allow that the learning rates $\eta_t(\nu)$ depend on variational parameters $\nu \in \boldsymbol{\nu}$. Since we have $3p_n + 1$ variational parameters, a single learning rate does not guarantee the convergence of all the variational parameters. Therefore, we use adaptive learning rates such as Adam described in Algorithm 1.1. After the optimal variational parameters $\boldsymbol{\nu}^*$ are achieved, we can empirically compute the marginal probability that $\gamma_i = 1$, $i = 1, \ldots, p_n$ based on the samples:

$$\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_S \sim q(\boldsymbol{\gamma}; b^*, \boldsymbol{a}^*),$$

where $\boldsymbol{\gamma}_s = (\gamma_{s,1}, \ldots, \gamma_{s,p_n})^\top$, $s = 1 \ldots, S$. Then the marginal probability of γ_i is:

$$q^*(\gamma_i = 1) = \frac{\sum_{s=1}^{S} \mathbb{1}(\gamma_{s,i} = 1)}{S}.$$

Based on the empirical marginal probabilities, we select the i-th feature if:

$$q^*(\gamma_i = 1) > T,$$

where T is a threshold (it could be a fixed number such as 0.5, or it could be the M-th largest marginal). For estimation of $\boldsymbol{\beta}$, we use the (variational) posterior mode $\boldsymbol{\mu}^*$.

3.3 Numerical results

In this section, we numerically investigate our variable selection algorithm equipped with two different adaptive learning rates, Adam (Kingma and Ba, 2014) and RMSprop which is an unpublished algorithm first proposed in the Coursera course. "Neural Network for Machine Learning" by Geoff Hinton. Also, we compare our method with a MCMC based method (Li and Zhang, 2010). Li and Zhang (2010) approached the structural regression problem through the Bayesian variable selection framework, where the covariates lie on an undirected graph and formulate an Ising prior on the model space for incorporating structural information. Li and Zhang (2010) adopt the Gibbs sampling algorithms, first suggested by George and McCulloch (1993).

3.3.1 Li and Zhang (2010)'s Gibbs sampling scheme

For the formulation of Li and Zhang (2010)'s Gibbs sampling method, we define $\boldsymbol{\gamma}_{(-i)} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_p)^\top$; let $I_{(-i)}$ be the set of indices $\{\gamma_j = 1 : j \neq i\}$; $I_{(i)} = I_{(-i)} \cup \{i\}$; $p_{(i)} = |I_{(i)}|$ and $p_{(-i)} = |I_{(-i)}|$. With an Ising prior on $\boldsymbol{\gamma} \in \{0,1\}^p$,

$$p(\boldsymbol{\gamma}) = \frac{1}{Z(a_0, b_0)} \exp\left(\frac{b_0}{2}\boldsymbol{\gamma}^\top J \boldsymbol{\gamma} + a_0 \sum_{i=1}^{p} \gamma_i\right),$$

the posterior distribution of $\boldsymbol{\gamma}$ given the data can be decomposed by Bayes formula as follows:

$$\mathbb{P}\left(\gamma_i = 1 \mid \boldsymbol{\gamma}_{(-i)}, \boldsymbol{y}\right) = \frac{\mathbb{P}\left(\gamma_i = 1 \mid \boldsymbol{\gamma}_{(-i)}\right)}{\mathbb{P}\left(\gamma_i = 1 \mid \boldsymbol{\gamma}_{(-i)}\right) + BF\left(i \mid \boldsymbol{\gamma}_{(-i)}\right)^{-1} \cdot \mathbb{P}\left(\gamma_i = -1 \mid \boldsymbol{\gamma}_{(-i)}\right)}, \tag{3.6}$$

where $BF\left(i \mid \boldsymbol{\gamma}_{(-i)}\right) = \frac{\mathbb{P}\left(\boldsymbol{y}|\gamma_i=1,\boldsymbol{\gamma}_{(-i)}\right)}{\mathbb{P}\left(\boldsymbol{y}|\gamma_i=0,\boldsymbol{\gamma}_{(-i)}\right)}$ is the Bayes factor, and

$$\mathbb{P}\left(\gamma_i \mid \boldsymbol{\gamma}_{(-i)}\right) = \frac{e^{\gamma_i\left(a_0+b_0\sum_{j\in I_{(-i)}}\gamma_j\right)}}{1+e^{a_0+b_0\sum_{j\in I_{(-i)}}\gamma_j}}.$$

The Bayes factor can be explicitly computed under the spike and slab prior on $\boldsymbol{\beta}$:

$$BF\left(i \mid \boldsymbol{\gamma}_{(-i)}\right) = \tau^{-1}\left(\frac{|K_{(-i)}|}{|K_{(i)}|}\right)^{1/2}\left(\frac{\boldsymbol{y}^\top\boldsymbol{y}-\boldsymbol{y}^\top X_{I_{(-i)}}K_{(-i)}^{-1}X_{I_{(-i)}}^\top\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}-\boldsymbol{y}^\top X_{I_{(i)}}K_{(i)}^{-1}X_{I_{(i)}}^\top\boldsymbol{y}}\right)^{n/2},$$

where $K_{(i)} = X_{I_{(i)}}^\top X_{I_{(i)}} + \tau^{-2}I_{p_{(i)}}$ and $K_{(-i)} = X_{I_{(-i)}}^\top X_{I_{(-i)}} + \tau^{-2}I_{p_{(-i)}}$. Based on Gibbs samples from the posterior probabilities in (3.6), Li and Zhang (2010) calculated the marginal posterior probabilities to find signal variables.

3.3.2 Hyper parameter selection

To implement the algorithm with the prior distributions (3.2), three hyper parameters are needed to be selected. We follow Li and Zhang (2010) for the hyper parameter choices. Li and Zhang (2010) considered reparametrizations for the hyper parameters (a_0, b_0) as follows:

$$a_0 = \log(r/w_0^2) \text{ and } b_0 = \log(w_1 \cdot w_0),$$

where $w_0 = rw_1 + 1 - r$. Li and Zhang (2010) used fixed $r = 0.03$ and various w_1. Table 3.1 shows examples of hyper parameter choices. Note that $w_1 = 1$ corresponds to the independent Bernoulli prior.

w_1	a_0	b_0
1	-3.5066	0
5	-3.7332	1.7228
9	-3.9368	2.4123

Table 3.1: Examples of hyper parameter choices

3.3.3 ROC curve

In this subsection, we compute ROC curves to see how performance changes as hyper parameters change in two scenarios.

Scenario 1: For the first scenario, we assume the γ_is are arranged on a circle clockwise. Then, each γ_i has two neighbors, γ_{i-1} and γ_{i+1}, where $\gamma_0 = \gamma_p$ and $\gamma_{p+1} = \gamma_1$, where $n = 100$ and $p = 2000$. Figure 3.2 shows the connectivity among $\boldsymbol{\gamma}$. For the true regression coefficients, we use:

$$\beta_i = \begin{cases} 0.3, \text{ if } i \in \mathcal{A} \text{ and } i \text{ is odd}, \\ 0.6, \text{ if } i \in \mathcal{A} \text{ and } i \text{ is even}, \\ 0, \text{ otherwise}, \end{cases}$$

where $\mathcal{A} = \{i : i \in [245, 260] \cup [745, 760] \cup [1245, 1260] \cup [1745, 1760]\}$.

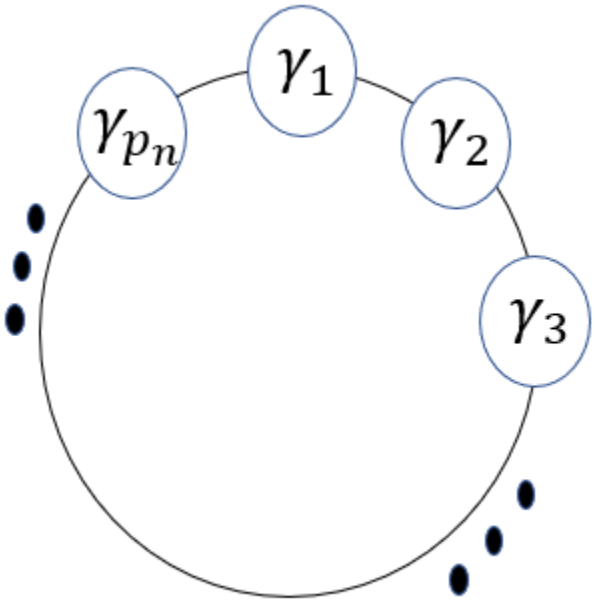

Figure 3.2: $\boldsymbol{\gamma}$ on a circle

For design matrix $X \in \mathbb{R}^{n \times p}$, we first assume each component follows independent standard Gaussian distribution, i.e., $X_{ij} \sim \mathcal{N}(0, 1)$. Figure 3.3 demonstrates the ROC curves corresponding to the hyper parameters $w_1 = 1$ and $w_1 = 3$. Figure 3.4 contains the results for higher w_1 with independent X.

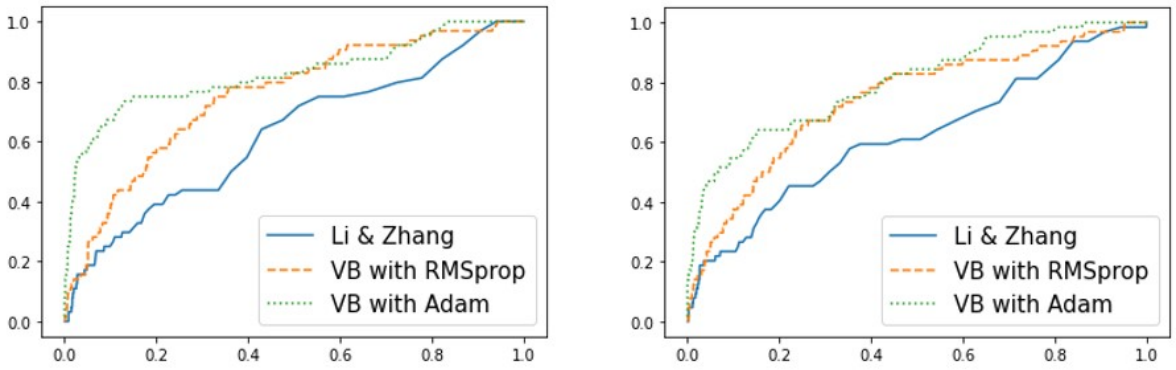

Figure 3.3: ROC curves for the three variable selection methods with the hyper parameter $w_1 = 1$ (left) and $w_1 = 3$ (right) respectively when the covariates are independent.

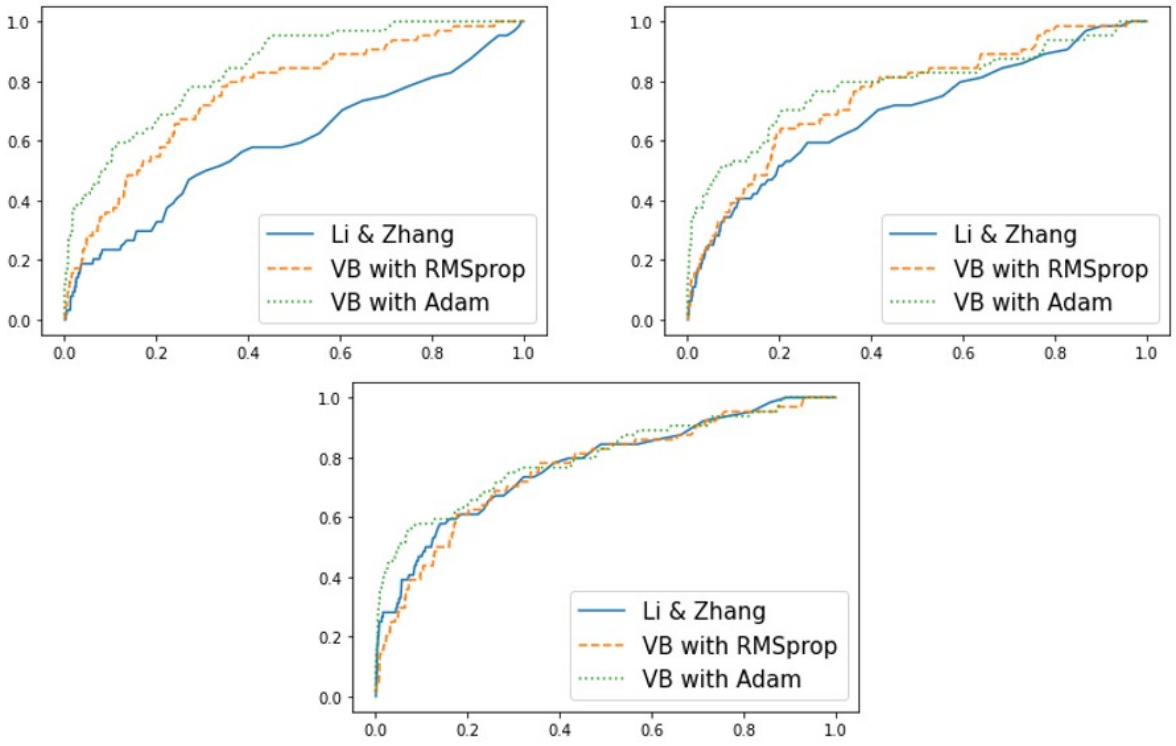

Figure 3.4: ROC curves for the three variable selection methods with the hyper parameter $w_1 = 5$ (left), $w_1 = 7$ (right), and $w_1 = 9$ (bottom) respectively when the covariates are independent.

For the second type of $X \in \mathbb{R}^{n \times p}$, we consider correlated X. In the blocks $[241, 265]$, $[741, 765]$, $[1241, 1265]$, and $[1741, 1765]$, we let $corr\,(X_{\cdot i}, X_{\cdot j}) = 0.75 - 0.03|i-j|$, where $X_{\cdot i}$ is i-th column of X. Also, as a noise, we let X be correlated as $corr\,(X_{\cdot i}, X_{\cdot j}) = 0.4 - 0.02|i-j|$, in four blocks which do not include a signal, $[41, 60]$, $[941, 960]$, $[1041, 1060]$, and $[1941, 1960]$.

With the hyper parameters $w_1 = 1$, $w_1 = 3$, $w_1 = 5$, $w_1 = 7$, and $w_1 = 9$, the corresponding ROC curves are demonstrated Figure 3.5 and Figure 3.6.

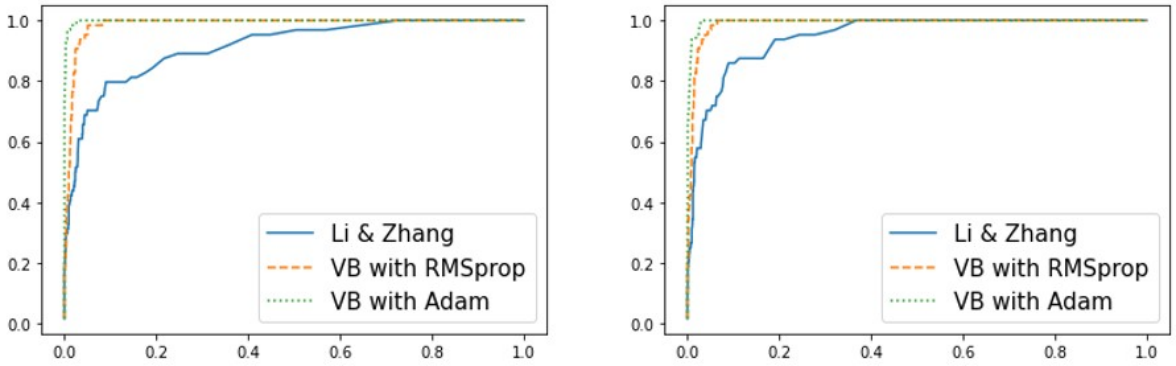

Figure 3.5: ROC curves for the three variable selection methods with the hyper parameter $w_1 = 1$ (left) and $w_1 = 3$ (right) respectively when the covariates are correlated.

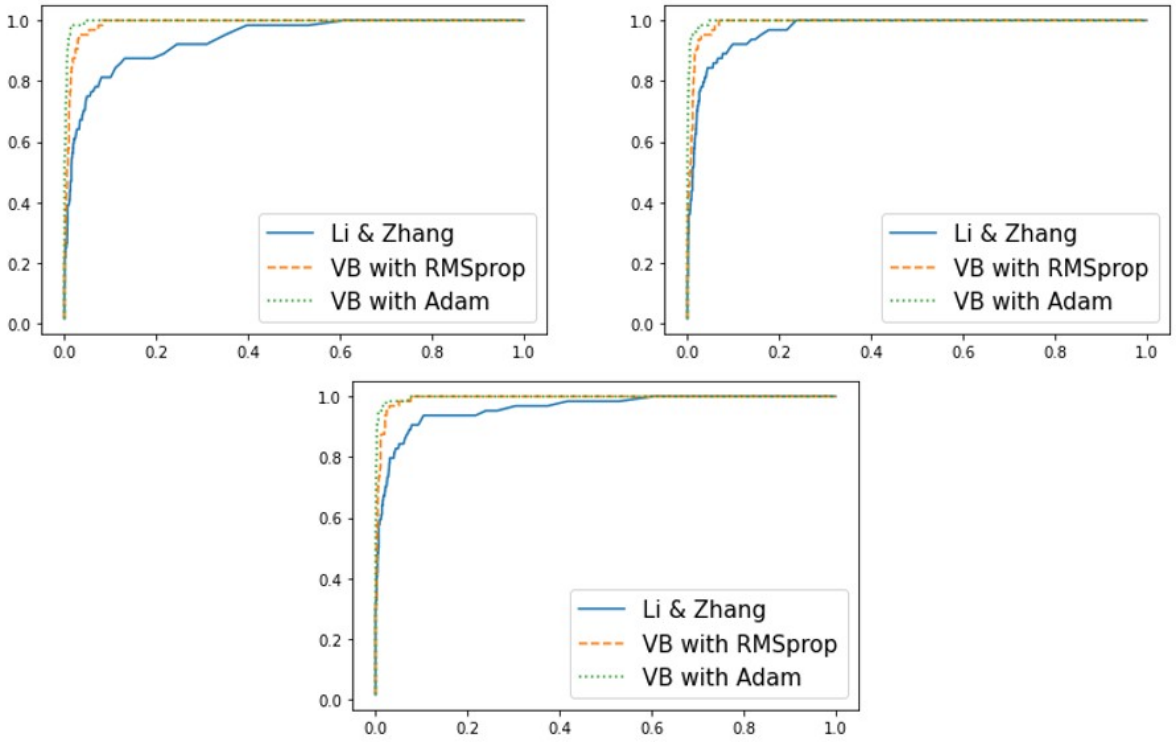

Figure 3.6: ROC curves for the three variable selection methods with the hyper parameter $w_1 = 5$ (left), $w_1 = 7$ (right), and $w_1 = 9$ (bottom) respectively when the covariates are correlated.

Scenario 2: For the second scenario, we assume the γ_is are arranged on a lattice such that each γ_i is connected with less then or equal to four neighbors. The four γ_is at the corner

of the lattice have two neighbors each, the γ_is at the boundary of the lattice have three neighbors, and the others have four neighbors. The connectivity is shown in Figure 3.7.

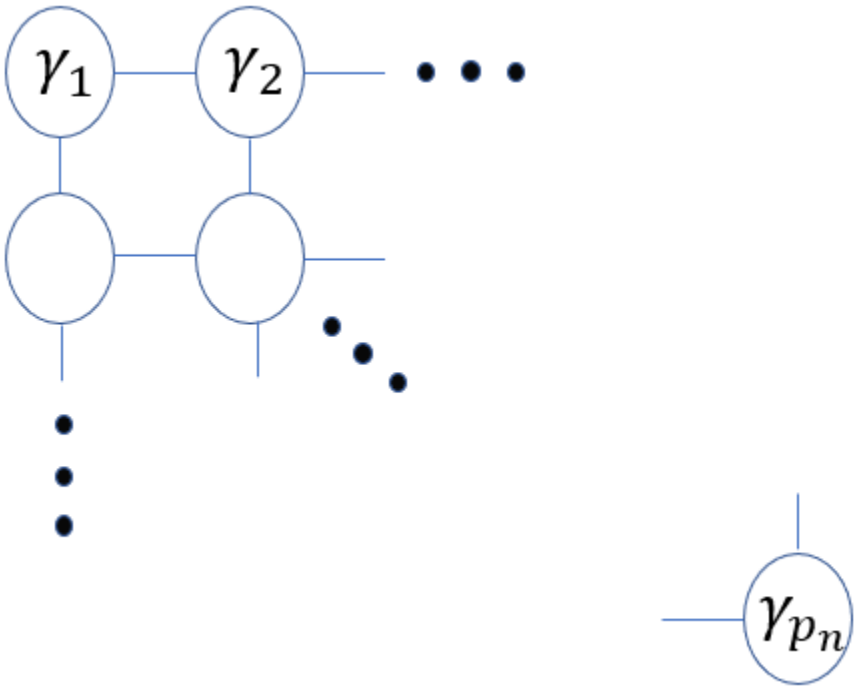

Figure 3.7: $\boldsymbol{\gamma}$ on a lattice.

For the true regression coefficients, we consider a vectorization of an image in which there are three signal areas. Each pixel in the signal areas takes a nonzero value, $\beta_i \in \{0.3, 0.6\}$, and all the other pixels are zeros. Figure 3.8 shows the signal areas in the image. The black pixels represent larger value of nonzero β_is (0.6), the grey pixels represent smaller value of nonzero β_is (0.3), and the white pixels represent zero β_is. We use the vecotrization of the image (matrix) as the vector of true regression coefficients in scenario 2.

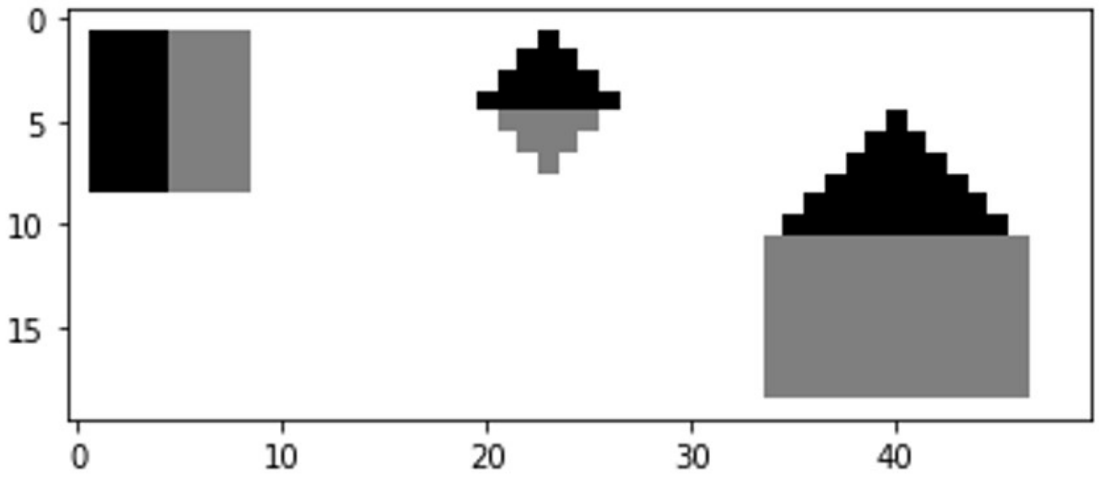

Figure 3.8: Signal areas in an image

For design matrix $X \in \mathbb{R}^{n \times p}$ in this scenario, we also consider two types of X. We first use independent standard Gaussian distribution as in the scenario 1. Figure 3.9 demonstrates the ROC curves corresponding to the hyper parameters $w_1 = 1$, $w_1 = 3$, and $w_1 = 5$.

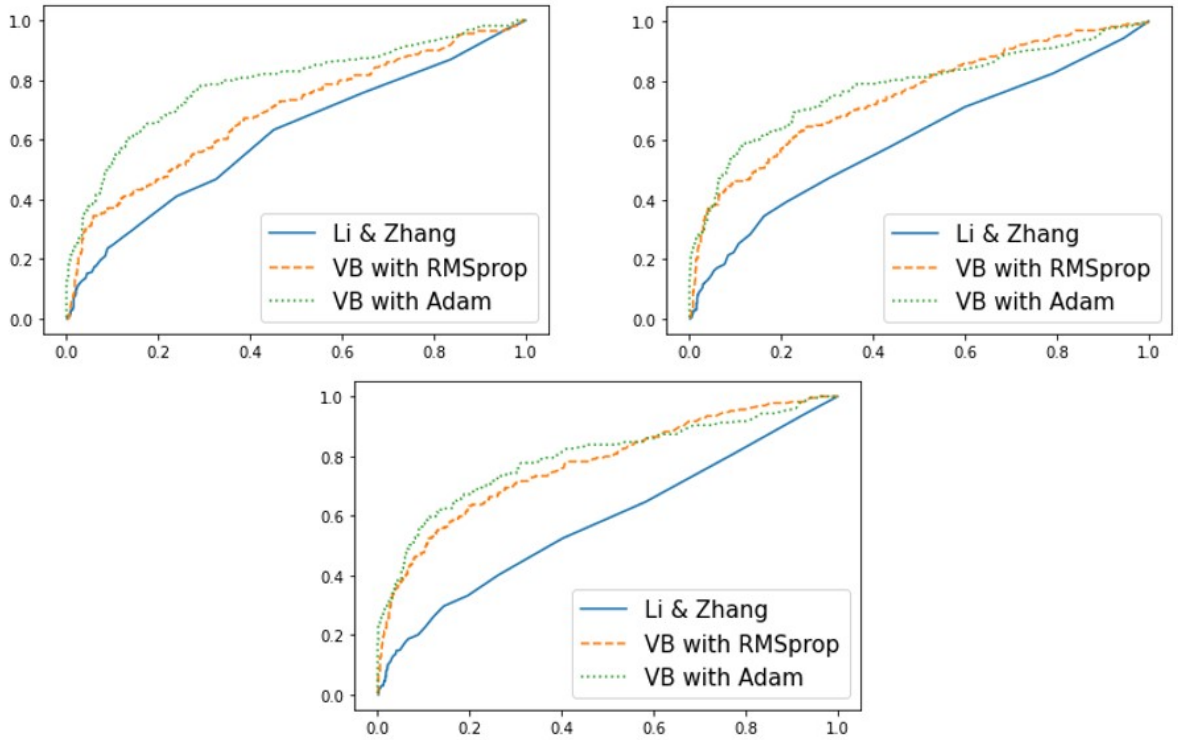

Figure 3.9: ROC curves in scenario 2 for the three variable selection methods with the hyper parameter $w_1 = 1$ (left), $w_1 = 3$ (right), and $w_1 = 5$ (bottom) respectively when the covariates are independent.

When w_1 is higher than 5, Li and Zhang (2010)'s approach does not work because of a phase transition. A phase transition can occur when the space of $\boldsymbol{\gamma}$ is two-dimensional (lattice). One can see Stanley (1971) for more details of phase transition in Ising model. Figure 3.10 contains the results for higher $w_1 = 7$ and $w_1 = 9$ when X is independent.

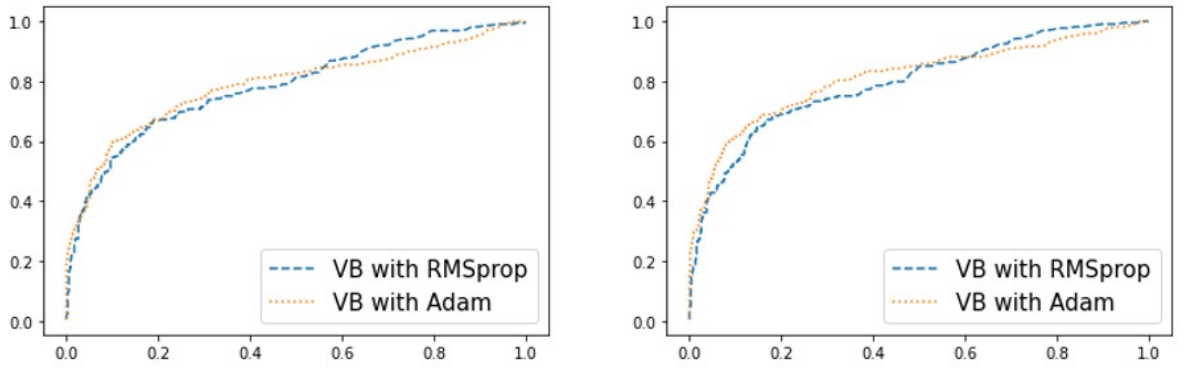

Figure 3.10: ROC curves in scenario 2 for the two VB methods with the hyper parameter $w_1 = 7$ (left) and $w_1 = 9$ (right) respectively when the covariates are independent.

For the second type of design matrix in scenario 2, we consider the same structure of $\boldsymbol{\gamma}$, that is, $corr\,(X_{\cdot i}, X_{\cdot j}) > 0$ if γ_i and γ_j are connected:

$$corr\,(X_{\cdot i}, X_{\cdot j}) = \begin{cases} 0.3 \text{ if } \gamma_i \text{ and } \gamma_j \text{ are connected,} \\ 0, \text{ otherwise.} \end{cases}$$

Using the correlated X, the ROC curves with $w_1 = 1$, $w_1 = 3$, and $w_1 = 5$ are shown in Figure 3.11 and the ROC curves for $w_1 = 7$ and $w_1 = 9$ are shown in Figure 3.12.

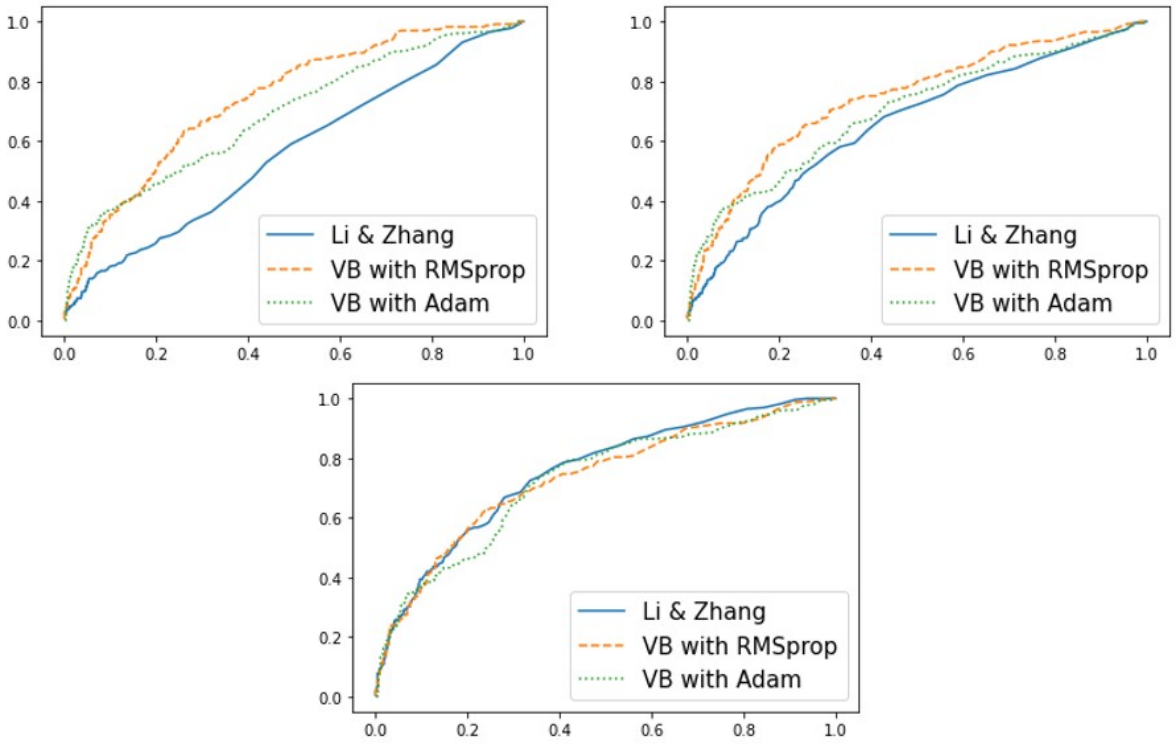

Figure 3.11: ROC curves in scenario 2 for the three variable selection methods with the hyper parameter $w_1 = 1$ (left), $w_1 = 3$ (right), and $w_1 = 5$ (bottom) respectively when the covariates are correlated.

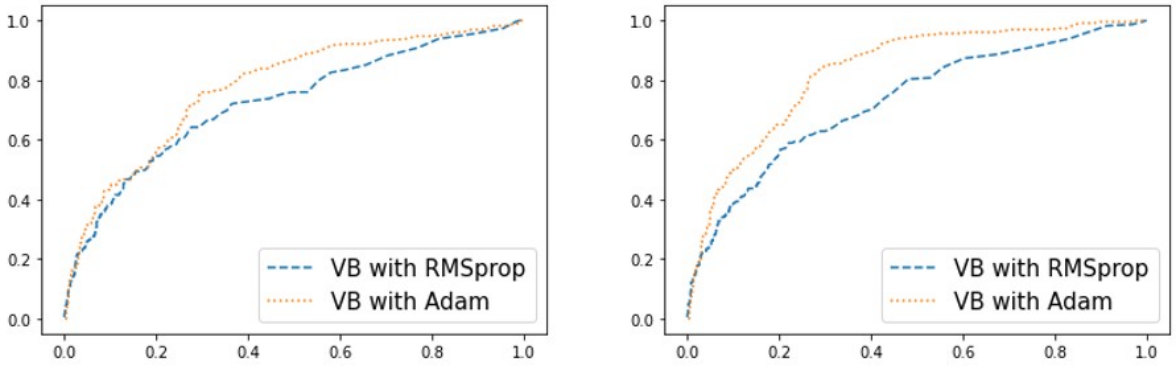

Figure 3.12: ROC curves in scenario 2 for the two VB methods with the hyper parameter $w_1 = 7$ (left) and $w_1 = 9$ (right) respectively when the covariates are correlated.

3.4 Theoretical results

In this section, we describe theoretical results. To establish the selection consistency of our variational Bayes method, we first observe the true posterior with the true Ising prior on $\boldsymbol{\gamma}$

as follows:

$$p(\boldsymbol{\gamma}) = \frac{1}{Z_n(a_0, b_0)} \exp\left(\frac{b_0}{2}\boldsymbol{\gamma}^\top J_n \boldsymbol{\gamma} + a_0 \sum_{i=1}^{p} \gamma_i\right),$$

where $J_n \in \mathbb{R}^{p_n \times p_n}$ is a given coupling matrix. We define an activation set $\mathcal{A}_{\boldsymbol{\gamma}} = \{i : \gamma_i = 1\}$ and, given $\boldsymbol{\gamma} = \boldsymbol{g}$, let $\boldsymbol{\beta}_{\boldsymbol{g}}$ be the vector of non-zero β_i's and $X_{\boldsymbol{g}}$ be the corresponding design matrix of size $n \times |\boldsymbol{\beta}_{\boldsymbol{g}}|$. Defining $|\boldsymbol{g}| := |\boldsymbol{\beta}_{\boldsymbol{g}}| = |\mathcal{A}_{\boldsymbol{g}}|$, the joint posterior is:

$$\begin{aligned}
&\pi(\boldsymbol{\beta}, \boldsymbol{\gamma} = \boldsymbol{g} \mid \boldsymbol{y}) \\
&\quad \propto p(\boldsymbol{\gamma} = \boldsymbol{g}) p(\boldsymbol{\beta} \mid \boldsymbol{\gamma} = \boldsymbol{g}) p(\boldsymbol{y} \mid \boldsymbol{\beta}_{\boldsymbol{g}}) \\
&\quad \propto \exp\left(a_0 \sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{g}^\top J_n \boldsymbol{g}\right) \exp\left(-\frac{1}{2}\left(\boldsymbol{y} - X_{\boldsymbol{g}}\boldsymbol{\beta}_{\boldsymbol{g}}\right)^\top \left(\boldsymbol{y} - X_{\boldsymbol{g}}\boldsymbol{\beta}_{\boldsymbol{g}}\right)\right) \\
&\qquad \times \prod_{i \in \mathcal{A}_{\boldsymbol{g}}} p(\beta_i \mid \gamma_i = 1) \prod_{i \in \mathcal{A}_{\boldsymbol{g}}^c} p(\beta_i \mid \gamma_i = -1) \\
&\quad \propto \exp\left(a_0 \sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{g}^\top J_n \boldsymbol{g}\right) \exp\left(-\frac{1}{2}\left(\boldsymbol{y} - X_{\boldsymbol{g}}\boldsymbol{\beta}_{\boldsymbol{g}}\right)^\top \left(\boldsymbol{y} - X_{\boldsymbol{g}}\boldsymbol{\beta}_{\boldsymbol{g}}\right)\right) \\
&\qquad \times \left(2\pi\tau^2\right)^{-\frac{|\boldsymbol{\beta}_{\boldsymbol{g}}|}{2}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}_{\boldsymbol{g}}^\top \boldsymbol{\beta}_{\boldsymbol{g}}\right) \prod_{i \in \mathcal{A}_{\boldsymbol{g}}^c} \mathbb{1}\left(\beta_i = 0\right) \\
&\quad \propto \exp\left(a_0 \sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{g}^\top J_n \boldsymbol{g}\right) \left(2\pi\tau^2\right)^{-\frac{|\boldsymbol{\beta}_{\boldsymbol{g}}|}{2}} \prod_{i \in \mathcal{A}_{\boldsymbol{g}}^c} \mathbb{1}\left(\beta_i = 0\right) \\
&\qquad \times \exp\left(-\frac{1}{2}\left(\left(\boldsymbol{y} - X_{\boldsymbol{g}}\boldsymbol{\beta}_{\boldsymbol{g}}\right)^\top \left(\boldsymbol{y} - X_{\boldsymbol{g}}\boldsymbol{\beta}_{\boldsymbol{g}}\right) + \frac{1}{\tau^2}\boldsymbol{\beta}_{\boldsymbol{g}}^\top \boldsymbol{\beta}_{\boldsymbol{g}}\right)\right) \\
&\quad \propto \exp\left(a_0 \sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{g}^\top J_n \boldsymbol{g}\right) \left(2\pi\tau^2\right)^{-\frac{|\boldsymbol{\beta}_{\boldsymbol{g}}|}{2}} \prod_{i \in \mathcal{A}_{\boldsymbol{g}}^c} \mathbb{1}\left(\beta_i = 0\right) \\
&\quad \propto \exp\left(a_0 \sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{g}^\top J_n \boldsymbol{g}\right) \left(2\pi\tau^2\right)^{-\frac{|\boldsymbol{\beta}_{\boldsymbol{g}}|}{2}} \prod_{i \in \mathcal{A}_{\boldsymbol{g}}^c} \mathbb{1}\left(\beta_i = 0\right) \\
&\qquad \times \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}_{\boldsymbol{g}} - \tilde{\boldsymbol{\beta}}_{\boldsymbol{g}}\right)^\top \left(X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}} + \frac{1}{\tau^2} I\right)\left(\boldsymbol{\beta}_{\boldsymbol{g}} - \tilde{\boldsymbol{\beta}}_{\boldsymbol{g}}\right)\right) \\
&\qquad \times \exp\left(-\frac{1}{2}\left(\boldsymbol{y}^\top \boldsymbol{y} - \tilde{\boldsymbol{\beta}}_{\boldsymbol{g}}\left(X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}} + \frac{1}{\tau^2} I\right)\tilde{\boldsymbol{\beta}}_{\boldsymbol{g}}\right)\right),
\end{aligned} \tag{3.7}$$

where $\tilde{\boldsymbol{\beta}}_{\boldsymbol{g}} = \left(X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}} + \frac{1}{\tau^2} I\right)^{-1} X_{\boldsymbol{g}}^\top \boldsymbol{y}$. To get the marginal posterior of $\boldsymbol{\gamma} = \boldsymbol{g}$, we integrate $\boldsymbol{\beta}_{\boldsymbol{g}}$ out first:

$$
\begin{aligned}
&\pi(\boldsymbol{\beta}_{-\boldsymbol{g}}, \boldsymbol{\gamma} = \boldsymbol{g} \mid \boldsymbol{y}) \\
&\quad = \int \pi(\boldsymbol{\beta}, \boldsymbol{\gamma} = \boldsymbol{g} \mid \boldsymbol{y}) d\boldsymbol{\beta}_{\boldsymbol{g}} \\
&\quad \propto \exp\left(-\frac{1}{2\sigma^2} R_{\boldsymbol{g}}\right) \exp\left(a_0 \sum_{i=1}^{p_n} g_i + \frac{b_n}{2} \boldsymbol{g}^\top J \boldsymbol{g}\right) \prod_{i \in \mathcal{A}_{\boldsymbol{g}}^c} \mathbb{1}\left(\beta_i = 0\right) \\
&\qquad \times \int \left(2\pi\sigma^2\tau^2\right)^{-\frac{|\boldsymbol{\beta}_{\boldsymbol{g}}|}{2}} \exp\left(-\frac{1}{2\sigma^2} \left(\boldsymbol{\beta}_{\boldsymbol{g}} - \tilde{\boldsymbol{\beta}}_{\boldsymbol{g}}\right)^\top \left(X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}} + \frac{1}{\tau^2} I\right) \left(\boldsymbol{\beta}_{\boldsymbol{g}} - \tilde{\boldsymbol{\beta}}_{\boldsymbol{g}}\right)\right) d\boldsymbol{\beta}_{\boldsymbol{g}} \\
&\quad \propto |X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}} + \frac{1}{\tau^2} I|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} R_{\boldsymbol{g}}\right) \exp\left(a_0 \sum_{i=1}^{p_n} g_i + \frac{b_n}{2} \boldsymbol{g}^\top J \boldsymbol{g}\right) \prod_{i \in \mathcal{A}_{\boldsymbol{g}}^c} \mathbb{1}\left(\beta_i = 0\right),
\end{aligned}
$$

where $R_{\boldsymbol{g}} = \boldsymbol{y}^\top \left(I - X_{\boldsymbol{g}} \left(X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}} + \frac{1}{\tau^2} I\right)^{-1} X_{\boldsymbol{g}}^\top\right) \boldsymbol{y}$. Now, we integrate $\boldsymbol{\beta}_{-\boldsymbol{g}}$ out:

$$
\begin{aligned}
&\pi(\boldsymbol{\gamma} = \boldsymbol{g} \mid \boldsymbol{y}) \\
&\quad = \int \pi(\boldsymbol{\beta}_{-\boldsymbol{g}}, \boldsymbol{\gamma} = \boldsymbol{g} \mid \boldsymbol{y}) d\boldsymbol{\beta}_{-\boldsymbol{g}} \\
&\quad \propto |X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}} + \frac{1}{\tau^2} I|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} R_{\boldsymbol{g}}\right) \exp\left(a_0 \sum_{i=1}^{p_n} g_i + \frac{b_0}{2} \boldsymbol{g}^\top J \boldsymbol{g}\right) \int \prod_{i \in \mathcal{A}_{\boldsymbol{g}}^c} \mathbb{1}\left(\beta_i = 0\right) d\boldsymbol{\beta}_{-\boldsymbol{g}} \\
&\quad = |X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}} + \frac{1}{\tau^2} I|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} R_{\boldsymbol{g}}\right) \exp\left(a_0 \sum_{i=1}^{p_n} g_i + \frac{b_0}{2} \boldsymbol{g}^\top J \boldsymbol{g}\right).
\end{aligned}
$$

Let $Q_{\boldsymbol{g}} = \left|X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}} + \frac{1}{\tau^2} I\right|^{-1/2}$. Then, the posterior ratio is:

$$
\begin{aligned}
&PR(\boldsymbol{g}, \boldsymbol{t}) \\
&\quad = \frac{\pi(\boldsymbol{\gamma} = \boldsymbol{g} \mid \boldsymbol{y})}{\pi(\boldsymbol{\gamma} = \boldsymbol{t} \mid \boldsymbol{y})} \\
&\quad = \underbrace{\frac{Q_{\boldsymbol{g}}}{Q_{\boldsymbol{t}}}}_{\text{(a)}} \underbrace{\exp\left(-\frac{1}{2\sigma^2} R_{\boldsymbol{g}} + \frac{1}{2\sigma^2} R_{\boldsymbol{t}}\right)}_{\text{(b)}} \underbrace{\exp\left(a_0 \sum_{i=1}^{p_n} g_i + \frac{b_0}{2} \boldsymbol{g}^\top J \boldsymbol{g} - a_0 \sum_{i=1}^{p_n} t_i - \frac{b_0}{2} \boldsymbol{t}^\top J \boldsymbol{t}\right)}_{\text{(c)}},
\end{aligned}
\tag{3.8}
$$

where $\boldsymbol{t}$ denotes the true model. Our first result is described in the following subsection.

3.4.1 True posterior consistency with true Ising prior

We let λ_{max}^n denote the maximum eigenvalue of the Gram matrix $X^\top X/n$, and for $\nu > 0$, as like Narisetty and He (2014), we define:

$$m_n(\nu) = p_n \wedge \frac{n}{(2+\nu)\log p_n} \quad \text{and} \quad \lambda_{min}^n(\nu) := \inf_{|\boldsymbol{g}| \le m_n(\nu)} \phi_{min}^{\#}\left(\frac{X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}}}{n}\right), \tag{3.9}$$

where $\phi_{min}^{\#}(A)$ denotes the minimum nonzero eigenvalue of a matrix A. To establish the selection consistency of the true posterior under the true Ising prior, we need following regularity conditions (Yang and Shen, 2017):

Condition 1. *(On dimension p_n). $p_n = e^{nd_n}$ for some $d_n \to 0$ as $n \to \infty$, that is, $\log p_n = o(n)$.*

Condition 2. *(On prior parameters). $n\tau^2 \sim (n \vee p_n^{2+3\delta})$, for some $\delta > 0$, $a_0 \sim -nd_n$, and $b_0 \sim \frac{1}{k_{max} p_n}$, where k_{max} is the maximum row sum of J_n, that is,*

$$k_{max} = \max_{i \in \{1,\ldots,p_n\}} \sum_{j=1}^{p_n} J_n(i,j).$$

Condition 3. *(On true model). $\boldsymbol{y} \mid X \sim \mathcal{N}_n(X_{\boldsymbol{t}}\boldsymbol{\beta}_{\boldsymbol{t}} + X_{\boldsymbol{t}^c}\boldsymbol{\beta}_{\boldsymbol{t}^c}, \sigma^2 I)$, where the size of the true model $|\boldsymbol{t}|$ is fixed. Besides, $|\boldsymbol{t}|/2 < r_{\boldsymbol{t}} < |\boldsymbol{t}|$, where $r_{\boldsymbol{t}}$ is the rank of $X_{\boldsymbol{t}}$.*

For any fixed $K > 0$, define

$$\Delta_n(K) := \inf_{\{\boldsymbol{g}: |\boldsymbol{g}| < K|\boldsymbol{t}|, \boldsymbol{g} \not\supset \boldsymbol{t}\}} ||(I - P_{\boldsymbol{g}})X_{\boldsymbol{t}}\boldsymbol{\beta}_{\boldsymbol{t}}||_2^2,$$

where $P_{\boldsymbol{g}}$ is the projection matrix onto the column space of $X_{\boldsymbol{g}}$.

Condition 4. *(Identifiability). There is $K > 1 + 8/\delta$ such that $\Delta_n(K) > \gamma_n := 5\sigma^2|\boldsymbol{t}|(1+\delta)\log(\sqrt{n} \vee p_n)$.*

Condition 5. *(Regularity of the design). For some $\nu < \delta$, $\kappa < (K-1)\delta/2$,*

$$\lambda_{max}^n \sim O(1) \quad \textit{and} \quad \lambda_{min}^n \succeq \left(\frac{n \vee p^{2+2\delta}}{n\tau^2} \vee p_n^{-\kappa}\right).$$

Theorem 2. *Under the Conditions 1 - 5, the posterior probability of the true model goes to 1, that is,*

$$\sum_{\boldsymbol{g}\neq\boldsymbol{t}} PR(\boldsymbol{g},\boldsymbol{t}) \to 0,$$

as n goes to infinity.

Narisetty and He (2014) separated the model space into four disjoint parts:

$$\begin{aligned}
M_1 &= \{\boldsymbol{g} : r_{\boldsymbol{g}} > m_n\},\\
M_2 &= \{\boldsymbol{g} : \boldsymbol{g} \supset \boldsymbol{t}, r_{\boldsymbol{g}} \leq m_n\},\\
M_3 &= \{\boldsymbol{g} : \boldsymbol{g} \not\supset \boldsymbol{t}, K|\boldsymbol{t}| < r_{\boldsymbol{g}} \leq m_n\},\\
M_4 &= \{\boldsymbol{g} : \boldsymbol{g} \not\supset \boldsymbol{t}, r_{\boldsymbol{g}} \leq K|\boldsymbol{t}|\}.
\end{aligned}$$

To prove the Theorem 2, we show $\sum_{\boldsymbol{g}} PR(\boldsymbol{g},\boldsymbol{t}) \to 0$ in each subspace.

Lemma 16. *For a universal constant $c' > 0$,*

$$\frac{Q_{\boldsymbol{g}}}{Q_{\boldsymbol{t}}} \leq c' \left(n\tau^2\lambda_{min}^n\right)^{-\left(r_{\boldsymbol{g}}^* - |\boldsymbol{t}|\right)/2} \left(\lambda_{min}^n\right)^{-|\boldsymbol{t}|/2}$$

Proof. From Lemma 11.1 in Narisetty and He (2014), we have

$$\begin{aligned}
Q_{\boldsymbol{g}} &= \left| I + \frac{1}{\tau^2} X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}} \right|^{-1/2}\\
&= \left| I + \tau^2 X_{\boldsymbol{g}} X_{\boldsymbol{g}}^\top \right|^{-1/2}
\end{aligned}$$

Note,

$$\text{(a)} = \frac{Q_{\boldsymbol{g}}}{Q_{\boldsymbol{t}}} = \frac{Q_{\boldsymbol{g}}}{Q_{\boldsymbol{g}\wedge\boldsymbol{t}}} \cdot \frac{Q_{\boldsymbol{g}\wedge\boldsymbol{t}}}{Q_{\boldsymbol{t}}},$$

$$
\begin{aligned}
\frac{Q_{\boldsymbol{g}}}{Q_{\boldsymbol{g}\wedge\boldsymbol{t}}} &= |I+\tau^2 X_{\boldsymbol{g}} X_{\boldsymbol{g}}^\top|^{-1/2} |I+\tau^2 X_{\boldsymbol{g}\wedge\boldsymbol{t}} X_{\boldsymbol{g}\wedge\boldsymbol{t}}^\top|^{1/2} \\
&\le \left(1+\tau^2\lambda^n_{\boldsymbol{g},min}\right)^{-r_{\boldsymbol{g}}/2} \left(1+\tau^2\lambda^n_{\boldsymbol{g}\wedge\boldsymbol{t},max}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}}/2} \\
&\le \left(\tau^2\lambda^n_{\boldsymbol{g},min}\right)^{-r_{\boldsymbol{g}}/2} \left(1+n\tau^2\lambda^n_{max}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}}/2} \\
&\le \left(n\tau^2\lambda^n_{min}\right)^{-r_{\boldsymbol{g}}/2} \left(1+n\tau^2\lambda^n_{max}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}}/2} \\
&\simeq \left(n\tau^2\lambda^n_{min}\right)^{-r_{\boldsymbol{g}}/2} \left(n\tau^2\lambda^n_{max}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}}/2} \text{ for sufficiently large } n \\
&= \left(n\tau^2\lambda^n_{min}\right)^{-(r_{\boldsymbol{g}}-r_{\boldsymbol{t}\wedge\boldsymbol{g}})/2} \left(\frac{\lambda^n_{max}}{\lambda^n_{min}}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}}/2} \\
&\le \left(n\tau^2\lambda^n_{min}\right)^{-\left(r^*_{\boldsymbol{g}}-r_{\boldsymbol{t}\wedge\boldsymbol{g}}\right)/2} \left(\frac{\lambda^n_{max}}{\lambda^n_{min}}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}}/2} \\
&\le \left(n\tau^2\lambda^n_{min}\right)^{-\left(r^*_{\boldsymbol{g}}-r_{\boldsymbol{t}}\right)/2} \left(\frac{\lambda^n_{max}}{\lambda^n_{min}}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}}/2} \\
&\le C\left(n\tau^2\lambda^n_{min}\right)^{-\left(r^*_{\boldsymbol{g}}-r_{\boldsymbol{t}}\right)/2} \left(\lambda^n_{min}\right)^{-r_{\boldsymbol{t}\wedge\boldsymbol{g}}/2},
\end{aligned}
$$

where $r^*_{\boldsymbol{g}} = r_{\boldsymbol{g}} \wedge m_n$.

$$
\begin{aligned}
\frac{Q_{\boldsymbol{g}\wedge\boldsymbol{t}}}{Q_{\boldsymbol{t}}} &= |I+\tau^2 X_{\boldsymbol{g}\wedge\boldsymbol{t}} X_{\boldsymbol{g}\wedge\boldsymbol{t}}^\top|^{-1/2} |I+\tau^2 X_{\boldsymbol{t}} X_{\boldsymbol{t}}^\top|^{1/2} \\
&= |I+\tau^2 X_{\boldsymbol{g}\wedge\boldsymbol{t}} X_{\boldsymbol{g}\wedge\boldsymbol{t}}^\top + \tau^2 X_{\boldsymbol{g}^c\wedge\boldsymbol{t}} X_{\boldsymbol{g}^c\wedge\boldsymbol{t}}^\top|^{1/2} |I+\tau^2 X_{\boldsymbol{g}\wedge\boldsymbol{t}} X_{\boldsymbol{g}\wedge\boldsymbol{t}}^\top|^{-1/2} \\
&= |\left(I+\tau^2 X_{\boldsymbol{g}\wedge\boldsymbol{t}} X_{\boldsymbol{g}\wedge\boldsymbol{t}}^\top + \tau^2 X_{\boldsymbol{g}^c\wedge\boldsymbol{t}} X_{\boldsymbol{g}^c\wedge\boldsymbol{t}}^\top\right)^{-1} \left(I+\tau^2 X_{\boldsymbol{g}\wedge\boldsymbol{t}} X_{\boldsymbol{g}\wedge\boldsymbol{t}}^\top\right)|^{-1/2} \\
&= |I+\tau^2 X_{\boldsymbol{g}^c\wedge\boldsymbol{t}}^\top \left(I+\tau^2 X_{\boldsymbol{g}\wedge\boldsymbol{t}} X_{\boldsymbol{g}\wedge\boldsymbol{t}}^\top\right)^{-1} X_{\boldsymbol{g}^c\wedge\boldsymbol{t}}|^{1/2} \\
&\le |I+\tau^2 X_{\boldsymbol{g}^c\wedge\boldsymbol{t}}^\top X_{\boldsymbol{g}^c\wedge\boldsymbol{t}}|^{1/2} \\
&= |I+\tau^2 X_{\boldsymbol{g}^c\wedge\boldsymbol{t}} X_{\boldsymbol{g}^c\wedge\boldsymbol{t}}^\top|^{1/2} \\
&\le \left(1+\tau^2\lambda^n_{\boldsymbol{g}^c\wedge\boldsymbol{t},max}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}^c}/2} \\
&\le \left(1+\tau^2\lambda^n_{max}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}^c}/2} \\
&\simeq \left(n\tau^2\lambda^n_{max}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}^c}/2} \text{ for sufficiently large } n \\
&= \left(n\tau^2\lambda^n_{min}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}^c}/2} \left(\frac{\lambda^n_{max}}{\lambda^n_{min}}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}^c}/2} \\
&\le C\left(n\tau^2\lambda^n_{min}\right)^{r_{\boldsymbol{t}\wedge\boldsymbol{g}^c}/2} \left(\lambda^n_{min}\right)^{-r_{\boldsymbol{t}\wedge\boldsymbol{g}^c}/2}, \ \lambda^n_{max} \sim O(1)
\end{aligned}
$$

From the above two inequalities,

$$\begin{aligned}
\text{(a)} = \frac{Q_{\boldsymbol{g}}}{Q_{\boldsymbol{t}}} &\leq c' \left(n\tau^2\lambda_{min}^n\right)^{-\left(r_{\boldsymbol{g}}^* - r_{\boldsymbol{t}\wedge\boldsymbol{g}} - r_{\boldsymbol{t}\wedge\boldsymbol{g}^c}\right)/2} \left(\frac{\lambda_{max}^n}{\lambda_{min}^n}\right)^{\left(r_{\boldsymbol{t}\wedge\boldsymbol{g}} + r_{\boldsymbol{t}\wedge\boldsymbol{g}^c}\right)/2} \\
&\leq c' \left(n\tau^2\lambda_{min}^n\right)^{-\left(r_{\boldsymbol{g}}^* - |\boldsymbol{t}\wedge\boldsymbol{g}| - |\boldsymbol{t}\wedge\boldsymbol{g}^c|\right)/2} \left(\frac{\lambda_{max}^n}{\lambda_{min}^n}\right)^{\left(|\boldsymbol{t}\wedge\boldsymbol{g}| + |\boldsymbol{t}\wedge\boldsymbol{g}^c|\right)/2} \\
&= c' \left(n\tau^2\lambda_{min}^n\right)^{-\left(r_{\boldsymbol{g}}^* - |\boldsymbol{t}|\right)/2} \left(\frac{\lambda_{max}^n}{\lambda_{min}^n}\right)^{|\boldsymbol{t}|/2} \\
&\leq c' \left(n\tau^2\lambda_{min}^n\right)^{-\left(r_{\boldsymbol{g}}^* - |\boldsymbol{t}|\right)/2} \left(\lambda_{min}^n\right)^{-|\boldsymbol{t}|/2}
\end{aligned}$$

□

Using the same argument of the Lemma A.1 in Narisetty and He (2014), we show that:

$$\begin{aligned}
R_{\boldsymbol{g}} = R_{\boldsymbol{g}} &= \boldsymbol{y}^\top \left(I - X_{\boldsymbol{g}} \left(X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}} + \frac{1}{\tau^2} I\right)^{-1} X_{\boldsymbol{g}}^\top\right) \boldsymbol{y} \\
&= \boldsymbol{y}^\top \left(I + \tau^2 X_{\boldsymbol{g}} X_{\boldsymbol{g}}^\top\right)^{-1} \boldsymbol{y} \geq 0, \forall \boldsymbol{g}
\end{aligned} \tag{3.10}$$

1. Models in M_1. From (3.10) and the supplement to Narisetty and He (2014),

$$\begin{aligned}
\mathbb{P}\left[R_{\boldsymbol{t}} - R_{\boldsymbol{g}} > n(1+2s)\sigma^2\right] &\leq \mathbb{P}\left[R_{\boldsymbol{t}} > n(1+2s)\sigma^2\right] \\
&\leq 2e^{-c'n}, \text{ uniformly for all } \boldsymbol{g}
\end{aligned} \tag{3.11}$$

Observe that on M_1, $r_{\boldsymbol{g}}^* = m_n \geq n/\log(p_n^{2+\nu}) \geq n/\log(p_n^{2+\delta})$. Therefore, on the high-

probability event $\{R_{\boldsymbol{t}} - R_{\boldsymbol{g}} \leq n(1+2s)\sigma^2\}$, by (3.11) and the regularity Conditions,

$$
\begin{aligned}
&\sum_{\boldsymbol{g}\in M_1} PR(\boldsymbol{g},\boldsymbol{t}) \\
&\quad \preceq \sum_{\boldsymbol{g}\in M_1} \left(n\tau^2\lambda_{min}^n\right)^{-\left(r_{\boldsymbol{g}}^* - |\boldsymbol{t}|\right)/2} \left(\frac{\lambda_{max}^n}{\lambda_{min}^n}\right)^{|\boldsymbol{t}|/2} \exp\left(\frac{n(1+2s)}{2}\right) \\
&\qquad\qquad \exp\left(a_0\sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{g}^\top J\boldsymbol{g} - a_0\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \\
&\quad \preceq \sum_{\boldsymbol{g}\in M_1} p_n^{-(1+\delta)(m_n - |\boldsymbol{t}|)} \left(\lambda_{min}^n\right)^{-|\boldsymbol{t}|/2} \exp\left(\frac{n(1+2s)}{2}\right) \\
&\qquad\qquad \exp\left(a_0\sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{g}^\top J\boldsymbol{g} - a_0\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right), \quad \text{by Condition 2 and 5} \\
&\quad = \sum_{\boldsymbol{g}\in M_1} \exp\left(-nd_n(1+\delta)(m_n - |\boldsymbol{t}|)\right) \left(\lambda_{min}^n\right)^{-|\boldsymbol{t}|/2} \exp\left(\frac{n(1+2s)}{2}\right) \\
&\qquad\qquad \exp\left(a_0\sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{g}^\top J\boldsymbol{g} - a_0\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right), \quad \text{by Condition 1} \\
&\quad \preceq \sum_{\boldsymbol{g}\in M_1} \exp\left(-\frac{n(1+\delta)}{2+\delta}\right) \left(\lambda_{min}^n\right)^{-|\boldsymbol{t}|/2} \exp\left(\frac{n(1+2s)}{2}\right) \\
&\qquad\qquad \exp\left(a_0\sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{g}^\top J\boldsymbol{g} - a_0\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right), \quad \text{because } m_n \geq n/\log(p_n^{2+\delta}) \\
&\quad \preceq \sum_{\boldsymbol{g}\in M_1} \exp\left(-\frac{n(1+\delta)}{2+\delta}\right) p_n^{\kappa|\boldsymbol{t}|/2} \exp\left(\frac{n(1+2s)}{2}\right) \\
&\qquad\qquad \exp\left(a_0\sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{g}^\top J\boldsymbol{g} - a_0\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right), \quad \text{by Condition 5} \\
&\quad \preceq \exp\left(-\frac{n(1+\delta)}{2+\delta} + \frac{n(1+2s)}{2}\right) p_n^{\kappa|\boldsymbol{t}|/2} \exp\left(-a_0\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \\
&\qquad\qquad \sum_{\boldsymbol{g}\in M_1} \exp\left(a_0\sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{g}^\top J\boldsymbol{g}\right), \quad \text{by Condition 5.}
\end{aligned}
$$

Since the conditions on a_0 and b_0 bounds the summation term, we have

$$
\begin{aligned}
\sum_{\boldsymbol{g}\in M_1} PR(\boldsymbol{g},\boldsymbol{t}) &\preceq w' \exp\left(\frac{nd_n\kappa|\boldsymbol{t}|}{2}\right) \exp\left(-\frac{n(1+\delta)}{2+\delta} + \frac{n(1+2s)}{2}\right), \quad \text{by Condition 1} \\
&\preceq \exp\left\{n\left(\frac{1+2s}{2} - \frac{1+\delta}{2+\delta}\right)\right\} \to 0,
\end{aligned}
$$

when $(1+2s)/2-(1+\delta)/(2+\delta))<0$, i.e., $s<\delta/2(2+\delta)$.

2. Models in M_2. Consider $0<s\leq\delta/8$, and define the events:

$$A(\boldsymbol{g}) := \{R_{\boldsymbol{t}} - R_{\boldsymbol{g}} > 2\sigma^2(1+2s)(r_{\boldsymbol{g}} - r_{\boldsymbol{t}})\log p_n\},$$
$$U(d) := \cup_{\{\boldsymbol{g}:r_{\boldsymbol{g}}=d\}} A(\boldsymbol{g}).$$

Let $P_{\boldsymbol{g}} = X_{\boldsymbol{g}}(X_{\boldsymbol{g}}^\top X_{\boldsymbol{g}})^{-1}X_{\boldsymbol{g}}^\top$. Since $R_{\boldsymbol{g}} \geq R_{\boldsymbol{g}}^* = \boldsymbol{y}^\top(I-P_{\boldsymbol{g}})\boldsymbol{y}$, we have

$$\begin{aligned}
\mathbb{P}[U(d)] &= \mathbb{P}\left[\cup_{\{\boldsymbol{g}:r_{\boldsymbol{g}}=d\}}\{R_{\boldsymbol{t}} - R_{\boldsymbol{g}} > 2\sigma^2(1+2s)(r_{\boldsymbol{g}}-r_{\boldsymbol{t}})\log p_n\}\right] \\
&\leq \mathbb{P}\left[\cup_{\{\boldsymbol{g}:r_{\boldsymbol{g}}=d\}}\{R_{\boldsymbol{t}} - R_{\boldsymbol{g}}^* > 2\sigma^2(1+2s)(r_{\boldsymbol{g}}-r_{\boldsymbol{t}})\log p_n\}\right] \\
&\leq \mathbb{P}\left[\cup_{\{\boldsymbol{g}:r_{\boldsymbol{g}}=d\}}\{R_{\boldsymbol{t}}^* - R_{\boldsymbol{g}}^* > \sigma^2(2+3s)(r_{\boldsymbol{g}}-r_{\boldsymbol{t}})\log p_n\}\right] \\
&\quad + \mathbb{P}\left[R_{\boldsymbol{t}} - R_{\boldsymbol{t}}^* > s\sigma^2(d-r_{\boldsymbol{t}})\log p_n\right] \\
&\leq c' p_n^{-(1+s)(d-r_{\boldsymbol{t}})} p_n^{(d-r_{\boldsymbol{t}})} + \exp\left(-c' n\log p_n\right) \\
&\leq 2c' p_n^{-s(d-r_{\boldsymbol{t}})}.
\end{aligned}$$

Next, we consider the union of all such events $U(d)$, that is,

$$\mathbb{P}\left[\cup_{\{d>r_{\boldsymbol{t}}\}}U(d)\right] \leq \frac{2c'}{p_n^s - 1} \to 0.$$

Observe that on M_2, we have $r_{\boldsymbol{g}}^* = r_{\boldsymbol{g}}$, $|\boldsymbol{t}\wedge\boldsymbol{g}| = |\boldsymbol{t}|$, and $|\boldsymbol{t}\wedge\boldsymbol{g}^c| = 0$. On the high-probability

event $\cap_{\{d>r_{\boldsymbol{t}}\}}U(d)^c$,

$$
\begin{aligned}
&\sum_{\boldsymbol{k}\in M_2} PR(\boldsymbol{k},\boldsymbol{t}) \\
&\quad\preceq \sum_{\boldsymbol{k}\in M_2} \left(n\tau^2\lambda_{min}^n\right)^{-\left(r_{\boldsymbol{k}}^*-r_{\boldsymbol{t}\wedge\boldsymbol{k}}\right)/2} \left(\frac{\lambda_{max}^n}{\lambda_{min}^n}\right)^{|\boldsymbol{t}|/2} \exp\left(-\frac{1}{2\sigma^2}\left(R_{\boldsymbol{k}}-R_{\boldsymbol{t}}\right)\right) \\
&\qquad\qquad \exp\left(a_n\sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{k}^\top J\boldsymbol{k} - a_0\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \\
&\quad\preceq \sum_{\boldsymbol{k}\in M_2} \left(n\tau^2\lambda_{min}^n\right)^{-(r_{\boldsymbol{k}}-r_{\boldsymbol{t}})/2} p_n^{(1+2s)(r_{\boldsymbol{k}}-r_{\boldsymbol{t}})} \left(\lambda_{min}^n\right)^{-|\boldsymbol{t}|/2} \\
&\qquad\qquad \exp\left(a_0\sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{k}^\top J\boldsymbol{k} - a_0\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \\
&\quad\preceq \sum_{\boldsymbol{k}\in M_2} \left(p_n^{2+2\delta}\vee n\right)^{-(r_{\boldsymbol{k}}-r_{\boldsymbol{t}})/2} p_n^{(1+2s)(r_{\boldsymbol{k}}-r_{\boldsymbol{t}})} p_n^{\kappa|\boldsymbol{t}|/2} \\
&\qquad\qquad \exp\left(a_0\sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{k}^\top J\boldsymbol{k} - a_0\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \\
&\quad\preceq \sum_{\boldsymbol{k}\in M_2} \left(p_n^{-1-\delta}\wedge 1/\sqrt{n}\right)^{(r_{\boldsymbol{k}}-r_{\boldsymbol{t}})} p_n^{(1+2s)(r_{\boldsymbol{k}}-r_{\boldsymbol{t}})} p_n^{\kappa|\boldsymbol{t}|/2} \\
&\qquad\qquad \exp\left(a_0\sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{k}^\top J\boldsymbol{k} - a_0\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \\
&\quad\preceq \sum_{\boldsymbol{k}\in M_2} \left(p_n^{-\delta+2s}\wedge\frac{p_n^{1+2s}}{\sqrt{n}}\right)^{(r_{\boldsymbol{k}}-r_{\boldsymbol{t}})} p_n^{\kappa|\boldsymbol{t}|/2} \\
&\qquad\qquad \exp\left(a_0\sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{k}^\top J\boldsymbol{k} - a_0\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \\
&\quad\preceq \left(p_n^{-3\delta/4}\wedge\frac{p_n^{1+\delta/4}}{\sqrt{n}}\right) p_n^{\kappa|\boldsymbol{t}|/2} \\
&\qquad\qquad \exp\left(-a_0\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \sum_{\boldsymbol{k}\in M_2} \exp\left(a_0\sum_{i=1}^{p_n} g_i + \frac{b_0}{2}\boldsymbol{k}^\top J\boldsymbol{k}\right) \\
&\quad\preceq \left(p_n^{\frac{\kappa|\boldsymbol{t}|}{2}-\frac{3}{4}\delta}\wedge\frac{p_n^{\frac{\kappa|\boldsymbol{t}|}{2}+1+\frac{\delta}{4}}}{\sqrt{n}}\right)\to 0, \text{ for some } \delta.
\end{aligned}
$$

3. Models in M_3. We define:

$$B(\boldsymbol{k}) := \{R_{\boldsymbol{t}} - R_{\boldsymbol{k}} > 2\sigma^2(1+2s)(r_{\boldsymbol{k}} - r_{\boldsymbol{t}})\log p_n\}.$$

Note that:

$$R_{\boldsymbol{k}} = \boldsymbol{y}^\top \left(I + \tau^2 X_{\boldsymbol{k}} X_{\boldsymbol{k}}^\top\right)^{-1} \boldsymbol{y}, \quad R_{\boldsymbol{k}\vee\boldsymbol{t}} = \boldsymbol{y}^\top \left(I + \tau^2 X_{\boldsymbol{k}} X_{\boldsymbol{k}}^\top + \tau^2 X_{\boldsymbol{k}^c\wedge\boldsymbol{t}} X_{\boldsymbol{k}^c\wedge\boldsymbol{t}}^\top\right)^{-1} \boldsymbol{y}.$$

Let $A = I + \tau^2 X_{\boldsymbol{k}} X_{\boldsymbol{k}}^\top$.

$$\begin{aligned}\left(I + \tau^2 X_{\boldsymbol{k}} X_{\boldsymbol{k}}^\top + \tau^2 X_{\boldsymbol{k}^c\wedge\boldsymbol{t}} X_{\boldsymbol{k}^c\wedge\boldsymbol{t}}^\top\right)^{-1} &= \left(A + \tau^2 X_{\boldsymbol{k}^c\wedge\boldsymbol{t}} X_{\boldsymbol{k}^c\wedge\boldsymbol{t}}^\top\right)^{-1} \\ &= A^{-1} - A^{-1} X_{\boldsymbol{k}^c\wedge\boldsymbol{t}} \left(I + X_{\boldsymbol{k}^c\wedge\boldsymbol{t}}^\top A^{-1} X_{\boldsymbol{k}^c\wedge\boldsymbol{t}}\right) X_{\boldsymbol{k}^c\wedge\boldsymbol{t}}^\top A^{-1}.\end{aligned}$$

Therefore, $R_{\boldsymbol{k}\vee\boldsymbol{t}} \le R_{\boldsymbol{k}}$. Let $V(d) := \cup_{\{\boldsymbol{k}: r_{\boldsymbol{k}} = d, \boldsymbol{k}\in M_3\}} B(\boldsymbol{k})$. From Narisetty and He (2014),

$$\begin{aligned}\mathbb{P}\left[V(d)\right] &\le \mathbb{P}\left[\cup_{\{\boldsymbol{k}: r_{\boldsymbol{k}} = d, \boldsymbol{k}\in M_3\}} \left\{R_{\boldsymbol{t}} - R_{\boldsymbol{k}\vee\boldsymbol{t}} > 2\sigma^2(1+2s)(r_{\boldsymbol{k}} - r_{\boldsymbol{t}})\log p_n\right\}\right] \\ &\le c' p_n^{-w'd}.\end{aligned}$$

Then,

$$\mathbb{P}\left[\cup_{\{d > K|\boldsymbol{t}|\}} V(d)\right] \le p_n^{-w'K|\boldsymbol{t}|} \to 0, \text{ as } n \to \infty.$$

Restricting our attention to the high probability event $\cap_{\{d>r_{\boldsymbol{t}}\}}V(d)^c$,

$$
\begin{aligned}
&\sum_{\boldsymbol{k}\in M_3} PR(\boldsymbol{k},\boldsymbol{t}) \\
&\quad\preceq \sum_{\boldsymbol{k}\in M_3} \left(n\tau^2\lambda_{min}^n\right)^{-(r_{\boldsymbol{k}}-r_{\boldsymbol{t}\wedge\boldsymbol{k}})/2} \left(\frac{\lambda_{max}^n}{\lambda_{min}^n}\right)^{|\boldsymbol{t}|/2} \exp\left(-\frac{1}{2\sigma^2}\left(R_{\boldsymbol{k}}-R_{\boldsymbol{t}}\right)\right) \\
&\qquad\qquad \exp\left(a_n\sum_{i=1}^{p_n} g_i + \frac{b_n}{2}\boldsymbol{k}^\top J\boldsymbol{k} - a_n\sum_{i=1}^{p_n} t_i - \frac{b_n}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \\
&\quad\preceq \sum_{\boldsymbol{k}\in M_3} \left(p_n^{1+\delta}\vee\sqrt{n}\right)^{-(r_{\boldsymbol{k}}-r_{\boldsymbol{t}})} \left(\lambda_{min}^n\right)^{-|\boldsymbol{t}|/2} p_n^{(1+2s)(r_{\boldsymbol{k}}-r_{\boldsymbol{t}})} \\
&\qquad\qquad \exp\left(a_n\sum_{i=1}^{p_n} g_i + \frac{b_n}{2}\boldsymbol{k}^\top J\boldsymbol{k} - a_n\sum_{i=1}^{p_n} t_i - \frac{b_n}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \\
&\quad\preceq \sum_{\boldsymbol{k}\in M_3} \left(p_n^{\delta-2s}\vee\sqrt{n}p_n^{-1-2s}\right)^{-(r_{\boldsymbol{k}}-r_{\boldsymbol{t}})} \left(\lambda_{min}^n\right)^{-|\boldsymbol{t}|/2} \\
&\qquad\qquad \exp\left(a_n\sum_{i=1}^{p_n} g_i + \frac{b_n}{2}\boldsymbol{k}^\top J\boldsymbol{k} - a_n\sum_{i=1}^{p_n} t_i - \frac{b_n}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \\
&\quad\preceq \sum_{\boldsymbol{k}\in M_3} \left(p_n^{-3\delta/4}\vee\frac{p_n^{1+\delta/4}}{\sqrt{n}}\right)^{(r_{\boldsymbol{k}}-r_{\boldsymbol{t}})} p_n^{\kappa|\boldsymbol{t}|/2} \\
&\qquad\qquad \exp\left(a_n\sum_{i=1}^{p_n} g_i + \frac{b_n}{2}\boldsymbol{k}^\top J\boldsymbol{k} - a_n\sum_{i=1}^{p_n} t_i - \frac{b_n}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \\
&\quad\preceq \left(p_n^{-3\delta/4}\vee\frac{p_n^{1+\delta/4}}{\sqrt{n}}\right)^{(K-1)r_{\boldsymbol{t}}+1} p_n^{\kappa|\boldsymbol{t}|/2} \\
&\qquad\qquad \exp\left(-a_n\sum_{i=1}^{p_n} t_i - \frac{b_n}{2}\boldsymbol{t}^\top J\boldsymbol{t}\right) \sum_{\boldsymbol{k}\in M_3} \exp\left(a_n\sum_{i=1}^{p_n} g_i + \frac{b_n}{2}\boldsymbol{k}^\top J\boldsymbol{k}\right) \\
&\quad\preceq \left(p_n^{-3\delta/4}\vee\frac{p_n^{1+\delta/4}}{\sqrt{n}}\right)^{(K-1)|\boldsymbol{t}|/2} p_n^{\delta(K-1)|\boldsymbol{t}|/4} \exp\left(-a_n|\boldsymbol{t}|\right) \\
&\quad\sim \left(p_n^{-3\delta/4}\vee\frac{p_n^{1+\delta/4}}{\sqrt{n}}\right)^{(K-1)|\boldsymbol{t}|/2} \exp\left(nd_n|\boldsymbol{t}|\right) \\
&\quad\sim \exp\left(-nd_n\frac{3\delta(K-1)|\boldsymbol{t}|}{8}\right)\exp\left(nd_n|\boldsymbol{t}|\right) \\
&\quad= \exp\left(-nd_n|\boldsymbol{t}|\left(\frac{3\delta(K-1)}{8}-1\right)\right) \to 0.
\end{aligned}
$$

4. Models in M_4. If $c \in (0,1)$,

$$\begin{aligned}\mathbb{P}\left[\cup_{\{\boldsymbol{k}\in M_4\}}\{R_{\boldsymbol{k}} - R_{\boldsymbol{t}} < \Delta_n(1-c)\}\right] &\leq \mathbb{P}\left[\cup_{\{\boldsymbol{k}\in M_4\}}\{R_{\boldsymbol{k}} - R_{\boldsymbol{k}\vee\boldsymbol{t}} < \Delta_n(1-c)\}\right] \\ &\leq 2\exp\left(-c'\Delta_n\right) \to 0.\end{aligned}$$

Observe that:

$$\begin{aligned}&\exp\left(a_0\sum_{i=1}^{p_n} k_i + \frac{b_0}{2}\boldsymbol{k}^\top J_n \boldsymbol{k} - a_n\sum_{i=1}^{p_n} t_i - \frac{b_0}{2}\boldsymbol{t}^\top J_n \boldsymbol{t}\right) \\ &= \exp\left(a_0\sum_{i=1}^{p_n} k_i^* + \frac{b_0}{2}\left(\boldsymbol{k}^*\right)^\top J_n\left(\boldsymbol{k}^*\right) - a_0\sum_{i=1}^{p_n} t_i^* - \frac{b_0}{2}\left(\boldsymbol{t}^*\right)^\top J_n\left(\boldsymbol{t}^*\right)\right) \\ &\times \exp\left(-b_0\boldsymbol{k}^\top J_n\mathbf{1} + b_0\boldsymbol{t}^\top J_n\mathbf{1}\right),\end{aligned}$$

where $\boldsymbol{k}^* = \boldsymbol{k} + \mathbf{1}$ and $\boldsymbol{t}^* = \boldsymbol{t} + \mathbf{1}$. By restricting to the high probability event $C_n :=$

$\{R_{\boldsymbol{k}} - R_{\boldsymbol{t}} \geq \Delta_n(1-c), \forall \boldsymbol{k} \in M_4\}$,

$$
\begin{aligned}
&\sum_{\boldsymbol{k}\in M_4} PR(\boldsymbol{k},\boldsymbol{t}) \\
&\quad \preceq \sum_{\boldsymbol{k}\in M_4} \left(n\tau_n^2\lambda_{min}^n\right)^{|\boldsymbol{t}|/2} \left(\frac{\lambda_{max}^n}{\lambda_{min}^n}\right)^{|\boldsymbol{t}|/2} \exp\left(-\frac{1}{2\sigma^2}\left(R_{\boldsymbol{k}} - R_{\boldsymbol{t}}\right)\right) \\
&\qquad \exp\left(a_n \sum_{i=1}^{p_n} k_i^* + \frac{b_n}{2}\left(\boldsymbol{k}^*\right)^\top J_n\left(\boldsymbol{k}^*\right) - a_n \sum_{i=1}^{p_n} t_i^* - \frac{b_n}{2}\left(\boldsymbol{t}^*\right)^\top J_n\left(\boldsymbol{t}^*\right)\right) \\
&\qquad \exp\left(-b_n \boldsymbol{k}^\top J_n \mathbf{1} + b_n \boldsymbol{t}^\top J_n \mathbf{1}\right) \\
&\quad \preceq \sum_{\boldsymbol{k}\in M_4} \left(n\tau_n^2\lambda_{min}^n\right)^{|\boldsymbol{t}|/2} \left(\lambda_{min}^n\right)^{-|\boldsymbol{t}|/2} \exp\left(-\Delta_n(1-c)/2\sigma^2\right) \\
&\qquad \exp\left(a_n \sum_{i=1}^{p_n} k_i^* + \frac{b_n}{2}\left(\boldsymbol{k}^*\right)^\top J_n\left(\boldsymbol{k}^*\right) - a_n \sum_{i=1}^{p_n} t_i^* - \frac{b_n}{2}\left(\boldsymbol{t}^*\right)^\top J_n\left(\boldsymbol{t}^*\right)\right) \\
&\quad \preceq \left(p_n^{2+3\delta} \vee n\right)^{|\boldsymbol{t}|/2} p_n^{\delta|\boldsymbol{t}|/2} \exp\left(-\Delta_n(1-c)/2\sigma^2\right) \\
&\qquad \exp\left(-a_n \sum_{i=1}^{p_n} t_i^* - \frac{b_n}{2}\left(\boldsymbol{t}^*\right)^\top J_n\left(\boldsymbol{t}^*\right)\right) \sum_{\boldsymbol{k}\in M_4} \exp\left(a_n \sum_{i=1}^{p_n} k_i^* + \frac{b_n}{2}\left(\boldsymbol{k}^*\right)^\top J_n\left(\boldsymbol{k}^*\right)\right) \\
&\quad \sim \exp\left(-\frac{1}{2\sigma^2}\left(\Delta_n(1-c) - \sigma^2|\boldsymbol{t}|\log\left(p_n^{2+3\delta} \vee n\right) - \sigma^2|\boldsymbol{t}|\delta\log p_n + 2\sigma^2 a_n \sum_{i=1}^{p_n} t_i^*\right)\right) \\
&\quad \sim \exp\left(-\frac{1}{2\sigma^2}\left(\Delta_n(1-c) - \sigma^2|\boldsymbol{t}|\log\left(p_n^{2+3\delta} \vee n\right) - \sigma^2|\boldsymbol{t}|(4+\delta)\log p_n\right)\right) \\
&\quad \preceq \exp\left(-\frac{1}{2\sigma^2}\left(\Delta_n(1-c) - \sigma^2|\boldsymbol{t}|(6+4\delta)\log p_n\right)\right) \\
&\quad \preceq \exp\left(-\frac{1}{2\sigma^2}\left(\Delta_n(1-c) - \sigma^2|\boldsymbol{t}|(4+4\delta)\log p_n\right)\right) \\
&\quad \sim \exp\left(-\frac{1}{2\sigma^2}\left(\Delta_n(1-c) - w'\gamma_n\right)\right) \to 0,
\end{aligned}
$$

where $w' \in (0,1)$ and $c < 1 - w'$.

3.4.2 True posterior consistency with pseudo Ising prior

In our variable selection algorithm, we used pseudo Ising prior on $\boldsymbol{\gamma}$ instead of the ture Ising prior. Replacing the true Ising prior with the Ising prior we used, the posterior ratio is:

$$\tilde{PR}(\boldsymbol{g},\boldsymbol{t}) = \frac{\tilde{\pi}(\boldsymbol{\gamma}=\boldsymbol{g} \mid X,\boldsymbol{y},\sigma^2)}{\tilde{\pi}(\boldsymbol{\gamma}=\boldsymbol{t} \mid X,\boldsymbol{y},\sigma^2)} \tag{3.12}$$

$$= \underbrace{\frac{Q_{\boldsymbol{g}}}{Q_{\boldsymbol{t}}}}_{\text{(a)}} \underbrace{\exp\left(-\frac{1}{2\sigma^2}R_{\boldsymbol{g}} + \frac{1}{2\sigma^2}R_{\boldsymbol{t}}\right)}_{\text{(b)}}$$

$$\times \exp\Bigg(a_0\mathbf{1}^\top\boldsymbol{g} + b_0\boldsymbol{g}^\top J_n\boldsymbol{g} - \mathbf{1}^\top\log\cosh\left(b_0J_n\boldsymbol{g} + a_0\mathbf{1}\right) \tag{3.13}$$

$$- a_0\mathbf{1}^\top\boldsymbol{t} - b_{0n}\boldsymbol{t}^\top J_n\boldsymbol{t} + \mathbf{1}^\top\log\cosh\left(b_0J_n\boldsymbol{t} + a_0\mathbf{1}\right)\Bigg). \tag{3.14}$$

Theorem 3 indicates the posterior above goes to zero for all $\boldsymbol{g} \neq \boldsymbol{t}$:

Theorem 3. *Under the Conditions 1 - 5, the posterior probability of the true model with pseudo Ising prior goes to 1, that is,*

$$\sum_{\boldsymbol{g}\neq\boldsymbol{t}} \tilde{PR}(\boldsymbol{g},\boldsymbol{t}) \to 0,$$

as n goes to infinity.

Note that the terms (a) and (b) in (3.12) are the same as in (3.8). Therefore, we need to be albe to control the last exponential term. Observe that:

$$\mathbf{1}^\top\log\cosh\left(b_0J_n\boldsymbol{t} + a_0\mathbf{1}\right) - \mathbf{1}^\top\log\cosh\left(b_0J_n\boldsymbol{g} + a_0\mathbf{1}\right) \to 0.$$

Above convergence is due to Condition 2. It allows us to use the same arguments to prove Theorem 3.

3.4.3 Bounded KL divergence

In this subsection, we provide an upper bound of KL divergence between $\tilde{q}(\boldsymbol{\beta},\boldsymbol{\gamma})$ and $\tilde{\pi}\left(\boldsymbol{\beta},\boldsymbol{\gamma} \mid \boldsymbol{y}\right)$ with appropriate choices of variational parameters, where $\tilde{q}(\boldsymbol{\beta},\boldsymbol{\gamma})$ is a variational distribution with pseudo-Ising on $\boldsymbol{\gamma}$ and $\tilde{\pi}\left(\boldsymbol{\beta},\boldsymbol{\gamma} \mid \boldsymbol{y}\right)$ is the true posterior with pseudo-Ising

on $\boldsymbol{\gamma}$:

$$\begin{aligned} & KL\left(\tilde{q}(\boldsymbol{\beta},\boldsymbol{\gamma}) \mid\mid \tilde{\pi}\left(\boldsymbol{\beta},\boldsymbol{\gamma} \mid \boldsymbol{y}\right)\right) \\ & = KL\left(\tilde{q}(\boldsymbol{\gamma}) \mid\mid \tilde{p}\left(\boldsymbol{\gamma}\right)\right) + \mathbb{E}_{\tilde{q}(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[\log q(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) - \log p(\boldsymbol{\beta} \mid \boldsymbol{\gamma})\right] \\ & - \mathbb{E}_{\tilde{q}(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[\log \frac{p(\boldsymbol{y} \mid \boldsymbol{\beta})}{p(\boldsymbol{y} \mid \boldsymbol{\beta_t})}\right] + C. \end{aligned} \tag{3.15}$$

Theorem 4. *There exists a variational distribution in the variational family* (3.3) *which satisfies*

$$KL\left(\tilde{q}(\boldsymbol{\beta},\boldsymbol{\gamma}) \mid\mid \tilde{\pi}\left(\boldsymbol{\beta},\boldsymbol{\gamma} \mid \boldsymbol{y}\right)\right) = o(n).$$

To prove Theorem 4, observe the first term in (3.15):

$$\begin{aligned} & KL\left(\tilde{q}(\boldsymbol{\gamma}) \mid\mid \tilde{p}\left(\boldsymbol{\gamma}\right)\right) \\ & = \sum_{\boldsymbol{\gamma}\in\{-1,1\}^{p_n}} \tilde{q}(\boldsymbol{\gamma})\left(\log \tilde{q}(\boldsymbol{\gamma}) - \log \tilde{p}(\boldsymbol{\gamma})\right) \\ & = \sum_{\boldsymbol{\gamma}\in\{-1,1\}^{p_n}} \tilde{q}(\boldsymbol{\gamma})\left(\sum_{i=1}^{p_n}\left(a_i - a_0\right)\gamma_i\right) \\ & + \sum_{\boldsymbol{\gamma}\in\{-1,1\}^{p_n}} \tilde{q}(\boldsymbol{\gamma})\left(\sum_{i=1}^{p_n}\left(\log\cosh\left(b_0 m_i(\boldsymbol{\gamma}) + a_0\right) - \log\cosh\left(b m_i(\boldsymbol{\gamma}) + a_i\right)\right)\right) \\ & + \sum_{\boldsymbol{\gamma}\in\{-1,1\}^{p_n}} \tilde{q}(\boldsymbol{\gamma})\left(b\boldsymbol{\gamma}^\top J_n \boldsymbol{\gamma} - b_0 \boldsymbol{\gamma}^\top J_n \boldsymbol{\gamma}\right). \end{aligned} \tag{3.16}$$

We consider variational parameters below:

$$b = \frac{1}{p_n} \text{ and } a_i = \begin{cases} nd_n \text{ if } i \in \mathcal{A}, \\ -nd_n \text{ if } i \in \mathcal{A}^c \end{cases}. \tag{3.17}$$

Lemma 17. *For any $\boldsymbol{g} \neq \boldsymbol{t}$, with the choices of variational parameters in* (3.17)*,*

$$\tilde{q}(\boldsymbol{\gamma} = \boldsymbol{g}) \to 0.$$

Proof. Since $b = \frac{1}{p_n}$, $b\boldsymbol{g}^\top J_n \boldsymbol{g}$ is bounded and $bm_i(\boldsymbol{g})$ is decreasing to zero at a rate $\frac{1}{p_n}$. Therefore,

$$\begin{aligned}
\tilde{q}(\boldsymbol{\gamma} = \boldsymbol{g}) &= 2^{-p_n} \exp\left(b\boldsymbol{g}^\top J_n \boldsymbol{g} + \sum_{i=1}^{p_n} \left(a_i g_i - \log\cosh\left(bm_i(\boldsymbol{g}) + a_i\right)\right)\right) \\
&= 2^{-p_n} C \exp\left(\sum_{i=1}^{p_n} \left(a_i g_i - \log\cosh\left(bm_i(\boldsymbol{g}) + a_i\right)\right)\right) \\
&\simeq 2^{-p_n} C \exp\left(\sum_{i=1}^{p_n} \left(a_i g_i - \log\cosh\left(a_i\right)\right)\right) \\
&= 2^{-p_n} C \exp\left(\sum_{i\in\mathcal{A}} a_i g_i + \sum_{i\in\mathcal{A}^c} a_i g_i - \sum_{i=1}^{p_n} \log\cosh\left(nd_n\right)\right) \\
&= 2^{-p_n} C \exp\left(\sum_{i\in\mathcal{A}} a_i g_i + \sum_{i\in\mathcal{A}^c} a_i g_i - \sum_{i=1}^{p_n} \log\cosh\left(nd_n\right)\right) \\
&= 2^{-p_n} C \exp\left(\sum_{i\in\mathcal{A}} nd_n g_i - \sum_{i\in\mathcal{A}^c} nd_n g_i - \sum_{i=1}^{p_n} \log\cosh\left(nd_n\right)\right).
\end{aligned}$$

Observe that $\sum_{i\in\mathcal{A}} nd_n g_i - \sum_{i\in\mathcal{A}^c} nd_n g_i = nd_n(|\mathcal{B}| - |\mathcal{B}^c|)$ where $\mathcal{B} = \{i : g_i = t_i\}$. When $|\mathcal{B}| - |\mathcal{B}^c| < 0$, it is easy to show (3.18) $\to 0$. Provided $\boldsymbol{g} \neq \boldsymbol{t}$, the upper bound of $|\mathcal{B}| - |\mathcal{B}^c|$ is $p_n - 2$ such that:

$$\begin{aligned}
&2^{-p_n} C \exp\left(\sum_{i\in\mathcal{A}} nd_n g_i - \sum_{i\in\mathcal{A}^c} nd_n g_i - \sum_{i=1}^{p_n} \log\cosh\left(nd_n\right)\right) \qquad (3.18) \\
&\leq 2^{-p_n} C \exp\left(nd_n(p_n - 2) - p_n \log\cosh\left(nd_n\right)\right) \\
&= 2^{-p_n} C \exp\left(p_n\left(nd_n - \log\cosh\left(nd_n\right)\right) - 2nd_n\right) \\
&= C\left(\frac{e^{nd_n - \log\cosh(nd_n)}}{2}\right)^{p_n} e^{-2nd_n} \\
&\simeq C\left(\frac{e^{\log 2}}{2}\right)^{p_n} e^{-2nd_n} \to 0.
\end{aligned}$$

To derive above convergence, we use the fact $u - \log\cosh(u) \simeq \log 2$ for large u. □

Using the Lemma 17,

$$\begin{aligned}
&\sum_{\boldsymbol{\gamma}\in\{-1,1\}^{p_n}} \tilde{q}(\boldsymbol{\gamma}) \left(\sum_{i=1}^{p_n} (a_i - a_0)\, \gamma_i \right) \\
&= \sum_{\boldsymbol{\gamma}\in\{-1,1\}^{p_n}} \tilde{q}(\boldsymbol{\gamma}) \left(\sum_{i\in\mathcal{A}} (a_i - a_{0n})\, \gamma_i + \sum_{i\in\mathcal{A}^c} (a_i - a_{0n})\, \gamma_i \right) \\
&= \sum_{\boldsymbol{\gamma}\in\{-1,1\}^{p_n}} \tilde{q}(\boldsymbol{\gamma}) \left(\sum_{i\in\mathcal{A}} (2nd_n)\, \gamma_i + \sum_{i\in\mathcal{A}^c} (0)\, \gamma_i \right) \\
&\leq (2nd_n|\mathcal{A}|) \left(\tilde{q}(\boldsymbol{t}) + \sum_{\boldsymbol{\gamma}\neq\boldsymbol{t}} \tilde{q}(\boldsymbol{\gamma}) \right) = o(n).
\end{aligned}$$

For the second term in (3.16), note that

$$\begin{aligned}
&\sum_{i=1}^{p_n} \left(\log\cosh\left(b_0 m_i(\boldsymbol{\gamma}) + a_0\right) - \log\cosh\left(b m_i(\boldsymbol{\gamma}) + a_i\right) \right) \\
&\quad \sim \sum_{i=1}^{p_n} \log \frac{\exp\left(\frac{1}{p_n} + a_0\right) + \exp\left(\frac{1}{p_n} - a_0\right)}{\exp\left(\frac{1}{p_n} + a_i\right) + \exp\left(\frac{1}{p_n} - a_i\right)}.
\end{aligned}$$

Combining the results above, we have:

$$KL\left(\tilde{q}(\boldsymbol{\gamma}) \;||\; \tilde{p}\left(\boldsymbol{\gamma}\right)\right) = o(n).$$

Next, to control the other terms, we construct the variational parameters as follows:

$$\sigma_i^2 = \begin{cases} 0, \text{ if } g_i = -1, \\ 1/p_n, \text{ if } g_i = 1, \end{cases} \tag{3.19}$$

and

$$\mu_i = \begin{cases} = 0, \text{ if } g_i = -1, \\ = \beta_{\boldsymbol{t},i}, \text{ if } g_i = 1 \text{ and } t_i = 1, \\ = 1/(n^{1/2} p_n), \text{ if } g_i = 1 \text{ and } t_i = -1. \end{cases} \tag{3.20}$$

Observe the second term in (3.15):

$$\begin{aligned}
&\mathbb{E}_{\tilde{q}(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[\log q(\boldsymbol{\beta}\mid\boldsymbol{\gamma})-\log p(\boldsymbol{\beta}\mid\boldsymbol{\gamma})\right]\\
&=\sum_{\boldsymbol{\gamma}\in\{-1,1\}^{p_n}}\tilde{q}(\boldsymbol{\gamma})\mathbb{E}_{q(\boldsymbol{\beta}|\boldsymbol{\gamma})}\left[\log q(\boldsymbol{\beta}\mid\boldsymbol{\gamma})-\log p(\boldsymbol{\beta}\mid\boldsymbol{\gamma})\right]\\
&=\frac{1}{2}\sum_{\boldsymbol{g}\in\{-1,1\}^{p_n}}\tilde{q}(\boldsymbol{\gamma}=\boldsymbol{g})\left(tr\left(\Sigma_{\boldsymbol{g}}\right)-|\boldsymbol{g}|+\boldsymbol{\mu}_{\boldsymbol{g}}^{\top}\boldsymbol{\mu}_{\boldsymbol{g}}-\log|\Sigma_{\boldsymbol{g}}|\right),
\end{aligned}\tag{3.21}$$

where $\boldsymbol{\mu}_{\boldsymbol{g}}$ is a vector of μ_is in (3.20), $\boldsymbol{\sigma}_{\boldsymbol{g}}$ is a vector of σ_is in (3.19), and $\Sigma_{\boldsymbol{g}}=diag\left(\boldsymbol{\sigma}_{\boldsymbol{g}}^2\right)$. With the variational parameters in (3.19) and (3.20), we can bound (3.21) by $o(n)$ using the Lemma 17. For the third term in (3.15), let $L_0\sim\mathcal{N}\left(X\boldsymbol{\beta}_{\boldsymbol{t}},\sigma^2I\right)$ and $L_{\boldsymbol{\beta}}\sim\mathcal{N}\left(X\boldsymbol{\beta},\sigma^2I\right)$. Then,

$$KL\left(L_0\mid\mid L_{\boldsymbol{\beta}}\right)=\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)^{\top}X^{\top}X\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{\boldsymbol{t}}\right),$$

$$\begin{aligned}
&\mathbb{E}_{\tilde{q}(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[KL\left(L_0\mid\mid L_{\boldsymbol{\beta}}\right)\right]\\
&=\sum_{\boldsymbol{\gamma}\in\{-1,1\}^{p_n}}\int\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)^{\top}X^{\top}X\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)\tilde{q}(\boldsymbol{\beta},\boldsymbol{\gamma})d\boldsymbol{\beta}\\
&=\sum_{\boldsymbol{\gamma}\in\{-1,1\}^{p_n}}\int\left\|X\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)\right\|_2^2\tilde{q}(\boldsymbol{\beta},\boldsymbol{\gamma})d\boldsymbol{\beta}\\
&=\sum_{\boldsymbol{g}\in\{-1,1\}^{p_n}}\tilde{q}(\boldsymbol{\gamma}=\boldsymbol{g})\mathbb{E}_{\tilde{q}(\boldsymbol{\beta}|\boldsymbol{\gamma}=\boldsymbol{g})}\left[\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)^{\top}X^{\top}X\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)\right]\\
&=\sum_{\boldsymbol{g}\in\{-1,1\}^{p_n}}\tilde{q}(\boldsymbol{\gamma}=\boldsymbol{g})\left(\left(\boldsymbol{\mu}_{\boldsymbol{g}}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)^{\top}X^{\top}X\left(\boldsymbol{\mu}_{\boldsymbol{g}}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)+\mathrm{tr}\left(X^{\top}X\Sigma_{\boldsymbol{g}}\right)\right).
\end{aligned}$$

Note,

$$\begin{aligned}
&\sum_{\boldsymbol{g}\in\{-1,1\}^{p_n}}\tilde{q}(\boldsymbol{\gamma}=\boldsymbol{g})\left(\boldsymbol{\mu}_{\boldsymbol{g}}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)^{\top}X^{\top}X\left(\boldsymbol{\mu}_{\boldsymbol{g}}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)\\
&=\tilde{q}(\boldsymbol{\gamma}=\boldsymbol{t})\left(\boldsymbol{\mu}_{\boldsymbol{t}}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)^{\top}X^{\top}X\left(\boldsymbol{\mu}_{\boldsymbol{t}}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)+\sum_{\boldsymbol{g}\neq\boldsymbol{t}}\tilde{q}(\boldsymbol{\gamma}=\boldsymbol{g})\left(\boldsymbol{\mu}_{\boldsymbol{g}}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)^{\top}X^{\top}X\left(\boldsymbol{\mu}_{\boldsymbol{g}}-\boldsymbol{\beta}_{\boldsymbol{t}}\right).
\end{aligned}$$

Let $\boldsymbol{w}:=\boldsymbol{\mu}_{\boldsymbol{g}}-\boldsymbol{\beta}_{\boldsymbol{t}}$. Then,

$$X\left(\boldsymbol{\mu}_{\boldsymbol{g}}-\boldsymbol{\beta}_{\boldsymbol{t}}\right)=X\boldsymbol{w}=\sum_{i=1}^{p_n}w_i\boldsymbol{x}_i,$$

where $\boldsymbol{x}_i$ is the i-th column of X.

$$\begin{aligned}\sum_{i=1}^{p_n} w_i \boldsymbol{x}_i &= \sum_{\{i:g_i=-1\}} w_i \boldsymbol{x}_i + \sum_{\{i:g_i=1 \text{ and } t_i=1\}} w_i \boldsymbol{x}_i + \sum_{\{i:g_i=1 \text{ and } t_i=-1\}} w_i \boldsymbol{x}_i \\ &= C + \sum_{\{i:g_i=1 \text{ and } t_i=-1\}} w_i \boldsymbol{x}_i.\end{aligned}$$

The order of the above summation is $o(n)$ due to (3.20).

3.4.4 Variational posterior consistency

In this subsection, we investigate a couple of ingredients for establishing variational posterior consistency, which is our ultimate goal.

Theorem 5. *Let q^* be the variational posterior obtained under the prior* (3.2) *and variational family* (3.3)*. Then for any $\boldsymbol{g} \neq \boldsymbol{t}$, we have*

$$q^*(\boldsymbol{\gamma} = \boldsymbol{g}) \to 0.$$

as n goes to infinity.

First, we consider:

$$\begin{aligned}Z^{\tilde{q}} := Z_n^{\tilde{q}}(\boldsymbol{a}, b) &= \sum_{\boldsymbol{\gamma} \in \{-1,1\}^{p_n}} \tilde{q}(\boldsymbol{\gamma}), \\ Z^{\tilde{p}} := Z_n^{\tilde{p}}(a_{0n}, b_{0n}) &= \sum_{\boldsymbol{\gamma} \in \{-1,1\}^{p_n}} \tilde{p}(\boldsymbol{\gamma}).\end{aligned}$$

$Z^{\tilde{q}}$ and $Z^{\tilde{p}}$ are the normalizing constants of $\tilde{q}(\boldsymbol{\gamma})$ and $\tilde{p}(\boldsymbol{\gamma})$ respectively. We define two valid probability mass functions $q_z(\boldsymbol{\gamma}) = \frac{1}{Z_n^{\tilde{q}}} \tilde{q}(\boldsymbol{\gamma})$ and $p_z(\boldsymbol{\gamma}) = \frac{1}{Z_n^{\tilde{p}}} \tilde{p}(\boldsymbol{\gamma})$. Then, the posterior distribution with $p_z(\boldsymbol{\gamma})$ is:

$$\begin{aligned}\pi_Z(\boldsymbol{\gamma}, \boldsymbol{\beta} \mid \boldsymbol{y}) &= \frac{\left(Z^{\tilde{p}}\right)^{-1} \tilde{p}(\boldsymbol{\gamma}) p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) L(\boldsymbol{\beta})}{\sum_{\boldsymbol{\gamma}} \int_{\boldsymbol{\beta}} \left(Z^{\tilde{p}}\right)^{-1} \tilde{p}(\boldsymbol{\gamma}) p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) L(\boldsymbol{\beta}) d\boldsymbol{\beta}} \\ &= \frac{\tilde{p}(\boldsymbol{\gamma}) p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) L(\boldsymbol{\beta})}{\sum_{\boldsymbol{\gamma}} \int_{\boldsymbol{\beta}} \tilde{p}(\boldsymbol{\gamma}) p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) L(\boldsymbol{\beta}) d\boldsymbol{\beta}} \\ &= \tilde{\pi}\left(\boldsymbol{\gamma}, \boldsymbol{\beta} \mid \boldsymbol{y}\right).\end{aligned}$$

From the relation above, we have:

$$
\begin{aligned}
&KL\left(q_Z(\boldsymbol{\gamma},\boldsymbol{\beta}) \mid\mid \pi_Z(\boldsymbol{\gamma},\boldsymbol{\beta} \mid \boldsymbol{y})\right) \\
&= \sum_{\boldsymbol{\gamma}\{-1,1\}^{p_n}} q_Z(\boldsymbol{\gamma}) \int_{\boldsymbol{\beta}} q(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) \left(\log q_Z(\boldsymbol{\gamma}) + \log q(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) - \log \pi_Z(\boldsymbol{\gamma},\boldsymbol{\beta} \mid \boldsymbol{y})\right) d\boldsymbol{\beta} \\
&= \sum_{\boldsymbol{\gamma}\{-1,1\}^{p_n}} \frac{1}{Z^{\tilde{q}}} \tilde{q}(\boldsymbol{\gamma}) \int_{\boldsymbol{\beta}} q(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) \left(\log \tilde{q}(\boldsymbol{\gamma}) - \log Z^{\tilde{q}} + \log q(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) - \log \tilde{\pi}\left(\boldsymbol{\gamma},\boldsymbol{\beta} \mid \boldsymbol{y}\right)\right) d\boldsymbol{\beta} \\
&= \frac{KL\left(\tilde{q}(\boldsymbol{\gamma},\boldsymbol{\beta}) \mid\mid \tilde{\pi}(\boldsymbol{\gamma},\boldsymbol{\beta} \mid \boldsymbol{y})\right)}{Z^{\tilde{q}}} - \log Z^{\tilde{q}}.
\end{aligned}
$$

Next, using Corollary 4.15 in Boucheron et al. (2013), we have

$$
\begin{aligned}
KL\left(q_Z(\boldsymbol{\gamma},\boldsymbol{\beta}) \mid\mid \pi_Z(\boldsymbol{\gamma},\boldsymbol{\beta} \mid \boldsymbol{y})\right) &\geq \int f dQ_Z - \log \int e^f d\Pi_Z\left(\mid \boldsymbol{y}\right) \\
&= \frac{1}{Z^{\tilde{q}}} \int f d\tilde{Q} - \log \int e^f d\tilde{\Pi}\left(\mid \boldsymbol{y}\right),
\end{aligned}
$$

where f is any function. Reorganizing the relation above, we have

$$
\int f d\tilde{Q} \leq KL\left(\tilde{q}(\boldsymbol{\gamma},\boldsymbol{\beta}) \mid\mid \tilde{\pi}(\boldsymbol{\gamma},\boldsymbol{\beta} \mid \boldsymbol{y})\right) - Z^{\tilde{q}} \log Z^{\tilde{q}} + Z^{\tilde{q}} \log \int e^f d\tilde{\Pi}\left(\mid \boldsymbol{y}\right). \tag{3.22}
$$

We need show that a lower bound of $Z^{\tilde{q}}$ is one to use the relation (3.22).

Proposition 1. *Let $\tilde{q}(\boldsymbol{\gamma})$ be a pseudo likelihood of Ising model characterized by parameters $\boldsymbol{a} = (a_1, \ldots, a_{p_n})$ and b. Then,*

$$
\sum_{\boldsymbol{\gamma} \in \{-1,1\}^{p_n}} \tilde{q}(\boldsymbol{\gamma}) \geq 1,
$$

for any $\boldsymbol{a}$ and b.

We provide numerical evidences to Proposition 1 (See Table 3.2).

$Z^{\tilde{q}}(\boldsymbol{a}, b)$	Parameters	
	$\boldsymbol{a} = 0.2 \cdot \mathbf{1} \quad b = 0.01$	$\boldsymbol{a} = -0.2 \cdot \mathbf{1} \quad b = 0.01$
$p_n = 4$	1.0004	1.0004
$p_n = 9$	1.0011	1.0011
$p_n = 16$	1.0022	1.0022
	$\boldsymbol{a} = \mathbf{1} \qquad b = 0.1$	$\boldsymbol{a} = -\mathbf{1} \qquad b = 0.1$
$p_n = 4$	1.006	1.006
$p_n = 9$	1.014	1.014
$p_n = 16$	1.025	1.025
	$\boldsymbol{a} = 0.2 \cdot \mathbf{1} \quad b = 0.5$	$\boldsymbol{a} = -0.2 \cdot \mathbf{1} \quad b = 0.5$
$p_n = 4$	1.616	1.616
$p_n = 9$	2.879	2.879
$p_n = 16$	6.941	6.941
	$\boldsymbol{a} = 0.5 \cdot \mathbf{1} \quad b = 0.5$	$\boldsymbol{a} = -0.3 \cdot \mathbf{1} \quad b = 0.5$
$p_n = 4$	1.417	1.562
$p_n = 9$	2.048	2.635
$p_n = 16$	3.603	5.879

Table 3.2: Exact normalizing constants with varying $\boldsymbol{a}$ and b.

With $f = 1$ in (3.22), we get:

$$Z^{\tilde{q}} \log Z^{\tilde{q}} \leq KL\left(\tilde{q}(\boldsymbol{\gamma}, \boldsymbol{\beta}) \mid\mid \tilde{\pi}(\boldsymbol{\gamma}, \boldsymbol{\beta} \mid \boldsymbol{y})\right) \tag{3.23}$$

With $f = n\mathbb{1}(\boldsymbol{\gamma} \neq \boldsymbol{t})$ in (3.22) and (3.23), we complete the Theorem 5 under the Proposition 1.

$$\begin{aligned}
n\mathbb{1}(\boldsymbol{\gamma} \neq \boldsymbol{t}) &\leq KL\left(\tilde{q}(\boldsymbol{\gamma}, \boldsymbol{\beta}) \mid\mid \tilde{\pi}(\boldsymbol{\gamma}, \boldsymbol{\beta} \mid \boldsymbol{y})\right) - Z^{\tilde{q}} \log Z^{\tilde{q}} + Z^{\tilde{q}} \log \int e^{n\mathbb{1}(\boldsymbol{\gamma} \neq \boldsymbol{t})} d\tilde{\Pi}\left(\mid \boldsymbol{y}\right) \\
&\leq KL\left(\tilde{q}(\boldsymbol{\gamma}, \boldsymbol{\beta}) \mid\mid \tilde{\pi}(\boldsymbol{\gamma}, \boldsymbol{\beta} \mid \boldsymbol{y})\right) + Z^{\tilde{q}} \log \int e^{n\mathbb{1}(\boldsymbol{\gamma} \neq \boldsymbol{t})} d\tilde{\Pi}\left(\mid \boldsymbol{y}\right) \\
&\leq o(n) + Z^{\tilde{q}} \log\left(1 + e^{n} \tilde{\pi}\left(\boldsymbol{\gamma} \neq \boldsymbol{t} \mid \boldsymbol{y}\right)\right) \\
&\implies \mathbb{1}(\boldsymbol{\gamma} \neq \boldsymbol{t}) \to 0.
\end{aligned}$$

CHAPTER 4

DISCUSSION AND FUTURE RESEARCH

4.1 Conclusion

Modeling discrete data is a basic problem in statistics and machine learning. Discrete data are rarely independent and a fundamental modeling task is to model the dependencies among variables. A graphical model is a flexible tool available for modeling such dependent discrete data. In this dissertation, we focused on a well known graphical model, Ising model. We provided a procesure for Ising model parameter estimation using variational Bayes methods. In order to tackle the issue of the intractable normalizing constant, we employed a pseudo-likelihood and placed it wherever the true likelihood is needed. We suggested two variational family choices and developed variational Bayes algorithms for each family of distributions under the pseudo-likelihood. In a variety of numerical studies, we compared our VB methods and two other existing methods, PMLE and a MCMC based method. Notably, the simulation results demonstrated the superiority of our VB methods in terms of accuracy and computational costs. In addition, we found that a two-parameter Ising model is suitable for characterizing a network data. Using the Facebook example, we applied the estimation procedures to characterize an overall strength of interaction and an external influence of the network. This thesis also provides a theoretical justification to the VB algorithm under mean-field family. We showed that the variational posterior is consistent as the data size increasing under three mild conditions on the coupling matrix A_n. Specifically, we established that the variational posterior concentrates around shrinking neighborhoods of the true parameter and we next establish the rate of contraction for the variational posterior with a suitable bound on the Kulback-Leibler divergence between the variational and the true posterior.

In addition to the Ising model parameter estimation, we developed a variable selection

technique in a high dimensional setup when the feature space is structurally dependent. We capture the structural dependencies using an pseudo-Ising prior and a pseudo-Ising variational distribution on latent binary variables with multiple threshold parameters in the variational family, which enable us to perform variable selection. Providing numerical experiments and some theoretical results, we validated the efficacy of the variable selection VB algorithm.

4.2 Directions for future research

Multiple observations from an Ising model: While we considered only one observation of $\boldsymbol{x}$ is available for the inference on two-parameter Ising model, another group of previous researches assumed that multiple observations of $\boldsymbol{x}$ are available. With the i.i.d copies of $\boldsymbol{x}$, we will be able to develop a procedure to estimate the structure of the underlying graph, i.e., we will be able to estimate all the edges in the graph. Besides, the multiple observations will enable inferences on multi-parameter Ising model beyond only one interaction parameter and only one threshold parameter.

Multivariate version of an Ising model: The Potts model is a versatile graphical model for discrete data that naturally extends from the Ising model. Consider an undirected graph. Without restricting to binary variables, each node of the graph represents a categorical variable such as blood types, hair colors, education levels, etc. Variational Bayes would be a useful tool for inference on a Potts model and its applications.

Multiresponse regression: If response variables in a regression model is multivariate, we need to consider the additional dependencies among the response variables. A researcher would use a graphical model to capture different types of dependencies in various ways. Also, to approximate more complex posterior distribution, a variational Bayes method will definitely help.

BIBLIOGRAPHY

BIBLIOGRAPHY

Anandkumar, A., Tan, V. Y., Huang, F., Willsky, A. S., et al. (2012). High-dimensional structure estimation in ising models: Local separation criterion. *The Annals of Statistics*, 40(3):1346–1375.

Andersen, M. R., Vehtari, A., Winther, O., and Hansen, L. K. (2017). Bayesian inference for spatio-temporal spike-and-slab priors. *The Journal of Machine Learning Research*, 18(1):5076–5133.

Andersen, M. R., Winther, O., and Hansen, L. K. (2014). Bayesian inference for structured spike and slab priors. *Advances in Neural Information Processing Systems*, 27.

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

Basak, A. and Mukherjee, S. (2017). Universality of the mean-field for the potts model. *Probability Theory and Related Fields*, 168(3):557–600.

Bhattacharya, B. B., Mukherjee, S., et al. (2018). Inference in ising models. *Bernoulli*, 24(1):493–525.

Bhattacharya, S. and Maiti, T. (2021). Statistical foundation of variational bayes neural networks. *Neural Networks*, 137:151–173.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bresler, G. (2015). Efficiently learning ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 771–782.

Brush, S. G. (1967). History of the lenz-ising model. *Reviews of modern physics*, 39(4):883.

Cao, X., Khare, K., and Ghosh, M. (2019). Posterior graph selection and estimation consistency for high-dimensional bayesian dag models. *The Annals of Statistics*, 47(1):319–348.

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis*, 7(1):73–108.

Castillo, I. and Roquain, É. (2020). On spike and slab empirical bayes multiple testing. *The Annals of Statistics*, 48(5):2548–2574.

Chang, C., Kundu, S., and Long, Q. (2018). Scalable bayesian variable selection for structured high-dimensional data. *Biometrics*, 74(4):1372–1382.

Chatterjee, S. and Dembo, A. (2016). Nonlinear large deviations. *Advances in Mathematics*, 299:396–450.

Chatterjee, S. et al. (2007). Estimation in spin glasses: A first step. *The Annals of Statistics*, 35(5):1931–1946.

Comets, F. (1992). On consistency of a class of estimators for exponential families of markov random fields on the lattice. *The Annals of Statistics*, pages 455–468.

Comets, F. and Gidas, B. (1991). Asymptotics of maximum likelihood estimators for the curie-weiss model. *The Annals of Statistics*, pages 557–578.

Fauske, J. (2009). An empirical study of the maximum pseudo-likelihood for discrete markov random fields. Master's thesis, Institutt for matematiske fag.

Gan, L., Narisetty, N. N., and Liang, F. (2019). Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, 114(527):1218–1231.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures.

Ghosal, P., Mukherjee, S., et al. (2020). Joint estimation of parameters in ising model. *Annals of Statistics*, 48(2):785–810.

Ghosh, S., Khare, K., and Michailidis, G. (2018). High-dimensional posterior consistency in bayesian vector autoregressive models. *Journal of the American Statistical Association*.

Gidas, B. (1988). Consistency of maximum likelihood and pseudo-likelihood estimators for gibbs distributions. In *Stochastic differential systems, stochastic control theory and applications*, pages 129–145. Springer.

Guyon, X. and Künsch, H. R. (1992). Asymptotic comparison of estimators in the ising model. In *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis*, pages

177–198. Springer.

Halim, S. (2007). Modified ising model for generating binary images. *Jurnal Informatika*, 8(2):115–118.

Haslbeck, J. M., Epskamp, S., Marsman, M., and Waldorp, L. J. (2021). Interpreting the ising model: The input matters. *Multivariate behavioral research*, 56(2):303–313.

Hurn, M. A., Husby, O. K., and Rue, H. (2003). A tutorial on image analysis. *Spatial statistics and computational methods*, pages 87–141.

Ising, E. (1924). *Beitrag zur theorie des ferro-und paramagnetismus*. PhD thesis, Grefe & Tiedemann.

Izenman, A. J. (2021). Sampling algorithms for discrete markov random fields and related graphical models. *Journal of the American Statistical Association*, pages 1–22.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, H. K. (2000). Consistency of posterior distributions for neural networks. *Neural Networks*, 13(6):629–642.

Leskovec, J. and Krevl, A. (2014). SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`.

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American statistical association*, 105(491):1202–1214.

Li, F., Zhang, T., Wang, Q., Gonzalez, M. Z., Maresh, E. L., Coan, J. A., et al. (2015). Spatial bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *The Annals of Applied Statistics*, 9(2):687–713.

Lokhov, A. Y., Vuffray, M., Misra, S., and Chertkov, M. (2018). Optimal structure and parameter learning of ising models. *Science advances*, 4(3):e1700791.

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

Milman, V. and Schechtman, G. (1986). Asymptotic theory of finite-dimensional normed spaces, lecture notes in mathematics 1200.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Molkaraie, M. (2014). An importance sampling algorithm for the ising model with strong couplings. *arXiv preprint arXiv:1404.5666*.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817.

Okabayashi, S., Johnson, L., and Geyer, C. J. (2011). Extending pseudo-likelihood for potts models. *Statistica Sinica*, pages 331–347.

Park, J., Jin, I. H., and Schweinberger, M. (2022). Bayesian model selection for high-dimensional ising models, with applications to educational data. *Computational Statistics & Data Analysis*, 165:107325.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822. PMLR.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319.

Ray, K. and Szabó, B. (2021). Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, pages 1–12.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Ročková, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics*, 46(1):401–437.

Ročková, V. and George, E. I. (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Ročková, V. and George, E. I. (2016). Fast bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American*

Statistical Association, 113(521):431–444.

Sriram, K., Ramamoorthi, R., Ghosh, P., et al. (2013). Posterior consistency of bayesian quantile regression based on the misspecified asymmetric laplace density. *Bayesian Analysis*, 8(2):479–504.

Stanley, H. E. (1971). *Phase transitions and critical phenomena*, volume 7. Clarendon Press, Oxford.

Van Der Vaart, A. and Van Zanten, H. (2011). Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12(6).

Wang, Y. and Blei, D. M. (2019). Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Xi, R., Li, Y., and Hu, Y. (2016). Bayesian quantile regression based on the empirical likelihood with spike and slab priors. *Bayesian Analysis*, 11(3):821–855.

Xue, L., Zou, H., Cai, T., et al. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics*, 40(3):1403–1429.

Yang, K. and Shen, X. (2017). On the selection consistency of bayesian structured variable selection. *Stat*, 6(1):131–144.