THE UNIVERSITY OF CHICAGO

THEORETICAL GUARANTEES OF VARIATIONAL INFERENCE AND ITS APPLICATIONS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
FENGSHUO ZHANG

CHICAGO, ILLINOIS
JUNE 2020

To my wife Yao Tong and my parents Hongxiu Qi and Libin Zhang

"All models are wrong, but some are useful" — *George Edward Pelham Box*

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

First of all, I would like to express my heartfelt gratitude to my advisor, Chao Gao, for his consistent encouragement and support in my academic life. His insightful guidance and continuous involvement motivate me a lot through my research. Besides, I'm greatly indebted to all professors and teachers in the Department of Statistics, especially Yali Amit, Rina Barber, Dan Nicolae, Matthew Stephens, Mei Wang, and Weibiao Wu. They instructed and helped me a lot during the first several years to develop my fundamental statistical knowledge. Aside from that, I am also thankful to my friends Fan Yang, Haoyang Liu, Bumeng Zhuo, for their helpful discussions on my research topics.

Last but not least, special thanks should go to my parents Libin Zhang and Hongxiu Qi, for their selfless love and care all these years and my beloved wife Yao Tong, who gives me great encouragement and constant support during my graduate study. They always stay behind me and be my reliable shield.

ABSTRACT

Variational Inference (VI) has become a popular technique to approximate difficult-to-compute posterior distributions for decades. It has been used in many applications and tends to be faster than classical methods, such as Monte Carlo Markov Chain. However, there are few theoretical understandings about it. In this thesis, our goal is to build a statistical guarantee for the variational inference method under high-dimensional or nonparametric settings. We apply our theoretical results to develop a general variational Bayes (VB) algorithm for a group of high dimensional linear structure models. At the end of this thesis, we point out the relations between variational Bayes and empirical Bayes and propose a general convergence result for empirical Bayes posterior distributions.

In Chapter 2, we develop a "prior mass and testing" framework to show the concentration results of the variational posterior distribution and then apply these results to the Gaussian sequence model, infinite-dimensional exponential family, and piecewise constant model. We also propose the convergence results of variational posterior distribution with model selection. At the end of this chapter, we provide some discussions on the properties of variational inference.

In Chapter 3, we propose a general VB algorithm for a group of high dimensional linear structured models. These models include but are not limited to stochastic block model, biclustering model, sparse linear regression, multiple regression with group sparsity, multi-task learning, and dictionary learning. Theoretically, we can prove an oracle type of contraction result for the variational posterior distribution. Empirically, the VB algorithm outperforms the classical spectral method in the stochastic block model and LASSO estimator in sparse linear regression as long as the signal-to-noise ratio is large.

In Chapter 4, we demonstrate that the empirical Bayes procedure can be viewed as the variational Bayes procedure with a particular variational set. Then, we propose a theorem for the concentration result of Empirical Bayes posterior distributions under the case when the true parameters are unbounded. Finally, this result is applied to the sparse sequence model,

sparse linear regression, and the general linear structured models discussed in Chapter 3.

CHAPTER 1

INTRODUCTION

1.1 Introduction to Variational Inference

Variational Bayes inference is a popular technique to approximate difficult-to-compute posterior distributions. Given a posterior distribution $\Pi(\cdot|X^{(n)})$ and a variational family $\mathcal{S}$, variational Bayes inference seeks a $\widehat{Q} \in \mathcal{S}$ that best approximates $\Pi(\cdot|X^{(n)})$ under the Kullback-Leibler divergence. Though it is not exact Bayes inference, the variational class $\mathcal{S}$ often gives computational advantages and leads to algorithms such as coordinate ascent that can be efficiently implemented on large-scale data sets. Researchers in many fields have used variational Bayes inference to solve real problems. Successful examples include statistical genetics [12, 49], natural language processing [11, 43], computer vision [59], and network analysis [8, 71], to name a few. We refer the readers to an excellent recent review [10] on this topic.

1.2 Convergence Theory of Posterior Distributions

Before discussing our results, we first introduce the convergence theory of posterior distributions. In a Bayes procedure, we assume the model is given by P_θ and a prior Π is put on the parameter θ. Then the posterior distribution is given by

$$\Pi\left(B|X^{(n)}\right) = \frac{\int_B p_\theta(X^{(n)})d\Pi(\theta)}{\int p_\theta(X^{(n)})d\Pi(\theta)}.$$

From a frequentist perspective, we can assume the observation $X^{(n)}$ is generated from a "true" distribution. Heuristically, if the prior put some mass around this true distribution and the true distribution P_{θ^*} can be well distinguished from P_θ for $\theta \neq \theta^*$ by the observation $X^{(n)}$, then the posterior will tend to put more mass around the true distribution. When the sample size is getting larger and larger, the testing power between θ^* from other parameters

is getting larger and larger, and then the posterior should be more and more concentrated around the true distribution. This theory is regarded as the convergence theory of the posterior distributions.

In order that posterior distributions concentrate around the true data generating distribution, "prior mass and testing" conditions are required: a) The prior is required to put a minimal amount of mass in a neighborhood of the true parameter; b) There exists a testing function that can distinguish the truth from the complement of its neighborhood by data. These conditions can be traced back to [55, 44, 5, 4]. However, under nonparametric settings, "prior mass and testing" conditions do not hold because the dimension of the parameter space is infinite. Due to this reason, no testing function is able to separate the truth from all the other parameters. At the beginning of this century, this issue is solved in [31, 57] by assuming the prior mass on the untestable parameter set to be sufficiently small. The rigorous theorem for the convergence result is given in Theorem 1.2.1.

Theorem 1.2.1 (Ghoshal, Ghosh, Van Der Vaart)**.** *Suppose $X_1, \cdots, X_n$ are i.i.d generated from P_0 and for a sequence ϵ_n with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ and a metric d, there exist constants $C, C_1, C_2 > 0$ and a set $\Theta_n \subset \Theta$ satisfying the following conditions.*

- *There exists a testing function ϕ_n, such that*

$$P_0^{(n)}\phi_n + \sup_{\substack{\theta\in\Theta_n \\ d(P_\theta,P_0)\geq C_1\epsilon_n}} P_\theta^{(n)}(1-\phi_n) \leq \exp(-C_2 n\epsilon_n^2). \tag{1.1}$$

- $$\Pi(\Theta_n^c) \leq \exp(-Cn\epsilon_n^2). \tag{1.2}$$

- $$\Pi\left(P : -P_0\left(-\log\frac{p}{p_0}\right) \leq \epsilon_n^2, \quad P_0\left(\log\frac{p}{p_0}\right)^2 \leq \epsilon_n^2\right) \geq \exp(-C_2 n\epsilon_n^2). \tag{1.3}$$

Then for sufficiently large M, we have $\Pi\left(P : d(P, P_0) \geq M\epsilon_n | X_1, \cdots, X_n\right) \to 0$ in P_0^n-

probability.

This theorem is a great breakthrough to understand the frequentist properties of Bayes analysis. There are lots of literatures about the convergence results for posterior distributions following it. Some representative work is given by [32, 52, 30, 42]. Aside from this, we also mention another line of work by [74, 67, 14, 34] that established posterior rates of convergence using other approaches. Similar "prior mass and testing" type of conditions are also discussed in [54] to establish the convergence results for Empirical Bayes posterior distributions. However, the "prior mass" condition usually does not hold when the true parameter is assumed unbounded. Usually, a long tail prior on parameters is applied to deal with this problem. We refer to [28, 16, 18] for related topics.

1.3 Structure of Thesis

In this thesis, we first establish a theory on the variational posterior distribution $\widehat{Q}$ in Chapter 2. Then we propose a general variational Bayes algorithm for high dimensional linear structured models in Chapter 3 and analyze the convergence rates of empirical Bayes posterior distributions in Chapter 4. Detailed results are introduced in the following paragraphs.

In Chapter 2, we show that under almost the same three conditions, the variational posterior $\widehat{Q}$ also converges to the true parameter, and the rate of convergence consists of two parts. The first part is the rate of convergence of the posterior distribution $\Pi(\cdot|X^{(n)})$, and the second term is the variational approximation error with respect to the class $\mathcal{S}$ under the data generating process $P_0^{(n)}$. Since we are able to generalize the "prior mass and testing" theory with the same old conditions, many well-studied problems in the literatures can now be revisited under our framework of variational Bayes inference with very similar proof techniques. This type of extensions will be illustrated with several examples considered in this thesis. Moreover, this general theorem is also specialized to the widely-used mean field variational distribution family and applied to several nonparametric models. At the end of

this chapter, we generalize the convergence theorem to a general variational inference with a model selection setting and a general misspecified model, which is the underlying support to the theory in Chapter 3 and Chapter 4.

In Chapter 3, we provide a general variational algorithm to the general linear structured model in [28]. The prior we apply is equivalent to the prior in [28] with a modification to a more hierarchical sampling procedure to assist the variational algorithm. With a point mass variational set for the distribution on structure label Z, we propose a similar oracle type of convergence result as in [28] for the variational posterior distribution with model selection. Then we propose a general variational algorithm to solve for the variational posterior distribution and specialize it into several linear structured models. At the end of this chapter, some simulation results are provided on the stochastic block model and sparse linear regression model. Simulation results show that the variational algorithm can improve the results from classical methods when the signal-to-noise ratio is large regardless of the singularity of the design matrix and the range of the true parameter.

In Chapter 4, we point out that in some cases, the empirical Bayes procedure is actually the variational Bayes procedure with a special variational distribution set $\mathcal{S}_{\mathrm{EB}}$. Besides, a convergence result for empirical Bayes can be directly obtained from the theory in the variational Bayes with model selection. However, this theorem cannot be generalized when the true parameter is not bounded. Thus, we propose weaker "prior mass ratio and testing" conditions to show the convergence of empirical Bayes posterior distribution. Finally, we apply this general theorem to the sparse sequence model, linear regression model, and the linear structured model that is considered in Chapter 3.

All deferred proofs are provided in Chapter 5.

1.4 Notation

We close this chapter by introducing notations that will be used later. For $a, b \in \mathbb{R}$, let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For a positive real number x, $\lceil x \rceil$ is the smallest

integer no smaller than x and $\lfloor x \rfloor$ is the largest integer no larger than x.

For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if $a_n \leq C b_n$ for all n with some constant $C > 0$ that does not depend on n. The relation $a_n \asymp b_n$ holds if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold. For an integer m, $[m]$ denotes the set $\{1, 2, ..., m\}$. Given a set S, $|S|$ denotes its cardinality, and $\mathbb{I}_S$ is the associated indicator function. In general, we also use $\mathbb{I}_A$ to denote the indicator function for the event A. The ℓ_p norm of a vector $v \in \mathbb{R}^m$ with $1 \leq m \leq \infty$ is defined as $\|v\|_p = \left(\sum_{j=1}^m |v_j|^p\right)^{1/p}$ for $1 \leq p < \infty$ and $\|v\|_\infty = \sup_{1 \leq k \leq m} |v_k|$. Moreover, we use $\|v\|$ to denote the ℓ_2 norm $\|v\|_2$ by convention. For any function f, the ℓ_p norm is defined in a similar way, i.e. $\|f\|_p = \left(\int f(x)^p dx\right)^{1/p}$. Specifically, $\|f\|_\infty = \sup_x |f(x)|$.

$\mathbb{P}$ and $\mathbb{E}$ are used to denote generic probability and expectation whose distribution is determined from the context. The notation $\mathbb{P}f$ also means expectation of f under $\mathbb{P}$ so that $Pf = \int f dP$. Throughout the paper, M, C, c, and their variants denote generic constants that do not depend on n. Their values may change from line to line.

$N(\mu, \Sigma)$ stands for the normal distribution with the mean μ and covariance matrix Σ. For a matrix Σ, $\Sigma \succ 0$ represents that Σ is a positive definite symmetric matrix. Bern(θ) stands for the Bernoulli distribution with probability θ. For a matrix $B = (b_1, \cdots, b_m) \in \mathbb{R}^{n \times m}$, we use Vec$(B)$ to denotes its column vectorization $(b_1^T, \cdots, b_m^T)^T \in \mathbb{R}^{mn}$. For any two matrices A and B, $\otimes$ denotes the Kronecker product between them and $A^{\otimes 2} = A \otimes A$.

CHAPTER 2

CONVERGENCE RATES OF VARIATIONAL POSTERIOR DISTRIBUTIONS

2.1 Introduction

During the past few decades, the theory of convergence rates for posterior distribution encounters a significant development. One of the most popular theories is the "prior mass and testing" framework to verify the convergence rates of the posterior distributions in the nonparametric setting. We refer [31, 57, 32] for details to these works. However, as the posterior distribution is usually computationally intractable, the theoretical property of its substitute, variational approximation of the posterior distribution (variational posterior thereafter), starts to get attention from statisticians.

In this chapter[1], we aim to build a convergence theory for the variational posterior. First of all, we show that under almost the same three conditions, the variational posterior $\widehat{Q}$ also converges to the true parameter with the rate of convergence given by

$$\epsilon_n^2 + \frac{1}{n} \inf_{Q \in \mathcal{S}} P_0^{(n)} D(Q \| \Pi(\cdot | X^{(n)})). \tag{2.1}$$

The first term ϵ_n^2 is the rate of convergence of the posterior distribution $\Pi(\cdot|X^{(n)})$. The second term is the variational approximation error with respect to the class $\mathcal{S}$ under the data generating process $P_0^{(n)}$. When $\mathcal{S}$ denotes to all distributions, the variational approximation error can be improved to be 0, which is consistent with theory for the convergence rate of posterior distribution.

Remarkably, for a special class of prior distributions and a corresponding variational class, the second term of (2.1) will be automatically dominated by ϵ_n^2 under a modified

1. The work presented in this chapter is published in [73].

"prior mass" condition. We illustrate this result by a prior distribution of product measure

$$d\Pi(\theta) = \prod_j d\Pi_j(\theta_j),$$

and a mean-field variational class

$$\mathcal{S}_{\mathrm{MF}} = \left\{ Q : dQ(\theta) = \prod_j dQ_j(\theta_j) \right\}.$$

As long as there exists a subset $\otimes_j \widetilde{\Theta}_j \subset \left\{ \theta : D_\rho \left(P_0^{(n)} \| P_\theta^{(n)} \right) \leq C_1 n \epsilon_n^2 \right\}$, such that the prior mass condition

$$\Pi \left(\otimes_j \widetilde{\Theta}_j \right) \geq \exp \left(-C_2 n \epsilon_n^2 \right) \tag{2.2}$$

holds together with the testing conditions, then the variational posterior distribution $\widehat{Q}$ converges to the true parameter with the rate ϵ_n^2. In other words, the variational approximation error term in (2.1) is dominated under this stronger prior mass condition (2.2). This is the result of Theorem 2.2.4. Here, $D_\rho(\cdot\|\cdot)$ stands for a Rényi divergence with some $\rho > 1$. The implication of the condition (2.2) is important. It says that as long as the prior satisfies a "prior mass" condition that is coherent with the structure of the variational class, the resulted variational approximation error will always be small compared with the statistical error from the true posterior. Therefore, the condition (2.2) offers a practical guidance on how to choose a good prior for variational Bayes inference. In addition, as a condition only on the prior mass, (2.2) is usually very easy to check. This mathematical simplicity is not just for independent priors and the mean-field class. In Section 2.4, a general model selection setting is studied that includes the setting of (2.2) as a special case.

With the general formulation of conditions, we study some popular nonparametric models such as the Gaussian sequence model, infinite-dimensional density exponential families, and piecewise constant model for their convergence behavior of variational posterior. Then we establish the convergence rate for variational distribution with model selection, in which

dimensions of parameter spaces are assumed different for different models and apply this theory on the variational posterior distribution of nonparametric mixture density estimation.

Finally, we would like to remark that the general rate (2.1) for variational posteriors is only an upper bound. It is *not* always true that the variational posterior has a slower convergence rate than the true posterior. Sometimes the variational posterior may not be a good approximation to the true posterior, but it can still contract faster to the true parameter if additional regularity is imposed by the variational class $\mathcal{S}$. We construct examples in Section 2.5.1 to illustrate this point.

Recently, statistical properties of variational posterior distributions have also been studied in the literature. A recent work by [68] established Bernstein-von Mises' type of results for parametric models. For nonparametric and high-dimensional models, recent work by [1, 69] studied variational approximation to tempered posteriors, where the likelihood $dP_\theta^{(n)}/dP_0^{(n)}$ is replaced by $\left(dP_\theta^{(n)}/dP_0^{(n)}\right)^\alpha$ for some $\alpha \in (0,1)$. Just as the convergence of tempered posteriors [66], the convergence of the variational approximation can also be established under generalizations of the prior mass condition. Besides, the paper [1] also studied convergence rates under model misspecification, and the paper [69] considered a more general setting that can handle latent variables, which is quite useful to analyze mixture models. We would like to point out that these results do not apply to the normal posterior distributions with $\alpha = 1$. Later on, similar results on $\alpha = 1$ have also been obtained independently by [46][2]. Early related work on this topic is by [74], where the results cover both posterior distributions and their variational approximations. However, the conditions in [74] are rather abstract and are not easy to check in applications.

The rest of this chapter is organized as follows. In Section 2.2, we formulate the problem and introduce the general conditions that characterize convergence rates of variational posteriors. This section also includes results for the mean-field variational class, where the variational approximation error can be explicitly analyzed. In Section 2.3, we apply our gen-

2. Some extensions of the results of [46] were later added in the revised version of [69] by the same authors.

eral theory to three nonparametric or high dimensional examples. Then, in Section 2.4, for a general class of prior distributions and a mean-field class under a model selection setting, we propose a new prior mass condition that leads to automatic control of the variational approximation error. In Section 2.5, we discuss possible situations where the variational posterior outperforms the true posterior in this section. An extension of the main results under model misspecification is also discussed in Section 2.5.

2.2 Main Results

2.2.1 Definitions and Settings

We start this section by introducing a class of divergence functions.

Definition 2.2.1 (Rényi divergence)**.** *Let $\rho > 0$ and $\rho \neq 1$. The ρ-Rényi divergence between two probability measures P_1 and P_2 is defined as*

$$D_\rho(P_1\|P_2) = \begin{cases} \frac{1}{\rho-1}\log\int\left(\frac{dP_1}{dP_2}\right)^{\rho-1}dP_1, & \text{if } P_1 \ll P_2, \\ +\infty, & \text{otherwise.} \end{cases}$$

The relations between the Rényi divergence and other divergence functions are summarized below.

1. When $\rho \to 1$, the Rényi divergence converges to the Kullback-Leibler divergence, defined as
$$D_1(P_1\|P_2) = \begin{cases} \int \log\left(\frac{dP_1}{dP_2}\right)dP_1, & \text{if } P_1 \ll P_2, \\ +\infty, & \text{otherwise.} \end{cases}$$
From now on, we use $D(P_1\|P_2)$ without the subscript to denote $D_1(P_1\|P_2)$.

2. When $\rho = 1/2$, the Rényi divergence is related to the Hellinger distance by
$$D_{1/2}(P_1\|P_2) = -2\log(1 - H(P_1, P_2)^2),$$

and the Hellinger distance is defined as

$$H(P_1, P_2) = \sqrt{\frac{1}{2}\int (\sqrt{dP_1} - \sqrt{dP_2})^2}.$$

3. When $\rho = 2$, the Rényi divergence is related to the χ^2-divergence by

$$D_2(P_1\|P_2) = \log(1 + \chi^2(P_1\|P_2)),$$

and the χ^2-divergence is defined as

$$\chi^2(P_1\|P_2) = \int \frac{(dP_1)^2}{dP_2} - 1.$$

Definition 2.2.2 (total variation)**.** *The total variation distance between two probability measures P_1 and P_2 is defined as*

$$\mathsf{TV}(P_1, P_2) = \frac{1}{2}\int |dP_1 - dP_2|\,.$$

The relation among the divergence functions defined above is given by the following proposition (see [64]).

Proposition 2.2.1. *With the above definitions, the following inequalities hold,*

$$TV(P_1, P_2)^2 \le 2H(P_1, P_2)^2 \le D_{1/2}(P_1\|P_2)$$
$$\le D(P_1\|P_2) \le D_2(P_1\|P_2) \le \chi^2(P_1\|P_2).$$

Moreover, the Rényi divergence $D_\rho(P_1\|P_2)$ is a non-decreasing function of ρ.

Now we are ready to introduce the variational posterior distribution. Given a statistical model $P_\theta^{(n)}$ parametrized by θ, and a prior distribution $\theta \sim \Pi$, the posterior distribution is

defined by

$$d\Pi(\theta|X^{(n)}) = \frac{dP_\theta^{(n)}(X^{(n)})d\Pi(\theta)}{\int dP_\theta^{(n)}(X^{(n)})d\Pi(\theta)}.$$

To address possible computational difficulty of the posterior distribution, variational approximation is a way to find the closest object in a class $\mathcal{S}$ of probability measures to $\Pi(\cdot|X^{(n)})$. The class $\mathcal{S}$ is usually required to be computationally or analytically tractable. The most popular mathematical definition of variational approximation is given through the KL-divergence.

Definition 2.2.3 (variational posterior)**.** *Let $\mathcal{S}$ be a family of distributions. The variational approximation of the posterior is defined as*

$$\widehat{Q} = \underset{Q\in\mathcal{S}}{\operatorname{argmin}}\, D(Q\|\Pi(\cdot|X^{(n)})). \tag{2.3}$$

Just like the posterior distribution $\Pi(\cdot|X^{(n)})$, the variational posterior $\widehat{Q}$ is a data-dependent measure that summarizes information from both the prior and the data. For a variational set $\mathcal{S}$, the corresponding variational posterior can be regarded as the projection of the true posterior onto $\mathcal{S}$ under KL-divergence. When $\mathcal{S}$ is the set of all distributions, $\widehat{Q}$ turns out to be the true posterior $\Pi(\cdot|X^{(n)})$. The choice of the class $\mathcal{S}$ usually determines the difficulty of the optimization (2.3). In this chapter, our main goal is to study the statistical property of the data-dependent measure $\widehat{Q}$ for a general $\mathcal{S}$.

2.2.2 Results for General Variational Posteriors

Assume the observation $X^{(n)}$ is generated from a probability measure $P_0^{(n)}$, and $\widehat{Q}$ is the variational posterior distribution driven by $X^{(n)}$. The goal of this chapter is to analyze $\widehat{Q}$ from a frequentist perspective. In other words, we study statistical properties of $\widehat{Q}$ under $P_0^{(n)}$. The first theorem gives conditions that guarantee convergence of the variational posterior $\widehat{Q}$.

Theorem 2.2.1. *Suppose ϵ_n is a sequence that satisfies $n\epsilon_n^2 \geq 1$. Consider a loss function $L(\cdot,\cdot)$, such that for any two probability measures P_1 and P_2, $L(P_1,P_2) \geq 0$. Let $C, C_1, C_2, C_3 > 0$ be constants such that $C > C_2 + C_3 + 2$. We assume*

- *For any $\epsilon > \epsilon_n$, there exists a set $\Theta_n(\epsilon)$ and a testing function ϕ_n, such that*

$$P_0^{(n)}\phi_n + \sup_{\substack{\theta\in\Theta_n(\epsilon) \\ L(P_\theta^{(n)},P_0^{(n)})\geq C_1 n\epsilon^2}} P_\theta^{(n)}(1-\phi_n) \leq \exp(-Cn\epsilon^2). \tag{C1}$$

- *For any $\epsilon > \epsilon_n$, the set $\Theta_n(\epsilon)$ above satisfies*

$$\Pi(\Theta_n(\epsilon)^c) \leq \exp(-Cn\epsilon^2). \tag{C2}$$

- *For some constant $\rho > 1$,*

$$\Pi\left(D_\rho(P_0^{(n)}\|P_\theta^{(n)}) \leq C_3 n\epsilon_n^2\right) \geq \exp(-C_2 n\epsilon_n^2). \tag{C3}$$

Then for the variational posterior $\widehat{Q}$ defined in (2.3), we have

$$P_0^{(n)}\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) \leq Mn(\epsilon_n^2 + \gamma_n^2), \tag{2.4}$$

for some constant M only depending on C_1, C and ρ, where the quantity γ_n^2 is defined as

$$\gamma_n^2 = \frac{1}{n}\inf_{Q\in\mathcal{S}} P_0^{(n)} D(Q||\Pi(\cdot|X^{(n)})).$$

Conditions (C1)-(C3) resemble the three conditions of "prior mass and testing" in [31]. Interestingly, Theorem 2.2.1 shows that with a slight modification, these three conditions also lead to the convergence of the variational posterior. The testing conditions (C1) and (C2) are required to hold for all $\epsilon > \epsilon_n$. In the prior mass condition (C3), the neighborhood of

$P_0^{(n)}$ is defined through a Rényi divergence with a $\rho > 1$, compared with the KL-divergence used in [31]. According to Proposition 2.2.1, $D_\rho(P_1\|P_2) \geq D(P_1\|P_2)$ for $\rho > 1$, so the condition (C3) in our paper is slightly stronger than that in [31]. This stronger "prior mass" condition ensures that the loss $L(P_\theta^{(n)}, P_0^{(n)})$ is exponentially integrable under the true posterior $\Pi(\cdot|X^{(n)})$, which is a key step in the proof of Theorem 2.2.1. In all the examples considered in this chapter, we will check (C3) with $D_2(P_0^{(n)}\|P_\theta^{(n)})$, which turns out to be a very convenient choice.

The convergence rate is the sum of two terms, ϵ_n^2 and γ_n^2. The first term ϵ_n^2 is the convergence rate of the true posterior $\Pi(\cdot|X^{(n)})$. The second term γ_n^2 characterizes the approximation error given by the variational set $\mathcal{S}$. A larger $\mathcal{S}$ means more expressive power given by the variational approximation, and thus the rate of γ_n^2 is smaller.

It is worth mentioning that we characterize the convergence of the variational posterior $\widehat{Q}$ through the expected loss $P_0^{(n)}\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)})$. Bounds for this quantity are also obtained by [46] independently with a stronger testing condition on the entire space. We remark that convergence in $P_0^{(n)}\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)})$ automatically implies that the entire variational posterior distribution concentrates in a neighborhood of the true distribution $P_0^{(n)}$ with a radius of the same rate. When the loss function is convex, it also implies the existence of a point estimator that enjoys the same convergence rate. We summarize these results in the next corollary.

Corollary 2.2.1. *Under the same setting of Theorem 2.2.1, for any diverging sequence $M_n \to \infty$, we have*

$$P_0^{(n)}\widehat{Q}\left(L(P_\theta^{(n)}, P_0^{(n)}) > M_n n(\epsilon_n^2 + \gamma_n^2)\right) \to 0.$$

Furthermore, if the loss $L(P_\theta^{(n)}, P_0^{(n)})$ is convex respect to θ, then the variational posterior mean $\widehat{\theta} = \widehat{Q}\theta$ satisfies

$$P_0^{(n)}L(P_{\widehat{\theta}}^{(n)}, P_0^{(n)}) \leq Mn(\epsilon_n^2 + \gamma_n^2),$$

where M is the same constant in (2.4).

Proof. The first result is an application of Markov's inequality

$$P_0^{(n)}\widehat{Q}\left(L(P_\theta^{(n)}, P_0^{(n)}) > M_n n(\epsilon_n^2+\gamma_n^2)\right) \le \frac{P_0^{(n)}\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)})}{M_n n(\epsilon_n^2+\gamma_n^2)} \le \frac{M}{M_n} \to 0.$$

The second result is directly implied by Jensen's inequality that

$$P_0^{(n)}L(P_{\widehat{Q}\theta}^{(n)}, P_0^{(n)}) \le P_0^{(n)}\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) \le Mn(\epsilon_n^2+\gamma_n^2).$$

□

To apply Theorem 2.2.1 to specific problems, we need to analyze the variational approximation error $\gamma_n^2 = \frac{1}{n}\inf_{Q\in\mathcal{S}} P_0^{(n)} D(Q||\Pi(\cdot|X^{(n)}))$ in each individual setting. However, this task may not be trivial for many problems. Now we borrow a technique in [74] to get a useful upper bound for γ_n^2. For any $Q \in \mathcal{S}$, we have

$$\begin{aligned} n\gamma_n^2 &\le P_0^{(n)} D(Q\|\Pi(\cdot|X^{(n)})) = D(Q\|\Pi) + Q\left[\int \log\left(\frac{dP_\Pi^{(n)}}{dP_\theta}\right) dP_0^{(n)}\right] \\ &= D(Q\|\Pi) + Q\left[D(P_0^{(n)}\|P_\theta^{(n)}) - D(P_0^{(n)}\|P_\Pi^{(n)})\right] \\ &\le D(Q\|\Pi) + Q\left[D(P_0^{(n)}\|P_\theta^{(n)})\right], \end{aligned}$$

where $P_\Pi^{(n)} = \int P_\theta^{(n)} d\Pi(\theta)$. Then, we obtain the upper bound

$$\gamma_n^2 \le \inf_{Q\in\mathcal{S}} R(Q),$$

where

$$R(Q) = \frac{1}{n}\left(D(Q\|\Pi) + Q\left[D(P_0^{(n)}\|P_\theta^{(n)})\right]\right). \tag{2.5}$$

Now, it is easy to see that a sufficient condition for the variational posterior to converge at the same rate as the true posterior is

$$\inf_{Q\in\mathcal{S}} R(Q) \lesssim \epsilon_n^2. \tag{C4}$$

We incorporate this condition into the next theorem.

Theorem 2.2.2. *Suppose ϵ_n is a sequence that satisfies $n\epsilon_n^2 \geq 1$, for which the conditions (C1), (C2), (C3), (C4) hold. Then, for the variational posterior $\widehat{Q}$ that is defined in (2.3), we have*

$$P_0^{(n)}\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) \lesssim n\epsilon_n^2. \tag{2.6}$$

We would like to remark that the quantity $\inf_{Q\in\mathcal{S}} R(Q)$ is easier to analyze compared with the original definition of γ_n^2. According to its definition given by (2.5), it is sufficient to find a distribution $Q \in \mathcal{S}$, such that

$$D(Q\|\Pi) \lesssim n\epsilon_n^2 \quad \text{and} \quad Q\left[D(P_0^{(n)}\|P_\theta^{(n)})\right] \lesssim n\epsilon_n^2. \tag{2.7}$$

These are exactly the two conditions formulated by [1] as a natural extension of the prior mass condition. The relation between the prior mass condition and (2.7) has also been discussed in [69].

One way to construct such a distribution Q that satisfies the above two inequalities is to focus on those whose supports are within the set $\mathcal{C} = \{\theta : D(P_0^{(n)}\|P_\theta^{(n)}) \leq Cn\epsilon_n^2\}$ for some constant $C > 0$. We summarize this method into the following theorem.

Theorem 2.2.3. *Suppose there exist constants $C_1, C_2 > 0$, such that*

$$\inf_{Q\in\mathcal{S}\cap\mathcal{E}} D(Q\|\Pi) \leq C_1 n\epsilon_n^2, \tag{C4*}$$

where $\mathcal{E} = \{Q : \text{supp}(Q) \subset \mathcal{C}\}$ *with* $\mathcal{C} = \{\theta : D(P_0^{(n)} \| P_\theta^{(n)}) \leq C_2 n\epsilon_n^2\}$. *Then, we have*

$$\inf_{Q \in \mathcal{S}} R(Q) \leq (C_1 + C_2)\epsilon_n^2.$$

2.2.3 Results for Mean-Field Variational Posteriors

A special choice of $\mathcal{S}$ is the mean-field class of distributions. Not only does this class leads to computationally efficient algorithms such as coordinate ascent, but in this section, we will also show that the structure of this class leads to a convenient convergence analysis. We begin with its definition.

Definition 2.2.4 (mean-field class)**.** *For parameters in a product space that can be written as* $\theta = (\theta_1, \theta_2, ..., \theta_m)$ *with some* $1 \leq m \leq \infty$*, the mean-field variational family is defined as*

$$\mathcal{S}_{\text{MF}} = \left\{ Q : dQ(\theta) = \prod_{j=1}^{m} dQ_j(\theta_j) \right\}.$$

The following theorem can be viewed as an application of Theorem 2.2.3 to the mean-field class.

Theorem 2.2.4. *Suppose there exists a* $\widetilde{Q} \in \mathcal{S}_{\text{MF}}$ *and a subset* $\otimes_{j=1}^m \widetilde{\Theta}_j$*, such that*

$$\otimes_{j=1}^m \widetilde{\Theta}_j \subset \left\{ \theta : D(P_0^{(n)} \| P_\theta^{(n)}) \leq C_1 n\epsilon_n^2, \quad \log \frac{d\widetilde{Q}(\theta)}{d\Pi(\theta)} \leq C_2 n\epsilon_n^2 \right\}, \tag{2.8}$$

and

$$-\sum_{j=1}^{m} \log \widetilde{Q}_j(\widetilde{\Theta}_j) \leq C_3 n\epsilon_n^2, \tag{2.9}$$

for some constants $C_1, C_2, C_3 > 0$*. Then, we have*

$$\inf_{Q \in \mathcal{S}_{\text{MF}}} R(Q) \leq (C_1 + C_2 + C_3)\epsilon_n^2.$$

Note that the condition (2.9) can also be written as

$$\widetilde{Q}\left(\otimes_{j=1}^{m}\widetilde{\Theta}_j\right) \geq \exp\left(-C_3 n\epsilon_n^2\right).$$

In other words, Theorem 2.2.4 gives an interesting "distribution mass" type of characterization for $\inf_{Q\in\mathcal{S}} R(Q)$. Checking (2.9) is very similar to checking the "prior mass" condition (C3), and is usually not hard in many examples. We only need to make sure that $\widetilde{Q}$ is not too far away from the prior Π in the sense of (2.8). In fact, if the prior Π belongs to the class $\mathcal{S}_{\mathrm{MF}}$, then one can take $\widetilde{Q} = \Pi$, and the conditions of Theorem 2.2.4 simply become a "prior mass" condition $\Pi\left(\otimes_{j=1}^{m}\widetilde{\Theta}_j\right) \geq \exp\left(-C_3 n\epsilon_n^2\right)$, with the choice of $\otimes_{j=1}^{m}\widetilde{\Theta}_j$ being a subset of the KL-neighborhood $\left\{\theta : D(P_0^{(n)}\|P_\theta^{(n)}) \leq C_1 n\epsilon_n^2\right\}$. A more general characterization of the variational approximation error under model selection setting through a prior mass condition will be studied in Section 2.4.

2.3 Applications

In this section, we consider several examples to illustrate the theory developed in Section 2.2.

2.3.1 Gaussian Sequence Model

Consider observations generated by a Gaussian sequence model,

$$Y_j = \theta_j + \frac{1}{\sqrt{n}}Z_j, \qquad Z_j \overset{i.i.d}{\sim} N(0,1), \qquad j \geq 1. \tag{2.10}$$

We use the notation $P_\theta^{(n)} = \otimes_j N(\theta_j, n^{-1})$ for the distribution above. Our goal is to use variational Bayes methods to estimate the true parameter θ^* that belongs to the following

Sobolev ball,

$$\Theta_\alpha(B) = \left\{ \theta = (\theta_j)_{j=1}^\infty : \sum_{j=1}^\infty j^{2\alpha}\theta_j^2 \le B^2 \right\}. \tag{2.11}$$

Here, the smoothness $\alpha > 0$ and the radius $B > 0$ are considered as constants throughout the paper. The loss function for this problem is $L(P_\theta^{(n)}, P_{\theta^*}^{(n)}) = n\|\theta - \theta^*\|^2$, which is a natural choice for the Gaussian sequence model.

The prior distribution $\theta \sim \Pi$ is described through the following sampling process.

1. Sample $k \sim \pi$;
2. Conditioning on k, sample $\theta_j \sim f_j$ for all $j \in [k]$, and set $\theta_j = 0$ for all $j > k$.

In other words, the prior on θ is a mixture of product measures,

$$d\Pi(\theta) = \sum_{k=1}^\infty \pi(k) \prod_{j=1}^k f_j(\theta_j) \prod_{j>k} \delta_0(\theta_j) d\theta. \tag{2.12}$$

Priors of similar forms are also considered in [52, 29, 28, 54]. Direct calculation implies that the posterior is also in the form of a mixture of product measures.

Consider the variational posterior $\widehat{Q}$ defined by (2.3) with $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}$. That is, we seek a data-dependent measure in a more tractable form of a product measure. In most cases, the variational posterior does not have a closed form and needs to be solved by coordinate ascent algorithms [10]. However, for the Gaussian sequence model (2.10) with the prior distribution (2.12), one can write down the exact form of the mean-field variational posterior distribution.

Theorem 2.3.1. *Consider the variational posterior $\widehat{Q}$ induced by the likelihood (2.10), the prior (2.12) and the mean-field variational set $\mathcal{S}_{\mathrm{MF}}$. The distribution $\widehat{Q}$ is a product measure with the density of each coordinate specified by*

$$q_j = \begin{cases} \widetilde{f}_j, & j < \widetilde{k}, \\ \widetilde{p}\delta_0 + (1-\widetilde{p})\widetilde{f}_{\widetilde{k}}, & j = \widetilde{k}, \\ \delta_0, & j > \widetilde{k}. \end{cases} \tag{2.13}$$

where

$$\widetilde{f}_j(\theta_j) \propto f_j(\theta_j) \exp\left(-\frac{n}{2}(\theta_j - Y_j)^2\right),$$

$$\widetilde{p} = \frac{\pi(k-1|Y)}{\pi(k-1|Y) + \pi(k|Y)},$$

and

$$\widetilde{k} = \underset{k}{\operatorname{argmax}} \left(\pi(k-1|Y) + \pi(k|Y)\right). \tag{2.14}$$

The number $\pi(k|Y)$ is the posterior probability of the model dimension, and according to Bayes formula, it is

$$\pi(k|Y) \propto \pi(k) \prod_{j \le k} \int f_j(\theta_j) \exp\left(-\frac{n(\theta_j - Y_j)^2}{2}\right) d\theta_j \prod_{j>k} \exp\left(-\frac{nY_j^2}{2}\right).$$

In other words, the mean-field variational posterior $\widehat{Q}$ is nearly equivalent to a thresholding rule. It estimates all θ_j^* by 0 after $\widetilde{k}$ and applies the usual posterior distribution for each coordinate before $\widetilde{k}$. A mixed strategy is applied to the $\widetilde{k}$th coordinate. The effective model dimension $\widetilde{k}$ is found in a data-driven way through (2.14).

Next, we will show that even though the posterior itself is not a product measure, using $\widehat{Q}$ from the mean-field class still gives us a rate-optimal contraction result. The conditions on the prior distributions are summarized below.

- There exist some constants $C_1, C_2 > 0$ such that

$$\sum_{j=k}^{\infty} \pi(j) \le C_1 \exp(-C_2 k), \text{ for all } k. \tag{2.15}$$

- There exist some constants $C_3, C_4 > 0$ such that for $k_0 = \left\lceil \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}} \right\rceil$,

$$\pi(k_0) \ge C_3 \exp(-C_4 k_0 \log k_0). \tag{2.16}$$

- For the k_0 defined above, there exist some constants $c_0 \in \mathbb{R}$ and $c_1 > 0$ such that

$$-\log f_j(x) \le c_0 + c_1 j^{2\alpha+1} x^2, \qquad \text{for all } j \le k_0 \text{ and } x \in \mathbb{R}. \tag{2.17}$$

These three conditions on Π include a large class of prior distributions. We remark that even though (2.17) involves α, it does not mean that one needs to know α when defining the prior Π. For example, the choice that $\pi(k) \propto e^{-\tau k}$ and f_j being $N(0, \sigma^2)$ for some constants $\tau, \sigma^2 > 0$ easily satisfies all the three conditions (2.15)-(2.17).

Conditions (2.15)-(2.17) will be used to derive the four conditions in Theorem 2.2.2. To be specific, (C1) and (C2) are consequences of (2.15) (see Lemma 5.1.7), and (C3) and (C4) can be derived from (2.16) and (2.17) (see Lemma 5.1.8). Then, by Theorem 2.2.2, we obtain the following result.

Theorem 2.3.2. *Consider the prior* Π *that satisfies (2.15)-(2.17). Then, for any* $\theta^* \in \Theta_\alpha(B)$*, we have*

$$P_{\theta^*}^{(n)} \widehat{Q} \|\theta - \theta^*\|^2 \lesssim n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^{\frac{2\alpha}{2\alpha+1}},$$

where $\widehat{Q}$ *is the variational posterior defined by (2.3) with* $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}$*.*

It is well known that the minimax rate of estimating θ^* in $\Theta_\alpha(B)$ is $n^{-\frac{2\alpha}{2\alpha+1}}$ [38]. Using a mean-field variational posterior, we achieve the minimax rate up to a logarithmic factor. In fact, the following proposition demonstrates that this rate cannot be improved for a very general class of priors.

Proposition 2.3.1. *Consider the prior* Π *specified in (2.12). Assume that* $\max_j \|f_j\|_\infty \le a$ *and* $\pi(k)$ *is nonincreasing over* k*. Then, we have*

$$\sup_{\theta^* \in \Theta_\alpha(B)} P_{\theta^*}^{(n)} \widehat{Q} \|\theta - \theta^*\|^2 \gtrsim n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^{\frac{2\alpha}{2\alpha+1}},$$

where $\widehat{Q}$ *is the variational posterior defined by (2.3) with* $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}$*.*

2.3.2 *Infinite Dimensional Exponential Families*

In this section, we study another interesting variational family. The Gaussian mean-field family is defined as

$$\mathcal{S}_{\mathrm{G}} = \left\{ Q = \otimes_j N(\mu_j, \sigma_j^2) : \mu_j \in \mathbb{R}, \sigma_j^2 \geq 0 \right\}. \tag{2.18}$$

This class offers better interpretability of the results because every distribution in $\mathcal{S}_G$ is fully determined by a sequence of mean and variance parameters. Note that we allow σ_j^2 to be zero and $N(\mu_j, 0)$ is understood as the delta measure δ_{μ_j} on μ_j.

The application of $\mathcal{S}_G$ is illustrated by an infinite dimensional exponential family model. We define the probability measure P_θ by

$$\frac{dP_\theta}{d\ell} = \exp\left(\sum_{j=0}^{\infty} \theta_j \phi_j - c(\theta) \right), \tag{2.19}$$

where ℓ denotes the Lebesgue measure on $[0,1]$, ϕ_j is the jth Fourier basis function of $L^2[0,1]$, and $c(\theta)$ is given by

$$c(\theta) = \log \int_0^1 \exp\left(\sum_{j=0}^{\infty} \theta_j \phi_j(x) \right) dx.$$

Since $\phi_0(x) = 1$ and θ_0 can take arbitrary values without changing P_θ, we simply set $\theta_0 = 0$. In other words, P_θ is fully parameterized by $\theta = (\theta_1, \theta_2, ...)$. Given i.i.d. observations from $P_{\theta^*}^n$, our goal is to estimate P_{θ^*}, where θ^* is assumed to belong to the Sobolev ball $\Theta_\alpha(B)$ defined in (2.11). The loss function is chosen as n times the squared Hellinger distance $L(P_\theta^n, P_{\theta^*}^n) = nH^2(P_\theta, P_{\theta^*})$.

We consider a prior distribution Π that is similar to the one used in Section 2.3.1. Its sampling process is described as follows.

1. Sample $k \sim \pi$;

2. Conditioning on k, sample $\theta_j \sim f_j$ for all $j \in [k]$, and set $\theta_j = 0$ for all $j > k$.

We impose the following conditions on the prior Π.

- There exist some constants $C_1, C_2 > 0$ such that

$$\sum_{j=k}^{\infty} \pi(j) \leq C_1 \exp(-C_2 k \log k), \text{ for all } k. \tag{2.20}$$

- There exist some constants $C_3, C_4 > 0$ such that for $k_0 = \left\lceil \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}} \right\rceil$

$$\pi(k_0) \geq C_3 \exp(-C_4 k_0 \log k_0). \tag{2.21}$$

- There exist some constants $c_0 \in \mathbb{R}$ and $c_1, \beta > 0$ such that

$$-\log f_j(x) \geq c_0 + c_1 |x|^{\beta}, \tag{2.22}$$

  for all $x \in \mathbb{R}$ and $j \in [k_0]$ with k_0 defined above.

- For the k_0 defined above, there exist some constants $c_0' \in \mathbb{R}$ and $c_1' > 0$ such that

$$-\log f_j(x) \leq c_0' + c_1' j^{2\alpha+1} x^2, \qquad \text{for all } j \leq k_0 \text{ and } x \in \mathbb{R}. \tag{2.23}$$

The conditions (2.20)-(2.23) are satisfied by a large class of prior distributions. For example, one can choose $k \sim \text{Poisson}(\tau)$ and f_j being the density of $N(0, \sigma^2)$ for some constants $\tau, \sigma^2 > 0$, and then the four conditions are easily satisfied.

Theorem 2.3.3. *Consider the prior Π that satisfies (2.20)-(2.23). Then, for any $\theta^* \in \Theta_\alpha(B)$ with some $\alpha > 1/2$, we have*

$$P_{\theta^*}^n \widehat{Q} H^2(P_\theta, P_{\theta^*}) \lesssim n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^{\frac{2\alpha}{2\alpha+1}},$$

where $\widehat{Q}$ is the variational posterior defined by (2.3) with $\mathcal{S} = \mathcal{S}_{\mathrm{G}}$.

The theorem shows that the Gaussian mean-field variational posterior is able to achieve the minimax rate $n^{-\frac{2\alpha}{2\alpha+1}}$ up to a logarithmic factor. We remark that the same result also holds for the mean-field variational posterior defined with $\mathcal{S}_{\mathrm{MF}}$. This is because $\mathcal{S}_{\mathrm{G}} \subset \mathcal{S}_{\mathrm{MF}}$, and thus $\inf_{Q\in\mathcal{S}_{\mathrm{MF}}} R(Q) \leq \inf_{Q\in\mathcal{S}_{\mathrm{G}}} R(Q)$. Compared with the class $\mathcal{S}_{\mathrm{MF}}$, the objective function using the parametric family $\mathcal{S}_{\mathrm{G}}$ can be optimized by algorithms such as stochastic gradient descent over the parameters (μ_j, σ_j^2).

2.3.3 Piecewise Constant Model

The previous two examples consider the mean-field variational set and its variant. In this section, we use another example to illustrate a situation where the mean-field variational set only gives a trivial rate. On the other hand, we show that alternative variational classes with appropriate dependence structures are able to achieve the optimal rate.

We consider the following piecewise constant model,

$$X_i = \theta_i + \sigma Z_i, \quad i \in [n], \tag{2.24}$$

where $Z_i \sim N(0,1)$ independently for all $i \in [n]$. We assume $n \geq 2$ throughout the section. The true parameter θ^* is assumed to belong to the class $\Theta_{k^*}(B) = \{\theta \in \Theta_{k^*} : \|\theta\|_\infty \leq B\}$, where for a general $k \in [n]$,

$$\begin{aligned}\Theta_k \;=\; \Big\{ & \theta \in \mathbb{R}^n : \text{ there exist } \{a_j\}_{j=0}^k \text{ and } \{\mu_j\}_{j=1}^k \text{ such that} \\ & 0 = a_0 \leq a_1 \leq \cdots \leq a_k = n, \text{ and } \theta_i = \mu_j \text{ for all } i \in (a_{j-1} : a_j] \Big\}. \end{aligned} \tag{2.25}$$

Here for any two integers $a < b$, we use $(a : b]$ to denote all integers from $a+1$ to b. We assume both $B > 0$ and $\sigma^2 > 0$ are constants throughout this section. A vector $\theta^* \in \Theta_{k^*}(B)$ is a piecewise constant signal with at most k^* pieces. We use $P_\theta^{(n)}$ to denote the probability distribution of $N(\theta, \sigma^2 I_n)$ in this section.

The piecewise constant model is widely studied in the literature of change-point analysis. Recently, the minimax rate of the class Θ_{k^*} is derived by [26]. When $2 < k^* \leq n^{1-\delta}$ for some constant $\delta \in (0,1)$, the minimax rate is $\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta_{k^*}} \mathbb{E}_{\theta^*}^{(n)} \|\widehat{\theta} - \theta^*\|^2 \asymp k^* \log n$. With an extra constraint on the infinity norm, the minimax rate for $\Theta_{k^*}(B)$ is still $k^* \log n$, with a slight modification of the proof in [26]. Since $D_\rho(P_\theta^{(n)}, P_{\theta'}^{(n)}) = \frac{\rho}{2\sigma^2}\|\theta - \theta'\|^2$ in this case, it is natural to choose the loss function as $L(P_\theta^{(n)}, P_{\theta^*}^{(n)}) = \|\theta - \theta^*\|^2$.

We put a prior distribution Π on the parameter θ. Consider Π that has the following sampling process.

1. Sample $w \sim \text{Beta}(\alpha_0, \beta_0)$;

2. Conditioning on w, sample $z_i \sim \text{Bernoulli}(w)$ for $i = 2, 3, ..., n$;

3. Conditioning on $(z_2, ..., z_n)$, sample $\theta_1 \sim g$, and then for $i = 2, 3, ..., n$, sample θ_i according to $\theta_i \sim g$ if $z_i = 1$ and $\theta_i = \theta_{i-1}$ if $z_i = 0$.

We first consider variational inference via the mean-field class, defined as

$$\mathcal{S}_{\text{MF}} = \left\{ Q : dQ(\theta) = \prod_{i=1}^{n} dQ_i(\theta_i) \right\}.$$

We also define $\mathcal{S} = \mathcal{S}_{\text{MF}}^{\text{joint}}$ on the joint distribution of (w, z, θ) by

$$\begin{aligned} \mathcal{S}_{\text{MF}}^{\text{joint}} \quad = \quad & \Bigg\{ Q : dQ(w,z,\theta) = dQ^{(w)}(w) dQ^{(z)}(z) dQ^{(\theta)}(\theta), \\ & \qquad dQ^{(z)}(z) = \prod_{i=2}^{n} dQ_i^{(z)}(z_i), Q^{(\theta)} \in \mathcal{S}_{\text{MF}} \Bigg\}. \end{aligned}$$

The variational posteriors $\widehat{Q}_{\text{MF}}$ and $\widehat{Q}_{\text{MF}}^{\text{joint}}$ are given by (2.3) with variational classes defined above respectively[3]. Interestingly, for the piecewise constant model, both $\widehat{Q}_{\text{MF}}$ and $\widehat{Q}_{\text{MF}}^{\text{joint}}$ give a trivial rate.

3. To be rigorous, the posterior distribution $\Pi(\cdot|X^{(n)})$ used in $D(Q\|\Pi(\cdot|X^{(n)}))$ are the marginal posterior of θ and the joint posterior of (w, z, θ), respectively.

Theorem 2.3.4. *For the prior* Π *specified above with any* g *absolutely continuous with respect to the Lebesgue measure, we have*

$$\sup_{\theta^* \in \Theta_{k^*}(B)} P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MF}} \|\theta - \theta^*\|^2 = \sup_{\theta^* \in \Theta_{k^*}(B)} P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MF}}^{\mathrm{joint}} \|\theta - \theta^*\|^2 \gtrsim n,$$

for any $k^* \in [n]$*, where* $\widehat{Q}_{\mathrm{MF}}$ *and* $\widehat{Q}_{\mathrm{MF}}^{\mathrm{joint}}$ *are the variational posteriors defined by (2.3) with* $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}$ *and* $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}^{\mathrm{joint}}$*, respectively.*

The result of Theorem 2.3.4 shows that the mean-field variational posteriors $\widehat{Q}_{\mathrm{MF}}$ and $\widehat{Q}_{\mathrm{MF}}^{\mathrm{joint}}$ are unable to achieve a better rate than simply estimating θ^* by the naive estimator $\widehat{\theta} = X$. The proof, given in Subsection 5.1.5, reveals the reason of this phenomenon. Since the independence structure of the two classes fails to capture the underlying dependence structure of the parameter space $\Theta_{k^*}(B)$, the variational posterior distributions are equivalent to the posterior distribution induced by the prior $\Pi = \otimes_{i=1}^n g$, and therefore the condition (C4) is violated. Note that this is the first negative result in the literature on the statistical convergence of the mean-field approximation.

In order to achieve the minimax rate of the space $\Theta_{k^*}(B)$, it is necessary to introduce some dependence structure in the variational class. One of the simplest classes of dependent distributions is the class of first-order Markov chains, defined by

$$\mathcal{S}_{\mathrm{MC}} = \left\{ Q : dQ(\theta) = dQ_1(\theta_1) \prod_{i=2}^{n} dQ_i(\theta_i | \theta_{i-1}) \right\}.$$

The class $\mathcal{S}_{\mathrm{MC}}$ introduces a natural dependence structure for the piecewise constant model, and it is compatible with the prior distribution Π, because conditioning on the change point pattern z, the prior distribution of $\theta | z$ belongs to the class $\mathcal{S}_{\mathrm{MC}}$. We also introduce a similar variational class on the joint distribution of (w, z, θ), defined by

$$\mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}} = \left\{ Q : dQ(w, z, \theta) = dQ^{(w)}(w) dQ^{(z)}(z) dQ^{(\theta)}(\theta), \right.$$

$$dQ^{(z)}(z) = \prod_{i=2}^{n} dQ_i^{(z)}(z_i), Q^{(\theta)} \in \mathcal{S}_{\mathrm{MC}} \Bigg\}.$$

Besides the distribution of θ restricted to $\mathcal{S}_{\mathrm{MC}}$, the distributions of w and z are both in the mean-field classes.

In order to derive the rates for the variational posterior distributions induced by $\mathcal{S}_{\mathrm{MC}}$ and $\mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$, we impose the following conditions on the prior distribution Π.

- There exist some constants $C_2 > C_1 > 1$ such that

$$(n + \alpha_0)n^{C_1} \leq \beta_0 \leq \alpha_0 n^{C_2} - n. \tag{2.26}$$

- There exists a constant $c > 0$ such that

$$g(x) \geq c, \text{ for all } |x| \leq B + 1. \tag{2.27}$$

According to Theorem 2.2.2, we get the following result.

Theorem 2.3.5. *Consider a prior distribution* Π *that satisfies (2.26) and (2.27). Then, for any* $\theta^* \in \Theta_{k^*}(B)$*, we have*

$$P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MC}} \|\theta - \theta^*\|^2 \lesssim k^* \log n,$$

$$P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}} \|\theta - \theta^*\|^2 \lesssim k^* \log n,$$

where $\widehat{Q}_{\mathrm{MC}}$ *and* $\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}$ *are the variational posterior distributions defined by (2.3) with* $\mathcal{S} = \mathcal{S}_{\mathrm{MC}}$ *and* $\mathcal{S} = \mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$*, respectively.*

Theorem 2.3.5 shows that both $\widehat{Q}_{\mathrm{MC}}$ and $\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}$ are able to achieve the minimax rate of the problem. This example illustrates the importance of the choice of the variational class. According to Theorem 2.2.1, the rate of a variational posterior is upper bounded by ϵ_n^2, the

rate of the true posterior, plus γ_n^2, the variational approximation error. The choice of $\mathcal{S}_{\mathrm{MF}}$ for the piecewise constant model leads to a very large γ_n^2, and thus a trivial rate in Theorem 2.3.4. On the other hand, the variational approximation errors given by the classes $\mathcal{S}_{\mathrm{MC}}$ and $\mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$ are small, which are dominated by the minimax rate.

Though the statistical properties of the two classes $\mathcal{S}_{\mathrm{MC}}$ and $\mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$ are both satisfactory, the class $\mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$ enjoys a computational advantage, and the solution $\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}$ can be computed exactly via dynamic programming. In order to characterize the solution $\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}$, we consider the following discrete optimization problem:

$$\max_{1\le k\le n}\left\{\max_{0=a_0<a_1<\cdots<a_k=n}\sum_{j=1}^{k}\log\int g(\theta)\exp\left(-\frac{1}{2}\sum_{i\in(a_{j-1}:a_j]}(X_i-\theta)^2\right)d\theta\right.$$
$$\left.+\log\left(\Gamma(k-1+\alpha_0)\Gamma(n-k+\beta_0)\right)\right\}. \tag{2.28}$$

The solution of (2.28) is denoted as the sequence $0=\widehat{a}_0<\widehat{a}_1<\cdots<\widehat{a}_{\widehat{k}}=n$. We remark that under the condition (2.26), the penalty term of (2.28) comes from the fact that

$$-\log\frac{\Gamma(k-1+\alpha_0)\Gamma(n-k+\beta_0)\Gamma(\alpha_0+\beta_0)}{\Gamma(n-1+\alpha_0+\beta_0)\Gamma(\alpha_0)\Gamma(\beta_0)}\asymp k\log n,$$

which coincides with the minimax rate.

Theorem 2.3.6. *Let the maximizer of (2.28) be* $(\widehat{a}_0,\widehat{a}_1,...,\widehat{a}_{\widehat{k}})$. *For* $d\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}(w,z,\theta)=d\widehat{Q}^{(w)}(w)d\widehat{Q}^{(z)}(z)d\widehat{Q}^{(\theta)}(\theta)$, *the distributions* $\widehat{Q}^{(w)}$, $\widehat{Q}^{(z)}$ *and* $\widehat{Q}^{(\theta)}$ *are specified as follows.*

1. *Under* $\widehat{Q}^{(z)}$, $z_{\widehat{a}_j+1}=1$ *for* $j=1,...,\widehat{k}-1$, *and* $z_i=0$ *elsewhere with probability* 1.
2. *We have* $\widehat{Q}^{(w)}=\mathrm{Beta}(\widehat{k}+\alpha_0-1,n-\widehat{k}+\beta_0)$.

3. We have $d\widehat{Q}^{(\theta)}(\theta) = d\widehat{Q}_1^{(\theta)}(\theta_1)\prod_{i=2}^n d\widehat{Q}_i^{(\theta)}(\theta_i|\theta_{i-1})$*, where*

$$\begin{cases} d\widehat{Q}_1^{(\theta)}(\theta_1) \propto g(\theta_1)\exp\left(-\frac{1}{2}\sum_{i\in(\widehat{a}_0:\widehat{a}_1]}(X_i-\theta_1)^2\right)d\theta_1, \\ d\widehat{Q}_i^{(\theta)}(\theta_i|\theta_{i-1}) \propto g(\theta_i)\exp\left(-\frac{1}{2}\sum_{l\in(\widehat{a}_{j-1}:\widehat{a}_j]}(X_l-\theta_i)^2\right)d\theta_i, & i=\widehat{a}_{j-1}+1, j>1, \\ d\widehat{Q}_i^{(\theta)}(\theta_i|\theta_{i-1}) = \delta_{\theta_{i-1}}(\theta_i)d\theta_i, & \text{otherwise.} \end{cases}$$

By Theorem 2.3.6, in order to get $\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}$, it is sufficient to solve (2.28). This can be done through a dynamic programming given in Algorithm 2.3.3. To simplify the notation, we define

$$S_{(a:b]} = \log\int g(\theta)\exp\left(-\frac{1}{2}\sum_{i\in(a:b]}(X_i-\theta)^2\right)d\theta, \tag{2.29}$$

for any integers $0 \le a < b \le n$.

Algorithm 1 Computation of (2.28)

Input : The data $X_1, ..., X_n$.
for $j = 1, 2, \ldots, n$ **do**
  set $A_{1,j} = \emptyset$, and compute $B_{1,j} = S_{(0:j]}$
end for
for $k = 2, \ldots, n$ **do**
  for $j = k, \ldots, n$ **do**
    Compute
      $B_{k,j} = \max_{k-1\le m\le j-1}\left\{B_{k-1,m} + S_{(m:j]}\right\}$,
      $a_{k,j} = \mathrm{argmax}_{k-1\le m\le j-1}\left\{B_{k-1,m} + S_{(m:j]}\right\}$,
      $A_{k,j} = A_{k-1,a_{k,j}} \cup \left\{a_{k,j}\right\}$,
  end for
end for
Compute
  $\widehat{k} = \mathrm{argmax}_{1\le k\le n}\left\{B_{k,n} + \log(\Gamma(k-1+\alpha_0)\Gamma(n-k+\beta_0))\right\}$.
Output : The set of knots $A_{\widehat{k},n} = \left\{\widehat{a}_1, \ldots, \widehat{a}_{\widehat{k}-1}\right\}$.

We note that the computational cost of the dynamic programming above is $O(n^3)$ (see [25]), and for any integers $0 \le a < b \le n$, (2.29) has a closed form as long as we use a conjugate $g(\cdot)$.

2.4 Variational Bayes with Model Selection

2.4.1 General Settings

In this section, we consider a general form of probability models

$$\mathcal{M} = \left\{ P^{(n)}_{k,\theta^{(k)}} : k \in \mathcal{K}, \theta^{(k)} \in \Theta^{(k)} \right\}.$$

Here, the probability $P^{(n)}_{k,\theta^{(k)}}$ is determined by an index k and a parameter $\theta^{(k)}$. We assume that the set $\mathcal{K}$ is either countable or finite. For a given k, the probability $P^{(n)}_{k,\theta^{(k)}}$ is parametrized by a $\theta^{(k)}$ in a parameter space $\Theta^{(k)}$ that is indexed by this k. Without loss of generality, we assume that the parameter $\theta^{(k)}$ can be written in a blockwise structure

$$\theta^{(k)} = (\theta^{(k)}_1, \cdots, \theta^{(k)}_{m_k}).$$

Note that the dimension of $\theta^{(k)}$ may vary with k.

The model $\mathcal{M}$ is very natural for many applications. One can think of k as a model dimension index, which determines the complexity of the parameter space $\Theta^{(k)}$. A leading example is the mixture density model, where k stands for the number of components.

To model the hierarchical structure of $(k, \theta^{(k)})$, one naturally uses a hierarchical prior distribution, which is specified through the following sampling process:

1. Firstly, sample $k \sim \pi$ from $\mathcal{K}$;

2. Conditioning on k, sample $\theta^{(k)}$ from the probability measure $\Pi^{(k)}$, and $\Pi^{(k)}$ has a product structure
$$d\Pi^{(k)}(\theta^{(k)}) = \prod_{j=1}^{m_k} d\Pi^{(k)}_j(\theta^{(k)}_j). \tag{2.30}$$

For variational inference, we consider a mean-field class that naturally takes advantage of the structure of the prior distribution. For a given $k \in \mathcal{K}$, the corresponding mean-field

class is defined as

$$\mathcal{S}_{\mathrm{MF}}^{(k)} = \left\{ Q^{(k)} : dQ^{(k)}(\theta^{(k)}) = \prod_{j=1}^{m_k} dQ_j^{(k)}(\theta_j^{(k)}) \right\}. \tag{2.31}$$

In order to select the best model from the data, we consider optimizing the evidence lower bound (ELBO). With the notation $p(X^{(n)}|\theta^{(k)})$ standing for the joint likelihood function, the marginal likelihood given a model $k \in \mathcal{K}$ is defined by

$$p(X^{(n)}|k) = \int p(X^{(n)}|\theta^{(k)}) d\Pi^{(k)}(\theta^{(k)}). \tag{2.32}$$

Then, a straightforward model selection procedure is to maximize $\log\left(p(X^{(n)}|k)\pi(k)\right)$ over $k \in \mathcal{K}$. In order to overcome the intractability of the integral (2.32), we instead optimize a lower bound, which is given by

$$\begin{aligned} & \log\left(p(X^{(n)}|k)\pi(k)\right) \\ \geq \; & \int \log p(X^{(n)}|\theta^{(k)}) dQ^{(k)}(\theta^{(k)}) - D\left(Q^{(k)} \| \Pi^{(k)}\right) + \log \pi(k), \end{aligned} \tag{2.33}$$

which can be derived by a direct application of Jensen's inequality. Denote the right hand side of (2.33) by $F(Q^{(k)}, k)$, and we will solve the following optimization problem,

$$\max_{k \in \mathcal{K}} \max_{Q^{(k)} \in \mathcal{S}_{\mathrm{MF}}^{(k)}} F(Q^{(k)}, k). \tag{2.34}$$

Finally, the solution to (2.34) leads to the variational posterior distribution $\widehat{Q} = \widehat{Q}^{(\widehat{k})}$ that we use in a model selection context. A similar variational approximation to the tempered posterior in the model selection setting was studied by [19].

2.4.2 Convergence Rates

Assume the observation $X^{(n)}$ is generated from a probability measure $P_0^{(n)}$, and $\widehat{Q} = \widehat{Q}^{(\widehat{k})}$ is the variational posterior that is a solution to (2.34). For the general settings described above, we show that the variational approximation error can be automatically controlled by a prior mass condition. Let Π be the prior distribution on $P_{k,\theta^{(k)}}$ induced by the sampling process of $(k, \theta^{(k)})$.

Theorem 2.4.1. *Suppose ϵ_n is a sequence that satisfies $n\epsilon_n^2 \geq 1$. Let $\rho > 1$ be a constant and $C_2, C_3 > 0$ be constants. We assume that there exists a $k_0 \in \mathcal{K}$ and a subset $\Theta^{(k_0)} = \otimes_{j=1}^{m_{k_0}} \Theta_j^{(k_0)} \subset \left\{\theta^{(k_0)} : D_\rho\left(P_0^{(n)} \| P^{(n)}_{k_0,\theta^{(k_0)}}\right) \leq C_3 n\epsilon_n^2\right\}$, such that*

$$-\log \pi(k_0) - \sum_{j=1}^{m_{k_0}} \log \Pi_j^{(k_0)}\left(\Theta_j^{(k_0)}\right) \leq C_2 n\epsilon_n^2, \tag{C3*}$$

where $\pi(k_0)$ and $\Pi_j^{(k_0)}$ are defined in the prior sampling procedure. Moreover, assume that the conditions (C1) and (C2) hold for all $\epsilon > \epsilon_n$ with respect to prior procedure Π and some constant $C > C_2 + C_3 + 2$. Then for the variational posterior $\widehat{Q}^{(\widehat{k})}$ defined as the solution of (2.34), we have

$$P_0^{(n)} \widehat{Q}^{(\widehat{k})} L(P^{(n)}_{\widehat{k},\theta^{(\widehat{k})}}, P_0^{(n)}) \lesssim n\epsilon_n^2. \tag{2.35}$$

Theorem 2.4.1 characterizes the convergence rate of mean-field variational posterior with model selection using the conditions (C1), (C2) and (C3*). Given the structure of the prior distribution, an equivalent way of writing (C3*) is

$$\Pi\left(\left\{P_{k,\theta^{(k)}} : k = k_0, \theta^{(k_0)} \in \Theta^{(k_0)}\right\}\right) \geq \exp\left(-C_2 n\epsilon_n^2\right),$$

for the factorized structure of $\Theta^{(k_0)}$. Therefore, our three conditions (C1), (C2) and (C3*) still fall into the "prior mass and testing" framework, and directly correspond to the three conditions in [31] for convergence rates of the true posterior.

An interesting special case is when the set $\mathcal{K}$ is a singleton. Then, for a product prior measure and the mean-field variational class, the condition (C3*) is reduced to (2.2) discussed in Section 2.1.

2.4.3 Density Estimation via Location-Scale Mixtures

In this section, we consider the location-scale mixture model as an application of the theory. The location-scale mixture density is defined as

$$p(x|k,\theta^{(k)}) = \sum_{j=1}^{k} w_j \psi_\sigma(x-\mu_j), \tag{2.36}$$

where $k \in \mathbb{N}_+$, $\theta^{(k)} = (\mu, w, \sigma)$ with $\sigma > 0$, $\mu = (\mu_1, \cdots, \mu_k) \in \mathbb{R}^k$, $w = (w_1, \cdots, w_k) \in \Delta_k = \left\{ w \in \mathbb{R}^k : w_j \geq 0 \text{ for } 1 \leq j \leq k \text{ and } \sum_{j=1}^k w_j = 1 \right\}$ and

$$\psi_\sigma(x) = \frac{1}{2\sigma\Gamma\left(1+\frac{1}{p}\right)} \exp(-(|x|/\sigma)^p), \tag{2.37}$$

for some positive even integer p. The kernel $\psi_\sigma(\cdot)$ has a pre-specified form, for example, Gaussian density when $p = 2$, while the parameters k and $\theta^{(k)} = (w, \mu, \sigma)$ are to be learned from the data.

The location-scale mixture model (2.36) can be written as a special example of the general probability models introduced in Section 2.4.1. In this case, the countable set $\mathcal{K}$ is the positive integer set $\mathbb{N}_+$. The parameter space indexed by k is defined as

$$\begin{aligned} \Theta^{(k)} \;=\; \Big\{ & \theta^{(k)} = (\mu, w, \sigma) : \mu = (\mu_1, \cdots, \mu_k) \in \mathbb{R}^k, \\ & w = (w_1, \cdots, w_k) \in \Delta_k, \sigma \in \mathbb{R}_+ \Big\}. \end{aligned} \tag{2.38}$$

Given i.i.d. observations $X_1, ..., X_n$ sampled from some density function f_0, our goal is to estimate the density f_0 through the location-scale mixture model (2.36). We denote

the probability distribution of the mixture density $p(x|k,\theta^{(k)})$ as $P_{k,\theta^{(k)}}$ and a probability distribution with a general density f as P_f. In the paper [42], a Bayesian procedure is proposed and a nearly minimax optimal convergence rate is derived for the true posterior distribution. We will follow the same setting in [42], but analyze the variational posterior.

We first specify the prior distribution Π through the following sampling process:

1. Sample the number of mixtures $k \sim \pi$;

2. Conditioning on k, sample the location parameters $\mu_1, \cdots, \mu_k$ independently from p_μ, sample the weights $w = (w_1, \cdots, w_k)$ from $p_w^{(k)}$, and then sample the precision parameter $\tau = \sigma^{-2}$ from p_τ.

In order to optimize (2.34) in the variational Bayes framework, we specify the blockwise structure (2.31) in this case as

$$\mathcal{S}_{\mathrm{MF}}^{(k)} = \left\{ Q^{(k)} : dQ^{(k)}(\theta^{(k)}) = dQ_\sigma(\sigma) dQ_w^{(k)}(w) \prod_{j=1}^{k} dQ_{\mu_j}(\mu_j) \right\}. \tag{2.39}$$

Note that we do not factorize $dQ_w^{(k)}(w)$ because of the constraint $\sum_{j=1}^k w_j = 1$. The variational posterior distribution is defined as $\widehat{Q} = \widehat{Q}^{(\widehat{k})}$ that solves (2.34). The loss function here is chosen as n times squared Hellinger distance, i.e., $L(P_f^n, P_{f_0}^n) = nH^2(P_f, P_{f_0})$.

In order that $\widehat{Q}$ enjoys a good convergence rate, we need conditions on the prior distribution and the true density function f_0. We first list the conditions on the prior.

1. There exist constants $C_1, C_2 > 0$, such that

$$\sum_{m=k}^{\infty} \pi(m) \le C_1 \exp(-C_2 k \log k), \tag{2.40}$$

   for all $m > 0$. There exist constants $t, C_3, C_4 > 0$, such that

$$\pi(k_0) \ge C_3 \exp(-C_4 k_0 \log k_0), \tag{2.41}$$

for all $n^{\frac{1}{2\alpha+1}} \leq k_0 \leq n^{\frac{1}{2\alpha+1}+t}$.

2. There exist constants $c_1, c_2, c_3 > 0$, such that

$$\int_{-\infty}^{-x_0} p_\mu(x)dx + \int_{x_0}^{\infty} p_\mu(x)dx \leq c_1 \exp(-c_2 x_0^{c_3}), \tag{2.42}$$

for all $x_0 > 0$ and constants c_4, c_5, c_6, such that

$$p_\mu(x) \geq c_4 \exp(-c_5|x|^{c_6}), \tag{2.43}$$

for all x.

3. There exist constants $t, d_1, d_2, d_3 > 0$, such that

$$\int_{w \in \Delta_{k_0}(w_0,\epsilon)} p_w^{(k_0)}(x)dx \geq d_1 \exp\left(-d_2 k_0 (\log k_0)^{d_3} \log\left(\frac{1}{\epsilon}\right)\right), \tag{2.44}$$

for all $w_0 \in \Delta_{k_0}$ and $n^{\frac{1}{2\alpha+1}} \leq k_0 \leq n^{\frac{1}{2\alpha+1}+t}$, where $\Delta_{k_0}(w_0, \epsilon) = \{w \in \Delta_{k_0} : \|w - w_0\|_1 \leq \epsilon\}$.

4. There exist constants $b_0, b_1, b_2, b_3 > 0$, such that

$$\|p_\tau\|_\infty < b_0, \qquad \int_{\tau_0}^{\infty} p_\tau(x)dx \leq b_1 \exp(-b_2|\tau_0|^{b_3}), \tag{2.45}$$

for all $\tau_0 > 0$. There exist constants $b_4, b_5 > 0$ and a constant $b_6 \in (0, 1]$ that satisfy

$$p_\tau(x) \geq b_4 \exp(-b_5|x|^{b_6}), \tag{2.46}$$

for all $x > 0$.

The conditions on the prior distribution are quite general. For example, one can choose $k \sim \text{Poisson}(\xi_0)$, $\mu_j \sim N(0, \sigma_0^2)$, $w \sim \text{Dir}(\alpha_0, \alpha_0, \cdots, \alpha_0)$ and $\tau \sim \Gamma(a_0, b_0)$ for some positive

constants $\xi_0, \sigma_0, \alpha_0, a_0, b_0$. Then, the conditions above are all satisfied.

Next, we list the conditions on the true density function f_0:

B1 (Smoothness) The logarithmic density function $\log f_0$ is assumed to be locally α-Hölder smooth. In other words, for the derivative $l_j(x) = \frac{d^j}{dx^j} \log f_0(x)$, there exists a polynomial $L(\cdot)$ and a constant $\gamma > 0$ such that,

$$|l_{\lfloor\alpha\rfloor}(x) - l_{\lfloor\alpha\rfloor}(y)| \leq L(x)|x - y|^{\alpha - \lfloor\alpha\rfloor}, \tag{2.47}$$

for all x, y that satisfies $|x - y| \leq \gamma$. Here, the degree and the coefficients of the polynomial $L(\cdot)$ are all assumed to be constants. Moreover, the derivative $l_j(x)$ satisfies the bound $\int |l_j(x)|^{\frac{2\alpha+\epsilon}{j}} f_0(x)dx < s_{\max}$ for all $j = 1, ..., \lfloor\alpha\rfloor$ with some constants $\epsilon, s_{\max} > 0$.

B2 (Tail) There exist positive constants T, ξ_1, ξ_2, ξ_3 such that

$$f_0(x) \leq \xi_1 e^{-\xi_2 |x|^{\xi_3}}, \tag{2.48}$$

for all $|x| \geq T$.

B3 (Monotonicity) There exist constants $x_m < x_M$ such that f_0 is nondecreasing on $(-\infty, x_m)$ and is nonincreasing on (x_M, ∞). Without loss of generality, we assume $f_0(x_m) = f_0(x_M) = c$ and $f_0(x) \geq c$ for all $x_m < x < x_M$ with some constant $c > 0$.

These conditions are exactly the same as in [42] and similar conditions are also considered in [45]. The conditions allow a well-behaved approximation to the true density by a location-scale mixture. There are many density functions that satisfy the conditions (B1)-(B3), for which we refer to [42].

The convergence rate of the variational posterior is given by the following theorem.

Theorem 2.4.2. *Consider i.i.d. observations generated by $P_{f_0}^n$, and the density function f_0 satisfies conditions (B1)-(B3). For the prior that satisfies (2.40)-(2.46), we have*

$$P_{f_0}^n \widehat{Q} H^2(P_{\widehat{k},\theta^{(\widehat{k})}}, P_{f_0}) \lesssim n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^{\frac{2\alpha r}{2\alpha+1}},$$

where $\widehat{Q} = \widehat{Q}^{(\widehat{k})}$ is the solution of (2.34), and $r = \frac{p}{\min\{p,\xi_3\}} + \max\{d_3+1, \frac{c_6}{\min\{p,\xi_3\}}\}$, with p, ξ_3, c_6, d_3 defined in (2.37), (2.48), (2.43) and (2.44), respectively.

The proof of Theorem 2.4.2 largely follows the arguments in [42] that are used to establish the corresponding result for the true posterior distribution, thanks to the fact that Theorem 2.4.1 requires three very similar "prior mass and testing" conditions to that of [31]. The only difference is that function approximations via location-scale mixtures need to be analyzed under a stronger divergence $D_\rho(\cdot\|\cdot)$ for some $\rho > 1$. For this reason, the proof of Theorem 2.4.2 relies on the construction of a surrogate density function $\widetilde{f}_0$. We first apply Theorem 2.4.1 and establish a convergence rate under $\widetilde{f}_0$. Then, the conclusion is transferred to f_0 with a change-of-measure argument. .

2.4.4 Dealing with Latent Variables

For the mixture model considered in Section 2.4.3, we discuss a variation of the variational Bayes approach (2.34) by including latent variables. This facilitates computation and leads to a simple coordinate ascent algorithm that has closed-form updates. In the setting of mixture model, our approach is adaptive to the unknown number of components, and can be regarded as an extension of [69, 46] for variational inference with latent variables.

Since $p(X^{(n)}|k,\theta^{(k)}) = \prod_{i=1}^n \sum_{j=1}^k w_j \psi_\sigma(X_i - \mu_j)$ with $\theta^{(k)} = (\mu, w, \sigma)$, we can write

$$p(X^{(n)}|\theta^{(k)}) = \sum_{z^{(k)} \in [k]^n} p(X^{(n)}|z^{(k)}, \theta^{(k)}) w^{(k)}(z^{(k)}),$$

where $p(X^{(n)}|z^{(k)},\theta^{(k)}) = \prod_{i=1}^n \prod_{j=1}^k \psi_\sigma(X_i - \mu_j)^{\mathbb{I}\left\{z_i^{(k)}=j\right\}}$, and the probability of $z_i^{(k)} = j$

is w_j under $w^{(k)}(\cdot)$. We use the notation $\bar{\Pi}^{(k)}$ for the joint distribution of $(z^{(k)}, \theta^{(k)})$, and then the marginal likelihood (2.32) can be written as

$$p(X^{(n)}|k) = \int p(X^{(n)}|z^{(k)}, \theta^{(k)}) d\bar{\Pi}^{(k)}(z^{(k)}, \theta^{(k)}).$$

Similar to (2.33), the evidence lower bound with the latent variables is given by

$$\begin{aligned} & \log\Big(p(X^{(n)}|k)\pi(k)\Big) \\ \geq & \int \log p(X^{(n)}|z^{(k)}, \theta^{(k)}) d\bar{Q}^{(k)}(z^{(k)}, \theta^{(k)}) - D(\bar{Q}^{(k)} \| \bar{\Pi}^{(k)}) + \log \pi(k). \end{aligned} \tag{2.49}$$

The right hand side of (2.49) is shorthanded by $\bar{F}(\bar{Q}^{(k)}, k)$. Define

$$\bar{\mathcal{S}}^{(k)}_{\mathrm{MF}} = \left\{ \bar{Q}^{(k)} : d\bar{Q}^{(k)}(z^{(k)}, \theta^{(k)}) = \prod_{i=1}^{n} dQ_z^{(k)}(z_i) dQ_\sigma(\sigma) dQ_w^{(k)}(w) \prod_{j=1}^{k} dQ_{\mu_j}(\mu_j) \right\}.$$

Then, we solve the following optimization problem,

$$\max_k \max_{\bar{Q}^{(k)} \in \bar{\mathcal{S}}^{(k)}_{\mathrm{MF}}} \bar{F}(\bar{Q}^{(k)}, k). \tag{2.50}$$

The solution to (2.50) leads to the variational posterior distribution $\widehat{Q} = \widehat{Q}^{(\widehat{k})}_{\mathrm{latent}}$. It is worth noting that even though $\widehat{Q}$ is a joint distribution of (z, μ, w, σ), the posterior inference only relies on the marginal of (μ, w, σ), since the parametrization of the density $f(\cdot)$ in (2.36) does not depend on the latent variables. The existence of the latent variables only facilitates computation.

Theorem 2.4.3. *Consider i.i.d. observations generated by $P^n_{f_0}$, and the density function f_0 satisfies conditions (B1)-(B3). For the prior that satisfies (2.40)-(2.46), we have*

$$P^n_{f_0} \widehat{Q} H^2(P_{\widehat{k}, \theta^{(\widehat{k})}}, P_{f_0}) \lesssim n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^{\frac{2\alpha r}{2\alpha+1}},$$

where $\widehat{Q} = \widehat{Q}^{(\widehat{k})}_{\text{latent}}$ *is the solution to (2.50), and* $r = \frac{p}{\min\{p,\xi_3\}} + \max\{d_3+1, \frac{c_6}{\min\{p,\xi_3\}}\}$*, with* p, ξ_3, c_6, d_3 *defined in (2.37), (2.48), (2.43) and (2.44), respectively.*

Theorem 2.4.3 shows that the variational posterior with latent variables achieves the same contraction rate as in Theorem 2.4.2. In fact, the two variational lower bounds (2.33) and (2.49) satisfy the following relation,

$$\log\Big(p(X^{(n)}|k)\pi(k)\Big) \geq \max_{Q^{(k)}\in\mathcal{S}^{(k)}_{\text{MF}}} F(Q^{(k)},k) \geq \max_{\bar{Q}^{(k)}\in\bar{\mathcal{S}}^{(k)}_{\text{MF}}} \bar{F}(\bar{Q}^{(k)},k),$$

which implies that the introduction of latent variables makes the variational approximation looser. On the other hand, Theorem 2.4.3 shows that the worse variational approximation does not compromise the statistical convergence rate. Moreover, with the help of latent variables, $\widehat{Q}^{(\widehat{k})}_{\text{latent}}$ can be computed via standard variational inference algorithms. .

2.5 Discussion

2.5.1 Variational Approximation as Regularization

According to Theorem 2.2.1, the convergence rate of the posterior is determined by the sum of ϵ_n^2, the rate of the true posterior, and γ_n^2, the variational approximation error. Since $\epsilon_n^2 + \gamma_n^2 \geq \epsilon_n^2$, it seems that the convergence rate of variational posterior is always no faster than that of the true posterior. However, Theorem 2.2.1 just gives an upper bound. In this section, we give two examples, and we show that it is possible for a variational posterior to have a faster convergence rate than that of the true posterior.

Example 1 We consider the setting of Gaussian sequence model (2.10). The true signal θ^* that generates the data is assumed to belong to the Sobolev ball $\Theta_\alpha(B)$. The prior

distribution is specified as

$$\theta \sim d\Pi = \prod_{j \le n} dN(0, j^{-2\beta-1}) \prod_{j>n} \delta_0.$$

Note that a similar Gaussian process prior is well studied in the literature [63, 13]. We force all the coordinates after n to be zero, so that the variational approximation through Kullback-Leibler divergence will not explode. For the specified prior, the posterior contraction rate is $n^{-\frac{2(\alpha\wedge\beta)}{2\beta+1}}$, and when $\beta = \alpha$, the optimal minimax rate $n^{-\frac{2\alpha}{2\alpha+1}}$ is achieved.

Consider the following variational class

$$\mathcal{S}_{[k]} = \left\{ Q : dQ = \prod_{j \le k} dQ_j \prod_{j=k+1}^{n} dN(0, e^{-jn}) \prod_{j>n} \delta_0 \right\},$$

for a given integer k. It is easy to see that the variational posterior $\widehat{Q}_{[k]}$ defined by (2.3) with $\mathcal{S} = \mathcal{S}_{[k]}$ can be written as

$$d\widehat{Q}_{[k]} = \prod_{j \le k} dN\left(\frac{n}{n + j^{2\beta+1}} Y_j, \frac{1}{n + j^{2\beta+1}}\right) \prod_{j=k+1}^{n} dN(0, e^{-jn}) \prod_{j>n} \delta_0.$$

In other words, the class $\mathcal{S}_{[k]}$ does not put any constraint on the first k coordinates and shrink all the coordinates after k to zero. Ideally, one would like to use δ_0 for the coordinates after k. However, that would lead to $D\left(Q \| \Pi(\cdot|Y)\right) = \infty$ for all $Q \in \mathcal{S}_{[k]}$ given that the support of δ_0 is a singleton. That is why we use $N(0, e^{-jn})$ instead. The rate of $\widehat{Q}_{[k]}$ for each k is given by the following theorem.

Theorem 2.5.1. *For the variational posterior $\widehat{Q}_{[k]}$, we have*

$$\sup_{\theta^* \in \Theta_\alpha(B)} \mathbb{P}_{\theta^*}^{(n)} \widehat{Q}_{[k]} \|\theta - \theta^*\|^2 \asymp \begin{cases} \frac{k}{n} + k^{-2\alpha}, & k \le n^{\frac{1}{2\beta+1}}, \\ n^{-\frac{2(\alpha\wedge\beta)}{2\beta+1}}, & k > n^{\frac{1}{2\beta+1}}, \end{cases}$$

where $\widehat{Q}_{[k]}$ is the variational posterior defined by (2.3) with $\mathcal{S} = \mathcal{S}_{[k]}$.

Note that Theorem 2.5.1 gives both upper and lower bounds for $\widehat{Q}_{[k]}$. This makes the comparison between variational posterior and true posterior possible. Observe that when $k = \infty$, we have $\widehat{Q}_{[\infty]} = \Pi(\cdot|Y)$, and the result is reduced to the posterior contraction rate $n^{-\frac{2(\alpha\wedge\beta)}{2\beta+1}}$ in [13].

Depending on the values of α, β and k, the rate for $\widehat{Q}_{[k]}$ can be better than that of the true posterior. For example, when $\beta < \alpha$, the choice $k = n^{\frac{1}{2\alpha+1}}$ leads to the minimax rate $n^{-\frac{2\alpha}{2\alpha+1}}$, which is always faster than $n^{-\frac{2(\alpha\wedge\beta)}{2\beta+1}}$. This is because for a $\beta < \alpha$, the true posterior distribution undersmooths the data, but the variational class $\mathcal{S}_{[k]}$ with $k = n^{\frac{1}{2\alpha+1}}$ helps to reduce the extra variance resulted from undersmoothing by thresholding all the coordinates after k. On the other hand, when $\beta \geq \alpha$, an improvement through the variational class $\mathcal{S}_{[k]}$ is not possible. In this case, the true posterior has already overly smoothed the data, and the information loss cannot be recovered by the variational class. In general, we plot the exponent value of the rate of $\widehat{Q}_{[k]}$ against the value of $\log_n(k)$ in Figure 2.2.

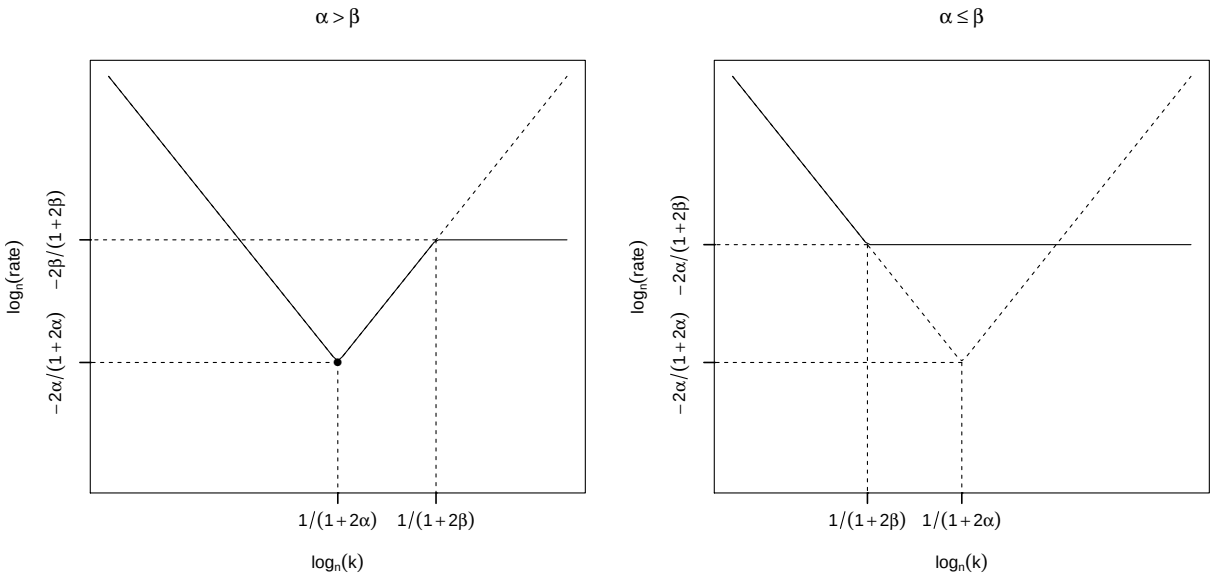

Figure 2.1: The exponent value of the rate of $\widehat{Q}_{[k]}$ against the value of $\log_n(k)$.

Example 2 Consider the problem of sparse linear regression $y \sim N(X\beta^*, I_n)$, where X is a design matrix of size $n \times p$ and β^* belongs to the sparse set $\mathcal{B}(s) = \{\beta \in \mathbb{R}^p : \sum_{j=1}^p \mathbb{I}_{\{\beta_j \neq 0\}} \leq s\}$ for some $s \in [p]$. The prior distribution on β is specified by the Laplace density

$$\frac{d\Pi(\beta)}{d\beta} = \prod_{j=1}^p \left(\frac{\lambda}{2} e^{-\lambda|\beta_j|}\right).$$

Though the posterior distribution has a close connection to LASSO, it is proved in [16] that the posterior distribution cannot adapt to the sparsity of β^*. In particular, the common choice of λ in the theoretical analysis of LASSO only leads to a dense posterior.

In fact, it is known in the literature (e.g. [9]) that the LASSO, which is the posterior mode, achieves a nearly optimal rate over the class $\mathcal{B}(s)$. We show that the posterior mode can be well approximated by applying a simple variational class. Consider the variational class

$$\mathcal{S}_{\tau^2} = \left\{N(\beta, \tau^2 I_p) : \beta \in \mathbb{R}^p\right\}.$$

Define $\widehat{Q}_{\tau^2}$ to be the minimizer of $\min_{Q \in \mathcal{S}_{\tau^2}} D(Q\|\Pi(\cdot|y))$.

Theorem 2.5.2. *For any $\lambda > 0$ and $\tau > 0$, we have $\widehat{Q}_{\tau^2} = N(\widehat{\beta}, \tau^2 I_p)$, where*

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} \left\{\frac{1}{2}\|y - X\beta\|^2 + \lambda \sum_{j=1}^p \tau h(\beta_j/\tau)\right\}. \tag{2.51}$$

The function h is defined by $h(x) = 2\phi(x) + x\left(\Phi(x) - \Phi(-x)\right)$ with $\Phi(x) = \mathbb{P}(N(0,1) \leq x)$ and $\phi(x) = \Phi'(x)$.

Theorem 2.5.2 shows that the variational approximation is characterized by the penalized least-squares estimator (2.51). Observe that h is a convex function, and it satisfies $\sup_{x \in \mathbb{R}} \left|\tau h(x/\tau) - |x|\right| = \tau\sqrt{\frac{2}{\pi}}$ (see Figure 2.2), and thus $\widehat{\beta}$ will get arbitrarily close to the LASSO estimator as $\tau \to 0$. Therefore, even though the posterior does not have a good frequentist property, its variational approximation can recover a sparse signal.

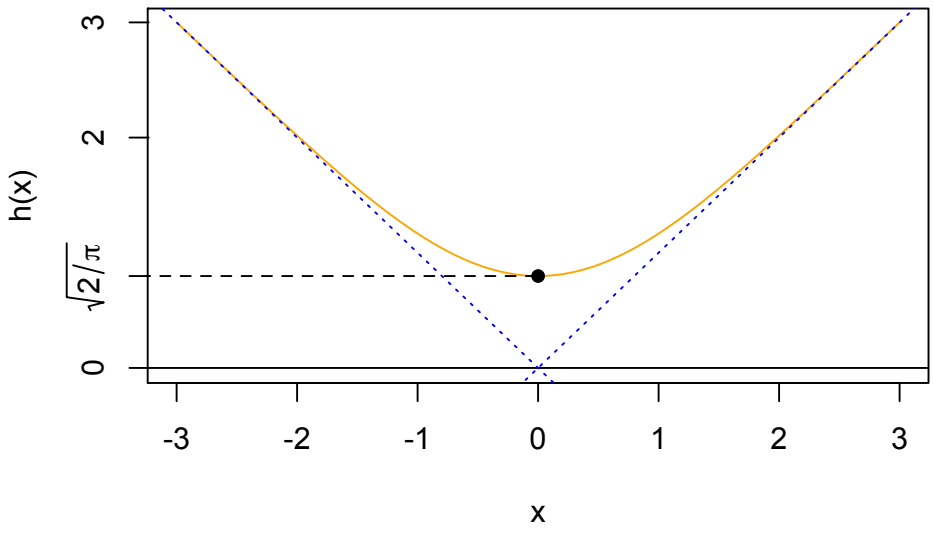

Figure 2.2: The functions $h(x)$ (orange) and $|x|$ (blue).

By the fact that $\widehat{Q}_{\tau^2} = N(\widehat{\beta}, \tau^2 I_p)$, we have

$$\widehat{Q}_{\tau^2}\|\beta - \beta^*\|^2 = \|\widehat{\beta} - \beta^*\|^2 + p\tau^2. \tag{2.52}$$

Hence, a risk bound for the penalized least-squares estimator (2.51) directly leads to the convergence of the variational posterior. To present a bound for $\|\widehat{\beta} - \beta^*\|^2$, we need to introduce some new notation. Let $S = \{j \in [p] : \beta_j^* \neq 0\}$ be the support of β^*. Define the restricted eigenvalue by

$$\kappa = \inf_{\{\Delta \neq 0 : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}} \frac{\frac{1}{\sqrt{n}}\|X\Delta\|}{\|\Delta\|}, \tag{2.53}$$

where $\|\Delta_S\|_1 = \sum_{j \in S} |\Delta_j|$ and $\|\Delta_{S^c}\|_1$ is defined similarly. The same quantity (2.53) also appears in the risk bound of LASSO [9].

Theorem 2.5.3. *Assume $\|X_{*j}\|/\sqrt{n} \leq L$ for all $j \in [p]$ and $\kappa \leq L$ with some constant $L > 0$. Choose $\lambda = C\sqrt{n \log p}$ and $\tau = O\left(\frac{1}{np}\right)$ for some sufficiently large constant $C > 0$.*

The solution to (2.51) satisfies

$$\|\widehat{\beta} - \beta^*\|^2 \lesssim \frac{s \log p}{n\kappa^4},$$

with probability at least $1 - p^{-C'}$ *uniformly over* $\|\beta^*\|_0 \leq s$ *for some constant* $C' > 0$. *As a consequence of (2.52), we also have*

$$\widehat{Q}_{\tau^2}\|\beta - \beta^*\|^2 \lesssim \frac{s \log p}{n\kappa^4},$$

with probability at least $1 - p^{-C'}$.

We note that $\frac{s \log p}{n\kappa^4}$ is the same rate of convergence of LASSO [9]. With τ chosen as small as $O\left(\frac{1}{np}\right)$, the statistical property of the variational posterior is very similar to that of the LASSO, and thus improves the original dense posterior distribution that is not suitable for sparse recovery.

2.5.2 Model Misspecification

In this section, we present an extension of Theorem 2.2.1 in the context of model misspecification. We consider a data generating process $X^{(n)} \sim P_*^{(n)}$ that may not satisfies the conditions (C1)-(C3). The following theorem shows that the convergence rate of the variational posterior will then have an extra term that characterizes the deviation of $P_*^{(n)}$ to the model specified by the likelihood.

Theorem 2.5.4. *Suppose* ϵ_n *is a sequence that satisfies* $n\epsilon_n^2 \geq 1$. *Assume that the conditions (C1)-(C3) hold with* $P_0^{(n)}$ *replaced by* $P_{\theta_0}^{(n)}$. *Then for the variational posterior* $\widehat{Q}$ *defined in (2.3), we have*

$$P_*^{(n)}\widehat{Q}L(P_\theta^{(n)}, P_{\theta_0}^{(n)}) \leq M\left(n\left(\epsilon_n^2 + \gamma_n^2\right) + D_2\big(P_*^{(n)} \| P_{\theta_0}^{(n)}\big)\right), \tag{2.54}$$

for some constant M *only depending on* C_1, C *and* ρ *in (C1)-(C3), where the quantity* γ_n^2 *is*

defined as

$$\gamma_n^2 = \frac{1}{n} \inf_{Q \in \mathcal{S}} P_*^{(n)} D(Q || \Pi(\cdot | X^{(n)})).$$

We note that here γ_n^2 is defined with respect to $P_*^{(n)}$ instead of $P_0^{(n)}$ in Theorem 2.2.1. Theorem 2.2.1 can be viewed as a special case of Theorem 2.5.4 with $P_0^{(n)} = P_*^{(n)} = P_{\theta_0}^{(n)}$. The extra term in the convergence rate that characterizes model misspecification is given by $D_2\big(P_*^{(n)} \| P_{\theta_0}^{(n)}\big)$. In fact, it can be replaced by any ρ-Rényi divergence with $\rho > 1$.

Convergence rates of variational approximation to tempered posterior distributions under model misspecification have been studied by [1] (See their Theorem 2.7). Our results complement theirs by considering variational approximation to the ordinary posterior.

The next theorem gives sufficient conditions so that the variational approximation error γ_n^2 is dominated by the sum of the other two terms in (2.54). It can be viewed as an extension of Theorem 2.2.3.

Theorem 2.5.5. *Suppose there are constants $C_1, C_2 > 0$, such that*

$$\inf_{Q \in \mathcal{S} \cap \mathcal{E}} D(Q \| \Pi) \leq C_1 \left(n\epsilon_n^2 + D_2\big(P_*^{(n)} \| P_{\theta_0}^{(n)}\big) \right), \tag{C4**}$$

where $\mathcal{E} = \{Q : \mathrm{supp}(Q) \subset \mathcal{C}\}$ with

$$\mathcal{C} = \left\{ \theta : D(P_*^{(n)} \| P_\theta^{(n)}) \leq C_2 \left(n\epsilon_n^2 + D_2\big(P_*^{(n)} \| P_{\theta_0}^{(n)}\big) \right) \right\}.$$

Then, we have

$$n\gamma_n^2 \leq (C_1 + C_2) \left(n\epsilon_n^2 + D_2\big(P_*^{(n)} \| P_{\theta_0}^{(n)}\big) \right).$$

To end this section, we apply Theorem 2.5.4 and Theorem 2.5.5 to the piecewise constant model discussed in Section 2.3.3 and derive oracle inequalities for the variational posterior distributions.

Theorem 2.5.6. *Consider a prior distribution Π that satisfies (2.26) and (2.27). Then, for*

any $\theta^* \in \mathbb{R}^n$*, we have*

$$P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MC}} \|\theta - \theta^*\|^2 \lesssim \min_{1 \le k \le n} \left\{ \inf_{\theta_0 \in \Theta_k(B)} \|\theta^* - \theta_0\|^2 + k \log n \right\},$$

$$P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}} \|\theta - \theta^*\|^2 \lesssim \min_{1 \le k \le n} \left\{ \inf_{\theta_0 \in \Theta_k(B)} \|\theta^* - \theta_0\|^2 + k \log n \right\},$$

where the definitions of $\widehat{Q}_{\mathrm{MC}}$ *and* $\widehat{Q}_{\mathrm{MC}}^{\mathrm{joint}}$ *are given in Theorem 2.3.5.*

CHAPTER 3

A GENERAL VARIATIONAL BAYES ALGORITHM

3.1 Introduction

Though the theoretical convergence results for the general variational posterior distribution have been established in Chapter 2, it is not easy to apply this theory in practice. For a statistical model, in order to to apply the variational algorithm with theoretical guarantee, the following two conditions must be satisfied: a) A prior must be well designed so that the prior mass condition is satisfied. b) The variational set should be computationally feasible. However, the first condition does not hold in the majority of scenarios if we do not assume the true parameter is bounded. Recently, many literatures are released to deal with this problem. In the sparse sequence model, [18] proves that with a spike and heavy tail slab prior, the posterior distribution will concentrate around the true parameter even though the true parameter is not bounded. This theorem is generalized in sparse linear regression by [16]. Recently, [28] provides a unified methodology and theory for a group of high-dimensional linear structured models and shows an oracle type of convergence result without assuming the true parameter is bounded. However, all above results are limited to the concentration of the exact posterior distribution, which is usually not computable. For the variational Bayes method, [71] provides a theoretical and computational convergence result of the variational algorithm for the stochastic block model. [51] gives the concentration results for variational posterior distribution in the sparse linear regression model without the boundedness condition of the true coefficients. Nevertheless, it is not sure whether these results can be generalized to more complicated cases.

In this chapter, we want to solve the problems (a) and (b) simultaneously by providing a general variational algorithm to obtain an approximation to the posterior distribution in the unified model proposed in [28] as well as develop the convergence result for the variational posterior distribution. In the variational algorithm, the prior we apply is slightly

modified from the prior in [28] to enable the computational conjugacy. To deal with the dimension heterogeneity given by the hyper index, we apply the variational Bayes with model selection procedure discussed in Section 2.4. As the traditional mean field variational class is computationally infeasible due to the exponentially large number of structures, we apply a modified mean field variational class under each hyper index. Finally, we can prove the corresponding variational posterior distribution has the same oracle type of convergence result as in [28], which provides the theoretical guarantee of the variational Bayes method.

Then we propose a general variational Bayes (VB) algorithm to solve for the variational posterior distribution. This general algorithm can be specialized to various models such as the stochastic block model, biclustering model, sparse linear regression, multiple regression with group sparsity, multi-task learning, and dictionary learning. At the end, we conduct some simulations in the stochastic block model and sparse linear regression. In the stochastic block model, the VB algorithm can outperform the traditional spectral method when the signal-to-noise ratio is large, and the network is unbalanced. In the sparse linear regression, when the signal-to-noise ratio is large, our VB algorithm can outperform LASSO under both FDR and ℓ_2 distance criteria, regardless of the collinearity of design matrix and the unboundedness of true coefficients.

The rest of this chapter is organized as follows. In Section 3.2, we reintroduce the model and the prior in [28], and propose our variational inference procedure with model selection for this general linear structured model. We also derive the convergence rate for the variational posterior distribution in this section. In Section 3.3, we develop a general VB algorithm. In Section 3.4, we specialize this general algorithm to six specific high dimensional linear structured models. In Section 3.5, we provide simulation results for the stochastic block model and sparse linear regression model to compare the performance of the novel VB algorithm and classical methods.

3.2 Main Results

3.2.1 Structured Linear Models

In the beginning, we first introduce a group of linear structured models proposed in [28]:

$$Y = \mathcal{X}_Z(B) + W, \tag{3.1}$$

where $W \in \mathbb{R}^N$ is a noise vector and $\mathcal{X}_Z$ is a linear operator depending on the structure label Z. For example, in the sparse linear regression $Y = X\gamma + W$ with a sparse coefficient $\gamma = (\gamma_S^T, 0_{S^c}^T)^T$ for some subset S, we have $B = \gamma_S$ and $Z = (z_1, \cdots, z_s)$ corresponding to indices in S. In this case $\mathcal{X}_Z = X_Z = (\vec{X}_{z_1}, \cdots, \vec{X}_{z_s})$ is the design matrix with selected columns by Z and $B = \gamma_S$. Then we have the representation $X\gamma = X_{\cdot S}\gamma_S = \mathcal{X}_Z(B)$. In general, we can assume the structure Z is in some discrete space $\mathcal{Z}_\tau$ determined by a index $\tau \in \mathcal{T}$ and the corresponding parameter space for B is $\mathbb{R}^{\ell_\tau}$.

For computational feasibility, we put one more assumption on the structure label Z compared to [28]. We assume Z is fully determined by several label vectors, i.e. $Z = (z_1, z_2, \cdots, z_m)$, $z_i = (z_{i1}, \cdots, z_{in_i})$ and $z_{ij} \in [k_i] = \{1, 2, \cdots, k_i\}$ for $1 \le j \le n_i$. In a sparse linear regression model, the structure Z is only determined by a label vector $z = (z_1, \cdots, z_s) \in [p]^s$, where p is the number of columns in the design matrix. In other words, the structure Z can be viewed as a component of s labels, where each label is selected from p columns. However, not all label vectors in $[p]^s$ is valid to construct a non-degenerating linear transformation $\mathcal{X}_Z$. For example, when $z_i = z_j$ for $i \neq j$, the column $\vec{X}_{z_i} = \vec{X}_{z_j}$ is selected twice and the effective sparsity of this model is smaller than s. Thus, the parameter space $\mathcal{Z}_s$ is defined as $\left\{z \in [p]^s : z_i \neq z_j \text{ for } i \neq j\right\}$, which is only a subset of $[p]^s$.

The framework (3.1) includes many models. We consider the following representative instances listed as below, where all specific models are reformulated in a similar way as discussed in [28] except that we require the structure Z to be a union of multiple label

vectors.

1. (Stochastic block model). Consider $\mathcal{X}_Z(B) \in [0,1]^{n\times n}$ to be the probability matrix to generate the adjacency matrix for a random graph with $[\mathcal{X}_Z(B)]_{ij} = B_{z_i z_j}$. $z = (z_1, \cdots, z_n) \in [k]^n$ is the label vector of the nodes. Moreover, it's easy to see the dimension of the parameter B is k^2 when we do not impose symmetry for B. Then the stochastic block model can be regarded as a special case of (3.1) with $Z = z$, $\tau = k$, $\mathcal{T} = [n]$, $\mathcal{Z}_k = [k]^n$, and $\ell_k = k^2$.

2. (Biclustering model). In a biclustering model, $\mathcal{X}_Z(B) \in \mathbb{R}^{n\times m}$ represents the model means with both row and column clustering structures. In other words, we can write $[\mathcal{X}_Z(B)]_{ij} = B_{z_i z_j}$ for some $z_1 \in [k]^n$ and $z_2 \in [l]^m$. The dimension of the parameter B is kl. Thus, the biclustering model can also be viewed as a special case of our general linear structured model (3.1) by setting $Z = (z_1, z_2)$, $\tau = (k, l)$, $\mathcal{T} = [n] \times [m]$, $\mathcal{Z}_{k,l} = [k]^n \times [l]^m$ and $\ell_{k,l} = kl$.

3. (Sparse Linear Regression). A sparse linear regression model is given by $Y = X\gamma + W$, where the design matrix is given by $X \in \mathbb{R}^{n\times p}$ and the coefficient $\gamma = (\gamma_S^T, 0_{S^c}^T)^T$ is a sparse vector with the support S as a subset of $[p]$. Then, we can formulate it as $X\gamma = \mathcal{X}_Z(B)$ with $\mathcal{X}_Z = X_Z = (\vec{X}_{z_1}, \cdots, \vec{X}_{z_s})$ and $B = \gamma_S$. In this case, it can be represented as (3.1) by letting $Z = z = (z_1, \cdots, z_s)$, $\tau = s$, $\mathcal{T} = [p]$, $\mathcal{Z}_s = \big\{z \in [p]^s : z_i \neq z_j \text{ for } i \neq j\big\}$ and $\ell_s = s$.

4. (Multiple linear regression with group sparsity). This model is similar to sparse linear regression, except that the coefficient vector γ is replaced by a matrix Γ. We also assume that indices of all nonzero rows of the coefficient matrix $\Gamma \in \mathbb{R}^{p\times m}$ are in a subset $S \subset [p]$. Then this model can be rewritten in the form of (3.1) in a similar way to the sparse linear regression model with $\ell_s = ms$.

5. (Multi-task learning). A multi-task learning model can be regarded as a collection of m linear regression problems with the coefficient vectors sharing a clustering structure.

We consider $X\Gamma$ as the linear structured signal for some $\Gamma \in \mathbb{R}^{p\times m}$. The j-th column of Γ can be expressed as $\Gamma_{*j} = B_{*z_j}$ for some $z \in [k]^m$ and $B \in \mathbb{R}^{p\times k}$. Therefore, this model can be reformulated as a special case of (3.1) with $Z = z$, $\tau = k$, $\mathcal{T} = [m]$, $\mathcal{Z}_k = [k]^m$ and $\ell_k = pk$.

6. (Dictionary learning). In this model, we consider a discrete sparse coding mode $\mathcal{X}_Z(B) = BZ \in \mathbb{R}^{n\times d}$ for some $Z = \{-1, 0, 1\}^{p\times d}$ and $B \in \mathbb{R}^{n\times p}$. Besides, we also assume that each column of Z is sparse. Then, dictionary learning can be viewed as multiple sparse linear regression models without knowing the design matrix. It can be written as form (3.1) by letting $\tau = (p, s)$, $\mathcal{T} = \{(p, s) \in [n \wedge d] \times [n] : s \leq p\}$, $\mathcal{Z}_{p,s} = \left\{Z \in \{-1, 0, 1\}^{p\times d} : \max_{j\in[d]} |\text{supp}(Z_{*j})| \leq s\right\}$ and $\ell_{p,s} = np$.

In some models above, the parameters and response have the matrix form. W e can define these models by vectorizing the matrix response and parameters. For example, in the stochastic block model, if we assume $Z = (\mathbb{I}_{z_i=j})_{1\leq i\leq n, 1\leq j\leq k} \in \{0, 1\}^{n\times k}$, then we have $\mathbb{E}(A) = \Theta = ZBZ^T$. However, this linear structured model can also be expressed as $\text{Vec}(\Theta) = (Z\otimes Z)\text{Vec}(B)$. Similar techniques can also be applied to other models. Therefore, without loss of generality, we can assume Y and B to be vectors.

3.2.2 The Prior Distribution

In this section, we want to build a prior that we can use in the variational algorithm. First of all, I want to introduce the prior proposed in [28]. The prior procedure has three steps. First of all, we sample the hyper index τ from a discrete measure. Then conditioning on τ, we sample Z from the non-degenerated structure set $\bar{\mathcal{Z}}_\tau \subseteq \mathcal{Z}_\tau$. As the linear transformation $\mathcal{X}_Z$ can also be understood as a matrix in $\mathbb{R}^{N\times \ell_\tau}$, the non-degenerated structure set $\bar{\mathcal{Z}}_\tau$ can be defined as

$$\bar{\mathcal{Z}}_\tau = \left\{Z \in \mathcal{Z}_\tau : \det(\mathcal{X}_Z^T \mathcal{X}_Z) > 0\right\}. \tag{3.2}$$

Lastly, given (τ, Z), we can sample the parameter B from an elliptical Laplace distribution with density function as:

$$f_{\tau,Z}(B) = \frac{\sqrt{\det(\mathcal{X}_Z^T \mathcal{X}_Z)}}{2} \left(\frac{\nu}{\sqrt{\pi}}\right)^{\ell_\tau} \frac{\Gamma(\ell_\tau/2)}{\Gamma(\ell_\tau)} \exp\left(-\nu \|\mathcal{X}_Z(B)\|\right). \tag{3.3}$$

As Z is sampled from non-degenerated structure set $\bar{\mathcal{Z}}_\tau$, we have $\det(\mathcal{X}_Z^T \mathcal{X}_Z) > 0$, and (3.3) is well-defined. Assume

$$\epsilon_\tau \geq \ell_\tau + \log|\bar{\mathcal{Z}}_\tau|. \tag{3.4}$$

The prior sampling procedure in [28] can be explicitly expressed as:

- Sample $\tau \sim \pi$ from $\mathcal{T}$, where $\pi(\tau) \propto \frac{\Gamma(\ell_\tau)}{\Gamma(\ell_\tau/2)} \exp(-D\epsilon_\tau)$;
- Conditioning on τ, sample Z uniformly from the set $\bar{\mathcal{Z}}_\tau = \{Z \in \mathcal{Z}_\tau : \det(\mathcal{X}_Z^T \mathcal{X}_Z) > 0\}$.
- Conditioning on (τ, Z) sample $B \sim g_{\ell_\tau, \mathcal{X}_Z, \alpha}$, where

$$g(B; \ell_\tau, \mathcal{X}_Z, \nu) = \frac{\sqrt{\det(\mathcal{X}_Z^T \mathcal{X}_Z)}}{2} \left(\frac{\nu}{\sqrt{\pi}}\right)^{\ell_\tau} \frac{\Gamma(\ell_\tau/2)}{\Gamma(\ell_\tau)} \exp\left(-\nu \|\mathcal{X}_Z(B)\|\right). \tag{3.5}$$

In the section 4 of [28], the authors assume the observation is generated by $Y = \theta^* + W$ with a sub-Gaussian noise W and then show that under some mild conditions, the posterior distribution have the following oracle type of convergence result:

$$P_{\theta^*}\Pi\left(\|\mathcal{X}_Z(B) - \theta^*\|^2 | Y\right) \leq (1+\delta)\|\mathcal{X}_{Z^*}(B^*) - \theta^*\|^2 + C\epsilon_{\tau^*},$$

for any $\tau^* \in \mathcal{T}$, $Z^* \in \bar{\mathcal{Z}}_{\tau^*}$, $B^* \in \mathbb{R}^{\ell_{\tau^*}}$ and any sufficiently small constant $\delta \in (0,1)$. The authors in [28] also point out that the rate ϵ_τ for each model they consider is minimax optimal. Thus, to derive the variational algorithm, a natural idea is to use the same prior as in [28].

However, from the computation point of view, the elliptical Laplace prior on B cannot be

conjugately updated with the normal likelihood of the model (3.1). The numerical estimation on the variational conditional mean will also slow down the computations especially when the dimension ℓ_τ is large. Thus, we need to modify the prior. Instead of sampling B directly from an elliptical Laplace prior, we sample B in a hierarchical way:

- First, conditioning on τ, sample $\lambda \sim \text{IG}\left(\frac{\ell_\tau+1}{2}, \beta\right)$, where $\text{IG}(\alpha, \beta)$ is the inverse gamma distribution with density:

$$\pi(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{-\alpha-1} \exp\left(-\frac{\beta}{\lambda}\right). \tag{3.6}$$

- Then, conditioning on (τ, Z, λ), sample $B \sim N(0_{\ell_\tau}, \lambda^{-1}(\mathcal{X}_Z^T \mathcal{X}_Z)^{-1})$ with the density

$$f_{\ell_\tau, Z, \lambda}(B) = \sqrt{\frac{\lambda^{\ell_\tau} \det(\mathcal{X}_Z^T \mathcal{X}_Z)}{(2\pi)^{\ell_\tau}}} \exp\left(-\frac{\lambda}{2} \|\mathcal{X}_Z(B)\|^2\right). \tag{3.7}$$

Then, our prior sampling procedure can be summarized as following:

1. Sample $\tau \sim \pi$ from $\mathcal{T}$, where $\pi(\tau) \propto \frac{\Gamma(\ell_\tau)}{\Gamma(\ell_\tau/2)} \exp\left(-D\epsilon_\tau\right)$;
2. Conditioning on τ, sample Z uniformly from the set $\bar{\mathcal{Z}}_\tau = \{Z \in \mathcal{Z}_\tau : \det(\mathcal{X}_Z^T \mathcal{X}_Z) > 0\}$.
3. Conditioning on τ, sample $\lambda \sim \text{IG}\left(\frac{\ell_\tau+1}{2}, \beta\right)$, where $\text{IG}(\alpha, \beta)$ is the inverse gamma distribution with density:

$$\pi(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{-\alpha-1} \exp\left(-\frac{\beta}{\lambda}\right). \tag{3.8}$$

4. Conditioning on (τ, Z, λ), sample $B \sim f_{\ell_\tau, \mathcal{X}_Z, \lambda}$ with

$$f_{\ell_\tau, \mathcal{X}_Z, \lambda}(B) = \sqrt{\frac{\lambda^{\ell_\tau} \det(\mathcal{X}_Z^T \mathcal{X}_Z)}{(2\pi)^{\ell_\tau}}} \exp\left(-\frac{\lambda}{2} \|\mathcal{X}_Z(B)\|^2\right). \tag{3.9}$$

In our modified prior, we introduce one more variable $\lambda \in \mathbb{R}$, and the conditional sampling

procedure of B is a normal distribution, which can be conjugately updated with the normal likelihood. However, if we write down the entire joint distribution of the prior and integrate the variable λ out, then we will get exactly the prior in [28]. This equivalence is illustrated in the following theorem.

Proposition 3.2.1. *The marginal sampling procedure of (3.8) and (3.9) is (3.5) with* $\nu = \sqrt{2\beta}$.

3.2.3 Variational Inference with Model Selection

As the dimension of the parameter B, and the parameter set of the structure label Z depend on τ. We consider the variational Bayes procedure with model selection approach proposed in Section 2.4. For a given τ, a natural idea is to apply the mean-field class as the variational set. The mean field class is defined as below:

$$\widetilde{\mathcal{S}}_{\mathrm{MF}}^{(\tau)} = \left\{ Q^{(\tau)} : Q^{(\tau)}\left(Z, B, \lambda\right) = Q_Z^{(\tau)}(Z) Q_B^{(\tau)}(B) Q_\lambda^{(\tau)}(\lambda) \right\}.$$

In this mean field class, the distributions of Z, B and λ are independent. According to equation (2.34), the variational inference with model selection method is equivalently to solve the following optimization problem:

$$\max_{\tau \in \mathcal{T}, \mathcal{Q}^{(\tau)} \in \widetilde{\mathcal{S}}_{\mathrm{MF}}^{(\tau)}} F(Q^{(\tau)}, \tau),$$

where $F(Q^{(\tau)}, \tau)$ is the evidence lower bound defined by

$$F(Q^{(\tau)}, \tau) = \int \log \phi_{\mathcal{X}_Z(B)}(Y) dQ^{(\tau)}(Z, B) - D\left(Q_Z^{(\tau)} Q_B^{(\tau)} Q_\lambda^{(\tau)} \| \Pi^{(\tau)}(Z, B, \lambda)\right) + \log \pi(\tau),$$

with $\phi_\theta(Y) = \frac{1}{(\sqrt{2\pi})^N} \exp\left(-\frac{1}{2}\|Y - \theta\|^2\right)$.

However, in the KL-divergence term $D\left(Q_Z^{(\tau)} Q_B^{(\tau)} Q_\lambda^{(\tau)} \| \Pi^{(\tau)}(Z, B, \lambda)\right)$, there is a term $Q_Z \log\det(\mathcal{X}_Z^T \mathcal{X}_Z)$. As the number of structures is exponentially large, we cannot estimate

the conditional expectation of $\log\det(\mathcal{X}_Z^T\mathcal{X}_Z)$ while updating Z coordinate-wise. Thus, we have to modify the mean field class to put more constraint on the distributions of Z. We also put some constraints on the distributions of B and λ to enable the conjugate computation procedure. The modified mean field variational class is defined below:

$$\begin{aligned}\mathcal{S}_{\mathrm{MF}}^{(\tau)} &= \Big\{Q^{(\tau)} : Q^{(\tau)}(Z,B,\lambda) = Q_Z^{(\tau)}(Z)Q_B^{(\tau)}(B)Q_\lambda^{(\tau)}(\lambda), \\ &\quad Z\in\bar{\mathcal{Z}}_\tau, B\in\mathbb{R}^{\ell_\tau}, \lambda\in\mathbb{R}, Q_Z^{(\tau)}\in\mathcal{S}_Z^{(\tau)}, Q_B^{(\tau)}\in\mathcal{S}_B^{(\tau)}, Q_\lambda^{(\tau)}\in\mathcal{S}_\lambda^{(\tau)}\Big\},\end{aligned} \tag{3.10}$$

where $\mathcal{S}_Z^{(\tau)}$, $\mathcal{S}_B^{(\tau)}$ and $\mathcal{S}_\lambda^{(\tau)}$ denote some distribution families for $Z\in\bar{\mathcal{Z}}_\tau$, $B\in\mathbb{R}^{\ell_\tau}$ and $\lambda\in\mathbb{R}$. To solve the computational issue when we update Z, we consider the following mass point distribution family for $\mathcal{S}_Z^{(\tau)}$ for each $\tau\in\mathcal{T}$.

$$\mathcal{S}_Z^{(\tau)} = \left\{Q_Z^{(\tau)} : Q_Z^{(\tau)}\left(Z=\widetilde{Z}\right)=1, \text{ for some } \widetilde{Z}\in\bar{\mathcal{Z}}_\tau\right\}.$$

Besides, $\mathcal{S}_B^{(\tau)}$ is selected as the normal distribution family and $\mathcal{S}^{(\tau)}$ is selected as a parametric distribution family with parameter $a>0$. They are defined as follows:

$$\mathcal{S}_B^{(\tau)} = \left\{Q_B^{(\tau)} : Q_B^{(\tau)} = N(\mu,\Sigma), \text{ for } \mu\in\mathbb{R}^{\ell_\tau}, \Sigma\in\mathbb{R}^{\ell_\tau\times\ell_\tau}, \Sigma\succ 0\right\}.$$

$$\mathcal{S}_\lambda^{(\tau)} = \left\{Q_\lambda^{(\tau)} : \frac{dQ_\lambda^{(\tau)}}{d\lambda} = \sqrt{\frac{\beta}{\pi}}\lambda^{-3/2}\exp\left(\sqrt{2a\beta}-\frac{\beta}{\lambda}-\frac{a\lambda}{2}\right), \text{ for some } a>0\right\}.$$

Then, the variational inference with model selection is actually to solve the following optimization problem

$$\max_{\tau\in\mathcal{T}, \mathcal{Q}^{(\tau)}\in\mathcal{S}_{\mathrm{MF}}^{(\tau)}} F(Q^{(\tau)},\tau), \tag{3.11}$$

where

$$F(Q^{(\tau)},\tau) = \int \log\phi_{\mathcal{X}_Z(B)}(Y)dQ^{(\tau)}(Z,B) - D\left(Q_Z^{(\tau)}Q_B^{(\tau)}Q_\lambda^{(\tau)}\|\Pi^{(\tau)}(Z,B,\lambda)\right) + \log\pi(\tau). \tag{3.12}$$

Suppose the solution of (3.11) is $(\widehat{\tau}, \widehat{Q}^{(\widehat{\tau})})$, then $\widehat{Q}^{(\widehat{\tau})}$ is the variational posterior distribution after model selection.

3.2.4 Data Generating Processing and Concentration Results

In this section, we want to show the concentration result for the variational posterior distribution $\widehat{Q}^{(\widehat{\tau})}$. This is an essential theoretical guarantee when we want to use the variational algorithm in practice. Just as in [28], we assume that data are generated from an arbitrary signal with sub-Gaussian noise:

$$Y = \theta^* + W,$$

where $W = Y - \theta^*$ is the noise vector with a sub-Gaussian tail satisfying

$$\mathbb{P}(|\langle W, K\rangle| > t) \leq e^{-\rho t^2/2} \text{ for all } \|K\| = 1. \tag{3.13}$$

For the hyper index τ, we put the same mild conditions on it as in [28]:

$$|\{\tau \in \mathcal{T} : t - 1 < \epsilon_\tau \leq t\}| \leq t. \tag{3.14}$$

Then for the variational posterior distribution $\widehat{Q}^{(\widehat{\tau})}$, we have the following theorem.

Theorem 3.2.1. *For the model defined in (3.1) and the modified hierarchical prior on (τ, Z, λ, B), we assume $\widehat{Q}^{(\widehat{\tau})}$ is the variational posterior distribution obtained from (3.11). If conditions (3.4), (3.13) and (3.14) are satisfied, for any $\tau^* \in \mathcal{T}$, $Z^* \in \mathcal{Z}_{\tau^*}$, $B^* \in \mathbb{R}^{\ell_{\tau^*}}$ and $\delta > 0$, there exists a constant $C > 0$, such that*

$$P_{\theta^*}\widehat{Q}^{(\widehat{\tau})}\|\mathcal{X}_Z(B) - \theta^*\|^2 \leq (1+\delta)\|\mathcal{X}_{Z^*}(B^*) - \theta^*\|^2 + C\epsilon_{\tau^*}, \tag{3.15}$$

when $D > D_{\beta,\rho,\delta}$ for a constant $D_{\beta,\rho,\delta}$ only depending on β, ρ, δ.

This theorem gives an oracle type of convergence rate for the variational posterior dis-

tribution $\widehat{Q}^{(\widehat{\lambda})}$. The upper bound is the same as in [28] for the true posterior distribution. Thus, there is no loss during the variational approximation procedure in the sense of convergence rate, which provides a strong theoretical guarantee when we apply this variational algorithm with model selection procedure in practice.

3.3 A General Variational Bayes Algorithm with Model Selection

In this section, we will derive a general variational Bayes (VB) algorithm for the linear structure model (3.1). The fundamental idea of the algorithm comes from coordinate ascend variational inference algorithm (CAVI) in [10]. For any fixed $\tau \in \mathcal{T}$, suppose $Q^{(\tau)} = Q_Z^{(\tau)} Q_B^{(\tau)} Q_\lambda^{(\tau)}$ is in the modified mean field class (3.10), where $Q_Z^{(\tau)} \in \mathcal{S}_Z^{(\tau)}$, $Q_B^{(\tau)} \in \mathcal{S}_B^{(\tau)}$, $Q_\lambda^{(\tau)} \in \mathcal{S}_\lambda^{(\tau)}$. Furthermore, we can assume that

$$Q_Z^{(\tau)}\left(Z = Z^{(\tau)}\right) = 1, \qquad Z^{(\tau)} \in \bar{\mathcal{Z}}_\tau,$$

$$Q_B^{(\tau)} = N(\mu^{(\tau)}, \Sigma^{(\tau)}), \qquad \mu^{(\tau)} \in \mathbb{R}^{\ell_\tau},\ \Sigma^{(\tau)} \in \mathbb{R}^{\ell_\tau \times \ell_\tau},$$

and $Q_\lambda^{(\tau)}$ has the density

$$\frac{dQ_\lambda^{(\tau)}}{d\lambda} = \sqrt{\frac{\beta}{\pi}} \lambda^{-3/2} \exp\left(\sqrt{2a^{(\tau)}\beta} - \frac{\beta}{\lambda} - \frac{a^{(\tau)}\lambda}{2}\right).$$

Then $Q^{(\tau)}$ is fully determined by $(Z^{(\tau)}, a^{(\tau)}, \mu^{(\tau)}, \Sigma^{(\tau)})$. And we can define the objective function $L(\tau, Z^{(\tau)}, a^{(\tau)}, \mu^{(\tau)}, \Sigma^{(\tau)}) = -F(Q^{(\tau)}, \tau)$. In this way, the optimization problem (3.11) is simplified as

$$\min_{\tau \in \mathcal{T}} \min_{\substack{Z^{(\tau)} \in \bar{\mathcal{Z}}_\tau, a^{(\tau)} \in \mathbb{R}_+, \mu^{(\tau)} \in \mathbb{R}^{\ell_\tau} \\ \Sigma^{(\tau)} \in \mathbb{R}^{\ell_\tau \times \ell_\tau}, \Sigma^{(\tau)} \succ 0}} L(\tau, Z^{(\tau)}, a^{(\tau)}, \mu^{(\tau)}, \Sigma^{(\tau)}). \tag{3.16}$$

To solve the optimization problem (3.16), we can follow the idea of CAVI to update each

parameter iteratively:

- For each τ, initialize $Z^{(\tau)} \in \bar{\mathcal{Z}}_\tau$, $a^{(\tau)} > 0$, $\mu^{(\tau)} \in \mathbb{R}^{\ell_\tau}$ and $\Sigma^{(\tau)} \in \mathbb{R}^{\ell_\tau \times \ell_\tau}$ such that $\Sigma^{(\tau)} \succ 0$.

- When the algorithm does not converge,

  - Update $Z^{(\tau)}$ by

$$Z^{(\tau)} \leftarrow \underset{Z \in \bar{\mathcal{Z}}_\tau}{\operatorname{argmin}} L(\tau, Z, a^{(\tau)}, \mu^{(\tau)}, \Sigma^{(\tau)}); \tag{3.17}$$

  - Update $a^{(\tau)}$ by

$$a^{(\tau)} \leftarrow \underset{a>0}{\operatorname{argmin}} L(\tau, Z^{(\tau)}, a, \mu^{(\tau)}, \Sigma^{(\tau)}); \tag{3.18}$$

  - Update $(\mu^{(\tau)}, \Sigma^{(\tau)})$ by

$$(\mu^{(\tau)}, \Sigma^{(\tau)}) \leftarrow \underset{\mu \in \mathbb{R}^{\ell_\tau}, \Sigma \in \mathbb{R}^{\ell_\tau \times \ell_\tau}, \Sigma \succ 0}{\operatorname{argmin}} L(\tau, Z^{(\tau)}, a^{(\tau)}, \mu, \Sigma). \tag{3.19}$$

- Assume the final result is $(\widehat{Z}^{(\tau)}, \widehat{a}^{(\tau)}, \widehat{\mu}^{(\tau)}, \widehat{\Sigma}^{(\tau)})$, then we select $\widehat{\tau}$ by

$$\widehat{\tau} = \underset{\tau \in \mathcal{T}}{\operatorname{argmin}} L\left(\tau, \widehat{Z}^{(\tau)}, \widehat{a}^{(\tau)}, \widehat{\mu}^{(\tau)}, \widehat{\Sigma}^{(\tau)}\right). \tag{3.20}$$

However, the biggest challenge for the above procedure is that the update of $Z^{(\tau)}$ cannot be computed explicitly because the number of structures is exponential large, making the computation NP-hard. Thus, we apply a coordinate-wise update method to update $Z^{(\tau)}$ in each iteration:

- Assume $Z^{(\tau)} = (z_1, \cdots, z_m)$, where $z_i = (z_{i1}, \cdots z_{in_i})$ and $z_{ij} \in [k_i]$ for $1 \le i \le m$ and $1 \le j \le n_i$.

- For each $1 \leq i \leq m$ and $1 \leq j \leq n_i$, we update z_{ij} by

$$z_{ij} \leftarrow \underset{c \in [k_i], Z_{z_{ij}}(c) \in \bar{\mathcal{Z}}_\tau}{\operatorname{argmin}} L\left(\tau, Z_{z_{ij}}(c), a^{(\tau)}, \mu^{(\tau)}, \Sigma^{(\tau)}\right), \tag{3.21}$$

  where $Z_{z_{ij}}(c)$ denotes the current structure with z_{ij} replaced by c.

With the procedure above, each time we only need to update one label z_{ij} and hence only need to choose the optimal structure from at most k_i choices.

As we apply the coordinate-wise update, the current structure Z changes after each coordinate gets updated. There are no general rules for the order of coordinates to update. In our algorithm, we update each coordinate of Z in a randomized order.

Now we present the variational inference algorithm for the general linear structure model (3.1) in an explicit form. First of all, we release the algorithm when τ is given. For a fixed τ, assume the parameters in the t-th iteration are $(Z^{[t]}, a^{[t]}, \mu^{[t]}, \Sigma^{[t]})$. We also denote $\delta^{[t]} = 1 + \sqrt{\frac{2\beta}{a^{[t]}}}$ and $L_t = L(\tau, Z^{[t]}, a^{[t]}, \mu^{[t]}, \Sigma^{[t]})$ for short. The algorithm stops as long as the iteration number t reach the predetermined maximum iteration number M, or the change L_t is smaller than the predetermined tolerance level ϵ. The general VB algorithm for the fixed τ is given in Algorithm 2.

Algorithm 2 General VB Algorithm for fixed τ

1: **Input:** Index $\tau \in \mathcal{T}$, initial structure $Z^{[0]} \in \bar{\mathcal{Z}}_\tau$, maximum iteration number M, tolerance level ϵ.

2: **Initialize:** Iteration number $t = 0$, objective function $L_0 = \infty$, change of objective function $\nabla L_0 = \infty$. Other parameters are initialized according to $Z^{[0]}$ as below:

$$a^{[0]} = \mathsf{Tr}\left(\mathcal{X}_{Z^{[0]}}^T \mathcal{X}_{Z^{[0]}}\right), \qquad \delta^{[0]} = 1 + \sqrt{\frac{2\beta}{a^{[0]}}},$$

$$\mu^{[0]} = \left(\delta^{[0]}\right)^{-1} \left(\mathcal{X}_{Z^{[0]}}^T \mathcal{X}_{Z^{[0]}}\right)^{-1} \mathcal{X}_{Z^{[0]}}^T Y, \qquad \Sigma^{[0]} = \left(\delta^{[0]}\right)^{-1} \left(\mathcal{X}_{Z^{[0]}}^T \mathcal{X}_{Z^{[0]}}\right)^{-1}.$$

3: **while** $t < M$ and $\nabla L_t > \epsilon$ **do** $t \leftarrow t + 1$.

Update Z:

for $i = 1, \cdots m$ **do**

for j from 1 to n_i in a randomized order **do**

Assume $Z_{z_{ij}}(c)$ is the current label with $z_{ij}^{[t-1]}$ replaced by c, then

$$z_{ij}^{[t]} = \underset{c\in[k_i], Z_{z_{ij}}(c)\in\bar{\mathcal{Z}}_\tau}{\operatorname{argmin}} \left\{ -\frac{1}{2}\log\det\left(\mathcal{X}_{Z_{z_{ij}}(c)}^T \mathcal{X}_{Z_{z_{ij}}(c)}\right) - Y^T \mathcal{X}_{Z_{z_{ij}}(c)} \mu^{[t-1]} + \frac{1}{2}\delta^{[t-1]}\left[\|\mathcal{X}_{Z_{z_{ij}}(c)}\mu^{[t-1]}\|^2 + \mathsf{Tr}\left(\mathcal{X}_{Z_{z_{ij}}(c)} \Sigma^{[t-1]} \mathcal{X}_{Z_{z_{ij}}(c)}^T\right)\right]\right\}, \tag{3.22}$$

end for

end for

Update a and δ

$$a^{[t]} = \left\|\mathcal{X}_{Z^{[t]}}\mu^{[t-1]}\right\|^2 + \mathsf{Tr}\left(\mathcal{X}_{Z^{[t]}}\Sigma^{[t-1]}\mathcal{X}_{Z^{[t]}}^T\right), \qquad \delta^{[t]} = 1 + \sqrt{\frac{2\beta}{a^{[t]}}} \tag{3.23}$$

Update μ and Σ:

$$\mu^{[t]} = \left(\delta^{[t]}\right)^{-1}\left(\mathcal{X}_{Z^{[t]}}^T\mathcal{X}_{Z^{[t]}}\right)^{-1}\mathcal{X}_{Z^{[t]}}^T Y, \qquad \Sigma^{[t]} = \left(\delta^{[t]}\right)^{-1}\left(\mathcal{X}_{Z^{[t]}}^T\mathcal{X}_{Z^{[t]}}\right)^{-1}. \tag{3.24}$$

Update L_t and ∇L_t

$$L_t = \frac{\ell_\tau}{2}\log\frac{\delta^{[t]}}{4\beta e^2} + \sqrt{\frac{a^{[t]}\beta}{2}} - \frac{1}{2}Y^T X_{Z^{[t]}}\mu^{[t]} + (D+1)\epsilon_\tau, \qquad \nabla L_t = |L_{t-1} - L_t|. \tag{3.25}$$

end while

Output: $\widehat{Z}^{(\tau)} = Z^{[t]}$, $\widehat{a}^{(\tau)} = a^{[t]}$, $\widehat{\mu}^{(\tau)} = \mu^{[t]}$, $\widehat{\Sigma}^{(\tau)} = \Sigma^{[t]}$, $\widehat{L}^{(\tau)} = L_t$.

The computational complexity for each iteration in this general algorithm is

$$O\left(\ell_\tau^2(\ell_\tau + N)\sum_{i=1}^{m} k_i n_i\right). \tag{3.26}$$

However, the actual complexity can be much smaller than (3.26) for a specific model, because the redundant computations can be removed.

Now, let me briefly explain how we get (3.25). To simply the formula, we use δ, Z, a, μ, Σ, $L(\tau)$ represent $\delta^{[t]}, a^{[t]}, Z^{[t]}, \mu^{[t]}, \Sigma^{[t]}, L_t$ in (3.25) for short. Then the original objective function derived from (3.12) is given by

$$\begin{aligned} L(\tau) &= \frac{1}{2}\left[\|\mathcal{X}_Z\mu - Y\|^2 + \mathsf{Tr}\left(\mathcal{X}_Z\Sigma\mathcal{X}_Z^T\right)\right] + \sqrt{\frac{a\beta}{2}} \\ &+\sqrt{\frac{a}{2\beta}}\left[\|\mathcal{X}_Z\mu\|^2 + \mathsf{Tr}\left(\mathcal{X}_Z\Sigma\mathcal{X}_Z^T\right)\right] + \frac{\ell_\tau}{2}\log\frac{\delta}{4\beta} + \log|\bar{\mathcal{Z}}_\tau| + D\epsilon_\tau. \end{aligned} \tag{3.27}$$

Based on the order of update in Algorithm 2, we have $\mu = \delta^{-1}\left(\mathcal{X}_Z^T\mathcal{X}_Z\right)^{-1}\mathcal{X}_Z^T Y$ and $\Sigma = \delta^{-1}\left(\mathcal{X}_Z^T\mathcal{X}_Z\right)^{-1}$ at the end of each iteration. Then the objective function $L(\tau)$ can be simplified as

$$L_t = \frac{\ell_\tau}{2}\log\frac{\delta}{4\beta} + \log|\bar{\mathcal{Z}}_\tau| + \sqrt{\frac{a\beta}{2}} - \frac{1}{2}Y^T X_Z\mu + D\epsilon_\tau. \tag{3.28}$$

However, as $\log|\bar{\mathcal{Z}}_\tau|$ is sometimes hard to compute, we further replace it by $\epsilon_\tau - \ell_\tau$, and then we can get the form (3.25).

The procedure in Algorithm 2 is random because we update each z_{ij} in a randomized order on j. As pointed out in [51], this order may have a significant effect, especially in variational inference. Therefore, it is highly recommended to run this algorithm multiple times and choose the best result among them.

We denote $\mathsf{VB}(\tau, Z, M, \epsilon)$ to Algorithm 2 with the input (τ, Z, M, ϵ) for simplification, then the model selection algorithm is given in Algorithm 3.

Algorithm 3 Variational Model Selection

1: **Input:** Index set $\mathcal{T}_0$, initial structure $Z^{[0](\tau)} \in \bar{\mathcal{Z}}_\tau$ for each $\tau \in \mathcal{T}_0$, number of trials R

2: **for** τ in $\mathcal{T}_0$ **do**

3: Run $\mathsf{VB}(\tau, Z^{[0](\tau)}, M, \epsilon)$, assign the result to $(\widehat{Z}^{(\tau)}, \widehat{a}^{(\tau)}, \widehat{\mu}^{(\tau)}, \widehat{\Sigma}^{(\tau)}, \widehat{L}^{(\tau)})$.

4: **end for**

5: Choose $\widehat{\tau}$ by $\widehat{\tau} = \operatorname{argmin}_{\tau \in \mathcal{T}_0} \widehat{L}^{(\tau)}$.

6: **Output:** Selected index $\widehat{\tau}$, corresponding parameters $\left(\widehat{Z}^{(\widehat{\tau})}, \widehat{a}^{(\widehat{\tau})}, \widehat{\mu}^{(\tau)}, \widehat{\Sigma}^{(\tau)}\right)$.

The derivations of Algorithm 2 and 3 are deferred in the proof of Theorem 5.2.1 in Section 5.2.

Now let's analyze the model selection criterion (3.27). When the iteration number M is large enough, the algorithm will converge to a stationary point at which $a = \|\mathcal{X}_Z\mu\|^2 + \mathsf{Tr}\left(\mathcal{X}_Z\Sigma\mathcal{X}_Z^T\right)$, then

$$\begin{aligned} L(\tau) &= \frac{1}{2}\left[\|\mathcal{X}_Z\mu - Y\|^2 + \mathsf{Tr}\left(\mathcal{X}_Z\Sigma\mathcal{X}_Z^T\right)\right] + \sqrt{\frac{\beta}{2}\left[\|\mathcal{X}_Z\mu\|^2 + \mathsf{Tr}\left(\mathcal{X}_Z\Sigma\mathcal{X}_Z^T\right)\right]} \\ &\quad + \frac{\ell_\tau}{2}\log\frac{\delta}{4\beta} + \log|\bar{\mathcal{Z}}_\tau| + D\epsilon_\tau. \end{aligned}$$

This objective function consists of three parts:

1.
$$L_1(\tau) = \frac{1}{2}\left[\|\mathcal{X}_Z\mu - Y\|^2 + \mathsf{Tr}\left(\mathcal{X}_Z\Sigma\mathcal{X}_Z^T\right)\right].$$

This is the average ℓ_2 loss between $\mathcal{X}_Z(B)$ and Y with $\mathcal{X}_Z(B)$ generated from variational posterior Q. It measures the accuracy of estimators generated from variational posterior distribution.

2.
$$L_2(\tau) = \sqrt{\frac{\beta}{2}\left[\|\mathcal{X}_Z\mu\|^2 + \mathsf{Tr}\left(\mathcal{X}_Z\Sigma\mathcal{X}_Z^T\right)\right]}.$$

This is the elliptical regularisation part on $\mathcal{X}_Z(B)$ that sampled from variational pos-

terior Q to ensure the final scale of estimator $\mathcal{X}_Z(B)$ cannot be too large.

3.
$$L_3(\tau) = \frac{\ell_\tau}{2} \log \frac{\delta}{4\beta} + \log |\bar{\mathcal{Z}}_\tau| + D\epsilon_\tau.$$

This part is the regularization on the complexity index τ so that the final model complexity will not be too large.

There is another point worth mentioning. When we update Z, the objective function has a term $-\log\det\left(\mathcal{X}_Z^T\mathcal{X}_Z\right)$. Thus, the VB algorithm can automatically adjust the penalty on the structure Z even when $\mathcal{X}_Z$ tends to be singular.

3.4 Applications

In this section, we specialize the general model (3.1) into the stochastic block model, biclustering model, sparse linear regression model, multiple linear regression with group sparsity, multi-task learning, and dictionary learning. Besides, we also present the VB algorithm corresponding to these models. To make the thesis concise, we only display the algorithm when the hyper index τ is fixed in each model. The derivations for these algorithms are given in Theorem 5.2.2 in Section 5.2.3.

3.4.1 Stochastic Block Model

Stochastic block model was first proposed in [35]. In this model, a symmetric adjacency matrix $A = A^T \in \{0,1\}^{n\times n}$ is given as the response. For each of node pair (i,j), we generate the edge by $A_{ij} \sim \text{Bern}(\theta_{ij})$ with $\theta_{ij} = B_{z_i z_j}$ for all $i > j$, where the label vector $z \in [k]^n$. We also assume that no self-circle in the graph, i.e. $A_{ii} = 0$ for $i \in [n]$. Our goal is to recover the true membership label z^* and the true signal θ^*.

Now we put this model into our general framework. In the stochastic block model, $Z = z$, $\tau = k$, $\mathcal{T} = [n]$ and $\mathcal{Z}_k = [k]^n$. Although B is assumed to be symmetric, we

do not need to put this assumption in our prior. Therefore, we have $\ell_k = k^2$, $|\mathcal{Z}_k| = k^n$ and we can choose $\epsilon_k = k^2 + n\log k$. In this case, it's not hard to see that $\bar{\mathcal{Z}}_k = \{z \in \mathcal{Z}_k : \sum_{i=1}^n \mathbb{I}\{z_i = t\} > 0, \text{ for all } 1 \le t \le k\}$. Then, the prior sampling procedure for the stochastic block model is given as follows:

1. Sample $k \sim \pi$ from $[n]$, where $\pi(k) \propto \frac{\Gamma(k^2)}{\Gamma(k^2/2)} \exp\big(-D(k^2 + n\log k)\big)$;
2. Conditioning on k, sample z uniformly from $\bar{\mathcal{Z}}_k$;
3. Conditioning on k, sample $\lambda \sim \mathrm{IG}((k^2+1)/2, \beta)$;
4. Conditioning on (k, z, λ), sample $B \sim f_{k,z,\lambda}$ with

$$f_{k,z,\lambda} \propto \exp\left(-\frac{\lambda}{2}\sum_{i=1}^n\sum_{j=1}^n B_{z_i z_j}^2\right);$$

5. Set $\theta_{ij} = B_{z_i z_j}$ for all $i \neq j$ and $\theta_{ii} = 0$ for all $i \in [n]$.

Assume $\left(\widehat{k}, \widehat{Q}^{(\widehat{k})}\right)$ is the solution of (3.11) for stochastic block model. If we assume there is no model misspecification, theoretically we will have the following concentration result.

Corollary 3.4.1. *Assume $\theta^*_{ij} = B^*_{z_i^* z_j^*}$ for $B^* \in [0,1]^{k^* \times k^*}$, $z^* \in [k^*]^n$ and $A_{ji} = A_{ij} \sim \mathrm{Bern}(\theta^*_{ij})$ for all $i < j$ and $A_{ii} = 0$ for all $i \in [n]$. Then*

$$P_{\theta^*}\widehat{Q}^{(\widehat{k})}\|\theta - \theta^*\|_F^2 \le M(k^{*2} + n\log k^*),$$

for any $D > D_{\beta,\rho}$ and some constant M only depending on β, ρ, D.

We can also derive the algorithm to solve for $(\widehat{k}, \widehat{Q}^{(\widehat{k})})$. First of all, we need to reformulate the model so that the signal $\mathcal{X}_Z(B)$, parameter B and the response Y are vectors. Specifically, the stochastic block model can be rewritten as $\mathrm{Vec}(A) = (Z \otimes Z)\mathrm{Vec}(B) + W$, where $Z = (\mathbb{I}\{z_i = j\})_{1 \le i \le n, 1 \le j \le k} \in [0,1]^{n \times k}$ to be the membership matrix corresponding to z and $\otimes$

represents Kronecker product. Then the design matrix is given by $\mathcal{X}_Z = Z \otimes Z$. Assume $n_c = \sum_{i=1}^{n} \mathbb{I}\{z_i = c\}$ for $1 \leq c \leq k$, then $\mathcal{X}_Z^T \mathcal{X}_Z = \text{diag}\left(n_1^2, n_1 n_2, \cdots, n_1 n_k, n_2 n_1, \cdots, \cdots, n_k^2\right)$.

Because $\mathcal{X}_Z^T \mathcal{X}_Z$ is a diagonal matrix, the updated Σ at each iteration is a diagonal matrix. We can simply assume $B_{cd} \sim N(\mu_{cd}, \Sigma_{cd})$ for $1 \leq c \leq k$ and $1 \leq d \leq k$ and update them pointwise. Algorithm 4 gives the variational Bayes algorithm for stochastic block model for a given k.

Algorithm 4 Variational Algorithm for Stochastic Block Model

1: **Input:** Number of clusters k, initial labels $z^{[0]} \in \bar{\mathcal{Z}}_k$, maximum iteration number M, tolerate level ϵ.

2: **Initialize:** Iteration number $t = 0$, objective function $L_0 = \infty$, change of objective function $\nabla L_0 = \infty$, $a^{[0]} = n^2$ and $\delta^{[0]} = 1 + \sqrt{\frac{2\beta}{n^2}}$. The initial μ and Σ are computed by followings:

$$\mu^{[0]} = \left(\delta^{[0]}\right)^{-1} \left(n_c^{[0]}\right)^{-1} \left(n_d^{[0]}\right)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \mathbb{I}\left\{z_i^{[0]} = c, z_j^{[0]} = d\right\}, \qquad 1 \leq c, d \leq k;$$

$$\Sigma_{cd}^{[0]} = \left(\delta^{[0]}\right)^{-1} \left(n_c^{[0]}\right)^{-1} \left(n_d^{[0]}\right)^{-1}, \qquad 1 \leq c, d \leq k;$$

where $n_c^{[0]} = \sum_{i=1}^{n} \mathbb{I}\left\{z_i^{[0]} = c\right\}$.

3: **while** $t < M$ and $\nabla L_t > \epsilon$ **do** $t \leftarrow t + 1$.

4: **Update** Z**:**

5: **for** $i = 1, \cdots, n$ in a randomized order **do**

6: **for** $c = 1, \cdots, k$ **do**

7: Set $z_j(i)^{[t]}$ to be the current label for z_j and $n_c(i) = \sum_{j \neq i} \mathbb{I}\{z_j(i)^{[t]} = c\}$.

8: Compute

$$v_{ic} = -k\log(1+n_c(i)^{-1}) - 2\sum_{j\neq i}^{n} A_{ij}\mu_{cz_j(i)^{[t]}}$$
$$+\delta^{[t-1]}\left[\sum_{j\neq i}\mu^2_{cz_j(i)^{[t]}} + \frac{1}{2}\mu^2_{cc}\right] + \frac{1}{2}\left(\sum_{r=1}^{k}\frac{n_r(i)}{n_r^{[t-1]}} + \frac{1}{n_c^{[t-1]}}\right)^2$$

9: **end for**

10: Set $z_i^{[t]} = \text{argmin}_{1\leq c\leq k}\{v_{ic}\}$.

11: **end for**

12: **Update a and δ:** compute $n_c^{[t]} = \sum_{i=1}^{n}\mathbb{I}\left\{z_i^{[t]} = c\right\}$, then

$$a^{[t]} = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mu^{[t-1]}_{z_i^{[t]}z_j^{[t]}}\right)^2 + \left(\delta^{[t-1]}\right)^{-1}\left(\sum_{c=1}^{k}\frac{n_c^{[t]}}{n_c^{[t-1]}}\right)^2, \quad \delta^{[t]} = 1+\sqrt{\frac{2\beta}{a^{[t]}}}$$

13: **Update μ and Σ:**

$$\mu_{cd}^{[t]} = \left(\delta^{[t]}\right)^{-1}\left(n_c^{[t]}\right)^{-1}\left(n_d^{[t]}\right)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij}\mathbb{I}\left\{z_i^{[t]} = c, z_j^{[t]} = d\right\}, \qquad 1\leq c,d\leq k,$$

$$\Sigma_{cd}^{[t]} = \left(\delta^{[t]}\right)^{-1}\left(n_c^{[t]}\right)^{-1}\left(n_d^{[t]}\right)^{-1}, \qquad 1\leq c,d\leq k.$$

14: **Update L_t and ∇L_t:**

$$L_t = \frac{k(k+1)}{4}\log\frac{\delta^{[t]}}{4\beta e^2} + \sqrt{\frac{a^{[t]}\beta}{2}} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij}\mu^{[t]}_{z_i^{[t]}z_j^{[t]}} + (D+1)\left(\frac{1}{2}k(k+1) + n\log k\right),$$

$$\nabla L_t = |L_{t-1} - L_t|.$$

15: **end while**

16: **Output:** $\widehat{z}^{(k)} = z^{[t]}$, $\widehat{a}^{(k)} = a^{[t]}$, $\widehat{\mu}^{(k)} = \mu^{[t]}$, $\widehat{\Sigma}^{(k)} = \Sigma^{[t]}$, $\widehat{L}^{(k)} = L_t$.

If we specialize the computational complexity from (3.26) for stochastic block model, it is $O(k^5(n^2+k^2))$ for each iteration. However, this is not true. The real computational complexity of Algorithm 4 for each iteration is $O(n^2k)$, which is much smaller than the order derived from the general algorithm. The main reason is that $\mathcal{X}_Z^T\mathcal{X}_Z$ is a diagonal matrix, and a lot of unnecessary computations can be removed.

In Algorithm 4, we can inductively show that $z^{[t]} \in \bar{\mathcal{Z}}_k$ all the time because once $n_c(i) = 0$, meaning no member is in cluster c excluding i, then $v_{ic} = -\infty$ and $z_i^{[t]} = c$. Therefore, during the iterations, each cluster must have at least one member as long as $z^{[0]} \in \bar{\mathcal{Z}}_k$.

3.4.2 Biclustering

Biclustering model can be viewed as an asymmetric version of stochastic block model. In the biclustering model, we assume $Y \in \mathbb{R}^{n\times m}$ is generated from a signal matrix $\theta = (\theta_{ij})$ with $\theta_{ij} = B_{z_{1i}z_{2j}}$ for some label vector $z_1 \in [k]^n$ and $z_2 \in [l]^m$. Our goal is to recover the labels z_1, z_2 and the true signal matrix θ^*.

Now we put the biclustering model into our general model framework. $Z = (z_1, z_2)$, $\tau = (k, l)$, $\mathcal{T} = [n] \times [m]$ and $\mathcal{Z}_{k,l} = [k]^n \times [l]^m$, $\ell_{k,l} = kl$, $|\bar{\mathcal{Z}}_{k,l}| \le |\mathcal{Z}_{k,l}| = k^n l^m$ and $\epsilon_{k,l} = kl + k\log n + l\log m$. Similarly, the non-degenerated structure set is given by $\bar{\mathcal{Z}}_{k,l} = \{(z_1, z_2) \in \mathcal{Z}_{k,l} : \sum_{i=1}^n \mathbb{I}\{z_{1i} = t\} > 0, \text{ for all } 1 \le t \le k \text{ and } \sum_{j=1}^m \mathbb{I}\{z_{1j} = s\} > 0, \text{ for all } 1 \le s \le l\}$.

The prior sampling procedure for biclustering model is given below:

1. Sample $(k, l) \sim \pi$ from $[n]\times[m]$, where $\pi(k, l) \propto \frac{\Gamma(kl)}{\Gamma(kl/2)} \exp\left(-D\left(kl + n\log k + m\log l\right)\right)$;
2. Conditioning on (k, l), sample (z_1, z_2) uniformly from $\bar{\mathcal{Z}}_{k,l}$;
3. Conditioning on (k, l), sample $\lambda \sim \text{IG}\left(\frac{kl+1}{2}, \beta\right)$;
4. Conditioning on $(k, l, z_1, z_2, \lambda)$, sample $B \sim f_{k,l,z_1,z_2,\lambda}$ with

$$f_{k,l,z_1,z_2,\lambda} \propto \exp\left(-\frac{\lambda}{2}\sum_{i=1}^n\sum_{j=1}^m B_{z_{1i}z_{2j}}^2\right);$$

5. Set $\theta_{ij} = B_{z_{1i}z_{2j}}$ for all $i \in [n]$ and $j \in [m]$.

Assume $\left(\widehat{k}, \widehat{l}, \widehat{Q}^{(\widehat{k},\widehat{l})}\right)$ is the solution of (3.11) to biclustering model. Theoretically, we have the following concentration result for the VB posterior distribution.

Corollary 3.4.2. *Assume $\theta^*_{ij} = B^*_{z^*_{1i}z^*_{2j}}$ for $B^* \in \mathbb{R}^{k^* \times l^*}$ and $(z^*_1, z^*_2) \in [k^*]^n \times [l^*]^m$ and $Y = \theta^* + W$ with W satisfying the condition (3.13), then*

$$P_{\theta^*}\widehat{Q}^{(\widehat{k},\widehat{l})}\|\theta - \theta^*\|_F^2 \leq M(k^*l^* + n\log k^* + m\log l^*),$$

for any $D > D_{\beta,\rho}$ and some constant M only depending on β, ρ, D.

Now we derive the VB algorithm to solve $(\widehat{k}, \widehat{l}, \widehat{Q}^{(\widehat{k},\widehat{l})})$ for biclustering model. We follow the same way to reformulate the model as discussed in the Section 3.4.1. In this model, Y is replaced by $\mathrm{Vec}(Y) \in \mathbb{R}^{nm}$, and B is replaced by $\mathrm{Vec}(B) \in \mathbb{R}^{kl}$. Then, if we assume $Z_1 \in [0,1]^{n\times k}$ and $Z_2 \in [0,1]^{m\times l}$ are the membership matrix corresponding to z_1 and z_2, then we have $\mathrm{Vec}(Y) = (Z_1 \otimes Z_2)\mathrm{Vec}(B) + W$. Thus, the design matrix for biclustering model is given by $\mathcal{X}_Z = Z_1 \otimes Z_2$ and $\mathcal{X}_Z^T\mathcal{X}_Z$ is also a diagonal matrix. Thus, we can also assume $B_{cd} \sim N(\mu_{cd}, \Sigma_{cd})$ for $1 \leq c \leq k$ and $1 \leq d \leq l$ and update them pointwise. For a given pair (k, l), the VB algorithm for the biclustering model is given in Algorithm 5.

Algorithm 5 Variational Algorithm for Biclustering Model

1: **Input:** Number of clusters (k, l), initial labels $\left(z_1^{[0]}, z_2^{[0]}\right) \in \bar{\mathcal{Z}}_{k,l}$, maximum iteration number M, tolerate level ϵ.

2: **Initialize:** Iteration number $t = 0$, objective function $L_0 = \infty$, change of objective function $\nabla L_0 = \infty$, $a^{[0]} = mn$ and $\delta^{[0]} = 1 + \sqrt{\frac{2\beta}{n^2}}$. The initial μ and Σ are computed by followings:

$$\mu_{cd}^{[0]} = \left(\delta^{[0]}\right)^{-1}(n_c^{[0]})^{-1}(m_d^{[0]})^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij}\mathbb{I}\{z_{1i}^{[0]} = c, z_{2j}^{[0]} = d\},$$

3:

$$\Sigma_{cd}^{[0]} = \left(\delta^{[0]}\right)^{-1} (n_c^{[0]})^{-1}(m_d^{[0]})^{-1}.$$

for $1 \le c \le k$ and $1 \le d \le l$, where $n_c^{[0]} = \sum_{i=1}^n \mathbb{I}\{z_{1i}^{[0]} = c\}$, $m_d^{[0]} = \sum_{i=1}^n \mathbb{I}\{z_{2i}^{[0]} = c\}$.

4: **while** $t < M$ and $\nabla L_t > \epsilon$ **do** $t \leftarrow t + 1$.

5: **Update** Z**:**

6: **for** $i = 1, \cdots, n$ in a randomized order **do**

7: **for** $c = 1, \cdots, k$ **do**

8: Set $n_c(i) = \sum_{j \neq i} \mathbb{I}\{z_{1j} = c\}$, where z_{1j} to be current label for j-th row.

$$\begin{aligned} v_{ic} &= -\frac{l}{2}\log(1 + n_c(i)^{-1}) \\ &\quad - \sum_{j=1}^m Y_{ij}\mu_{cz_{2j}^{[t-1]}} + \frac{1}{2}\delta^{[t-1]}\sum_{j=1}^m \mu^2_{cz_{2j}^{[t-1]}} + \frac{l}{2n_c^{[t-1]}}. \end{aligned}$$

9: where $z_j(i)^{[t]}$ denotes the current label for z_j such that $j \neq i$.

10: **end for**

11: Set $z_i^{[t]} = \text{argmin}_{1 \le c \le k}\{v_{ic}\}$.

12: **end for**

13: Compute $n_c^{[t]} = \sum_{i=1}^n \mathbb{I}\{z_{1i}^{[t]} = c\}$ for all $1 \le c \le k$.

14: **for** $j = 1, \cdots, m$ in a randomized order **do**

15: **for** $d = 1, \cdots, l$ **do**

16: Set $m_d(j) = \sum_{i \neq j} \mathbb{I}\{z_{2i} = d\}$. z_{2i} refers to current label for i-th column.

17: Compute

$$\begin{aligned} w_{jd} &= -\frac{k}{2}\log(1 + m_d(j))^{-1} \\ &\quad - \sum_{i=1}^n Y_{ij}\mu_{z_{1i}^{[t]}d} + \frac{1}{2}\delta^{[t-1]}\sum_{i=1}^n \mu^2_{z_{1i}^{[t]}d} + \frac{1}{2m_d^{[t-1]}}\sum_{c=1}^k \frac{n_c^{[t]}}{n_c^{[t-1]}}, \end{aligned}$$

end for

end for

Compute $m_d^{[t]} = \sum_{i=1}^{m} \mathbb{I}\{z_{2i}^{[t]} = d\}$ for all $1 \le d \le l$.

Update a and δ:

$$a^{[t]} = \sum_{i=1}^{n}\sum_{j=1}^{m} \mu^2_{z_{1i}^{[t]} z_{2j}^{[t]}} + \left(\delta^{[t-1]}\right)^{-1} \left(\sum_{c=1}^{k} \frac{n_c^{[t]}}{n_c^{[t-1]}}\right)\left(\sum_{d=1}^{l} \frac{m_d^{[t]}}{m_d^{[t-1]}}\right), \quad \delta^{[t]} = 1 + \sqrt{\frac{2\beta}{a^{[t]}}}$$

Update μ and Σ:

$$\mu_{cd}^{[t]} = \left(\delta^{[t]}\right)^{-1}\left(n_c^{[t]}\right)^{-1}\left(m_d^{[t]}\right)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij}\mathbb{I}\{z_{1i}^{[t]} = c, z_{2j}^{[t]} = d\},$$

$$\Sigma^{[t]} = \left(\delta^{[t]}\right)^{-1}\left(n_c^{[t]}\right)^{-1}\left(m_d^{[t]}\right)^{-1}.$$

Update L_t and ∇L_t:

$$L_t = \frac{kl}{2}\log\frac{\delta^{[t]}}{4\beta e^2} + \sqrt{\frac{a^{[t]}\beta}{2}} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} Y_{ij}\mu^{[t]}_{z_{1i}^{[t]} z_{1j}^{[t]}} + (D+1)\left(kl + n\log k + m\log l\right),$$

$$\nabla L_t = |L_{t-1} - L_t|.$$

end while

Output: $\widehat{z}^{(k,l)} = z^{[t]}$, $\widehat{a}^{(k,l)} = a^{[t]}$, $\widehat{\mu}^{(k,l)} = \mu^{[t]}$, $\widehat{\Sigma}^{(k,l)} = \Sigma^{[t]}$, $\widehat{L}^{(k,l)} = L_t$.

The computational complexity for each iteration of Algorithm 5 is $O((m+n)(kn+ml))$. With the same discussions as in Section 3.4.1 we can also show that $z^{[t]} \in \bar{\mathcal{Z}}_{k.l}$ during the iterations.

3.4.3 Sparse Linear Regression

In the sparse linear regression, we assume $Y = \theta + W$, where $\theta = X\gamma$ with $X \in \mathbb{R}^{n\times p}$ as a fixed design matrix. The regression coefficients are assumed to be sparse, i.e. $\gamma^T = (\gamma_S^T, \mathbf{0}_{S^c}^T)$ for some $S \in [p]$. Our goal is to recover θ and S from (X, Y).

Now we formulate the sparse linear regression model into form (3.1). For sparse linear regression, we can consider it as a special form of clustering problem, in which the number of nodes is s, whereas the number of clusters is p that could be much larger than s. However, we also need to assume that all the nodes have different labels so that s distinct columns in the design matrix are selected. Thus, in this model, $\tau = s$, $\mathcal{T} = [p]$, $Z = z = (z_1, \cdots, z_s)$, $z_i \in [p]$, $\ell_s = s$, $\mathcal{Z}_s = \left\{z \in [p]^s : z_i \neq z_j \text{ for all } i \neq j\right\}$. Note that $|\mathcal{Z}_s| = p(p-1)\cdots(p-s+1)$, we can choose $\epsilon_s = s + s\log p > \ell_s + \log|\mathcal{Z}_s|$. Assume the design matrix $X = \left(\vec{X}_1, \cdots, \vec{X}_p\right)$, then $\mathcal{X}_Z = X_Z = \left(\vec{X}_{z_1}, \vec{X}_{z_2}, \cdots, \vec{X}_{z_s}\right)$, $B = \gamma_S \in \mathbb{R}^s$ and $\theta = \mathcal{X}_Z B$.

The prior sampling procedure for the sparse linear regression model is given below:

1. Sample $s \sim \pi$ from $[p]$, where $\pi(s) \propto \frac{\Gamma(s)}{\Gamma(s/2)} \exp\left(-Ds\left(1 + \log p\right)\right)$;
2. Conditioning on s, sample z uniformly from $\left\{z \in \mathcal{Z}_s : \det\left(X_Z^T X_Z\right) > 0\right\}$;
3. Conditioning on s, sample $\lambda \sim \mathrm{IG}((s+1)/2, \beta)$;
4. Conditioning on (s, z, λ), sample $B \sim f_{s,z,\lambda}$ with

$$f_{s,z,\lambda} \propto \exp\left(-\frac{\lambda}{2}\|\mathcal{X}_Z B\|^2\right);$$

5. Set $\theta = X_Z B$.

Assume $(\widehat{s}, \widehat{Q}^{(\widehat{s})})$ is the solution of (3.11) to sparse regression model, then we have the following result.

Corollary 3.4.3. *Assume $\theta^* = X_{Z^*}B^*$ for $B^* \in \mathbb{R}^s$, $Z^* = z^*$ and $z^* \in \bar{\mathcal{Z}}_{s^*}$. $Y = \theta^* + W$ with W satisfying the condition (3.13), then*

$$P_{\theta^*}\widehat{Q}^{(\widehat{s})}\|\theta - \theta^*\|^2 \leq Ms^* \log p,$$

for any $D > D_{\beta,\rho}$ and some constant M only depending on β, ρ, D.

In sparse linear regression, the Gram matrix $\mathcal{X}_Z^T\mathcal{X}_Z = X_Z^T X_Z$ is no longer diagonal, so we need to store the whole matrix Σ. The VB algorithm for sparse linear regression for a given s is Algorithm 6.

Algorithm 6 Variational Algorithm for Sparse Linear Regression Model

1: **Input:** sparsity level s, initial support $z^{[0]} \in \bar{\mathcal{Z}}_s$, maximum iteration number M, tolerate level ϵ.

2: **Initialize:** Iteration number $t = 0$, objective function $L_0 = \infty$, change of objective function $\nabla L_0 = \infty$. The initial a, δ, μ and Σ are computed by followings:

$$a^{[0]} = \sum_{j=1}^{s} \|\vec{X_{z_j}}\|^2, \qquad \delta^{[0]} = 1 + \sqrt{\frac{2\beta}{a^{[0]}}}$$

$$\mu^{[0]} = \left(\delta^{[0]}\right)^{-1} (X_{Z^{[0]}}^T X_{Z^{[0]}})^{-1} X_{Z^{[0]}}^T Y, \qquad \Sigma^{[0]} = \left(\delta^{[0]}\right)^{-1} (X_{Z^{[0]}}^T X_{Z^{[0]}})^{-1}.$$

3: **while** $t < M$ and $\nabla L_t > \epsilon$ **do** $t \leftarrow t + 1$

4: **Update Z:**

5: **for** $j = 1, \cdots, s$ in a randomized order **do**

6: Assume $X_{Z(-j)} \in \mathbb{R}^{n\times(s-1)}$ is the current design matrix without j-th column.

7: Compute the projection matrix $H_{Z(-j)} = X_{Z(-j)} \left(X_{Z(-j)}^T X_{Z(-j)}\right)^{-1} X_{Z(-j)}^T$.

8: **for** $c = 1, \cdots, p$ **do**

9: Compute

$$\begin{aligned} v_{jc} &= -\frac{1}{2}\log\left(\|X_c\|^2 - X_c^T H_{Z(-j)} X_c\right) - \mu_j^{[t-1]} Y^T X_c \\ &+\frac{1}{2}\delta^{[t-1]}\left[\left(\mu_j^{[t-1]}\right)^2 \|X_c\|^2 + 2\mu_j^{[t-1]}(X_{Z(-j)}\mu_{-j})^T X_c\right] \\ &+\frac{1}{2}\delta^{[t-1]}\left[2\Sigma_{j,-j}^{[t-1]} X_{Z(-j)}^T X_c + \Sigma_{j,j}^{[t-1]}\|X_c\|^2\right] \end{aligned}$$

10: **end for**

$$z_j^{[t]} = \underset{1\le c\le p}{\operatorname{argmin}}\{v_{jc}\}.$$

11: **end for**

12: **Update a and δ:**

$$a^{[t]} = \|X_{Z^{[t]}}\mu^{[t-1]}\|^2 + \mathsf{Tr}\left(\Sigma^{[t-1]} X_{Z^{[t]}}^T X_{Z^{[t]}}\right), \qquad \delta^{[t]} = 1 + \sqrt{\frac{2\beta}{a^{[t]}}}.$$

13: **Update μ and Σ:**

$$\mu^{[t]} = \left(\delta^{[t]}\right)^{-1}(X_{Z^{[t]}}^T X_{Z^{[t]}})^{-1} X_{Z^{[t]}}^T Y, \qquad \Sigma^{[t]} = \left(\delta^{[t]}\right)^{-1}(X_{Z^{[t]}}^T X_{Z^{[t]}})^{-1}.$$

14: **Compute L_t and ∇L_t:**

$$L_t = \frac{s}{2}\log\frac{\delta^{[t]}}{4\beta e^2} + \sqrt{\frac{a^{[t]}\beta}{2}} - \frac{1}{2}Y^T X_{Z^{[t]}}\mu^{[t]} + (D+1)\left[s + s\log\left(\frac{ep}{s}\right)\right],$$

$$\nabla L_t = |L_{t-1} - L_t|.$$

15: **end while**

16: **Output:** $\widehat{z}^{(s)} = z^{[t]}$, $\widehat{a}^{(s)} = a^{[t]}$, $\widehat{\mu}^{(s)} = \mu^{[t]}$, $\widehat{\Sigma}^{(s)} = \Sigma^{[t]}$, $\widehat{L}^{(s)} = L_t$.

For each iteration, the computational complexity of Algorithm 6 is $O\left(s^4 + (n+s)s^2 p)\right)$. As $H_{Z(-j)}$ is a projection matrix, $\|X_c\|^2 - X_c^T H_{Z(-j)} X_c = \|X_c - H_{Z(-j)} X_c\|^2$. Thus, If

X_c can be linearly represented by $X_{Z(-j)}$, then $X_c - H_{Z(-j)}X_c = 0$, then $v_{jc} = \infty$ and therefore c cannot be selected as $z_j^{[t]}$. Thus, when $z \in \bar{\mathcal{Z}}_s$ before z_j is updated, then it will also be so after z_j is updated. Therefore, iteratively, we can show that $z^{[t]} \in \bar{\mathcal{Z}}_s$.

3.4.4 Multiple Linear Regression with Group Sparsity

Now we consider multiple linear regression with group sparsity proposed in [70]. In this model, $Y = \Theta + W$, where $\Theta = X\Gamma$ and $X \in \mathbb{R}^{n\times p}$, $\Gamma \in \mathbb{R}^{p\times m}$, $W \in \mathbb{R}^{n\times m}$. We assume that there exists a subset $S \subset [p]$ such that $\Gamma^T = \left(\Gamma_S^T, \mathbf{0}_{S^c}^T\right)$, where S denotes the index set of non-zero rows in parameter matrix Γ. When $m = 1$, the multiple linear regression with group sparsity reduces to the ordinary sparse linear regression problem. Our goal is to recover Θ, B and S from (X, Y).

Now we formulate this problem into our general formula (3.1), which follows a similar way as the sparse linear regression model. In this case, $\tau = s$, $\mathcal{T} = [p]$, $Z = (z_1, \cdots, z_s)$, $z_i \in [p]$, $\ell_s = ms$, $\mathcal{Z}_s = \left\{z \in [p]^s : z_i \neq z_j, \text{ for all } i \neq j\right\}$. Similarly, we can choose $\epsilon_s = s\,(m + \log p)$. The prior sampling procedure specialized in multiple linear regression with group sparsity is given as follows:

1. Sample $s \sim \pi$ from $[p]$, where $\pi(s) \propto \frac{\Gamma(s)}{\Gamma(s/2)} \exp\left(-Ds\,(m + \log p)\right)$;
2. Conditioning on s, sample S uniformly from $\bar{\mathcal{Z}}_s = \left\{z \in \mathcal{Z}_s : \det\left(X_Z^T X_Z\right) > 0\right\}$;
3. Conditioning on s, sample λ from $\mathrm{IG}((ms+1)/2, \beta)$;
4. Conditioning on (s, z, λ), sample $B \in \mathbb{R}^{s\times m}$ from $f_{s,z,\lambda}$ with

$$f_{s,z,\lambda} \propto \exp\left(-\frac{\lambda}{2}\|X_Z B\|_F^2\right);$$

5. Set $\Theta = X_Z B$.

Assume $\left(\widehat{s}, \widehat{Q}^{(\widehat{s})}\right)$ is the solution of (3.11) to multiple linear regression model with group sparsity, then

Corollary 3.4.4. *Assume* $\Theta^* = X_{Z^*}B^* \in \mathbb{R}^{n\times m}$ *for* $B^* \in \mathbb{R}^{s\times m}$, $Z^* = z^*$ *and* $z^* \in \bar{\mathcal{Z}}_{s^*}$. $Y = \Theta^* + W$ *with* W *satisfying the condition (3.13), then*

$$P_{\Theta^*}\widehat{Q}^{(\widehat{s})}\|\Theta - \Theta^*\|_F^2 \leq Ms^*\left(m + \log p\right),$$

for any $D > D_{\beta,\rho}$ *and some constant* M *only depending on* β, ρ, D.

To derive the algorithm, we first need to vectorize the response Y, the signal $X_Z B$ and the parameter B. We replace Y by $\mathrm{Vec}(Y) \in \mathbb{R}^{nm}$, Θ by $\mathrm{Vec}(\Theta) \in \mathbb{R}^{nm}$ and B by $\mathrm{Vec}(B) \in \mathbb{R}^{ms}$. Then we have $\mathrm{Vec}(\Theta) = (I_m \otimes X_Z)\mathrm{Vec}(B)$. Thus, the design matrix in the multiple regression model with group sparsity can be written as $\mathcal{X}_Z = I_m \otimes X_Z$.

In this model, $\mathcal{X}_Z^T\mathcal{X}_Z = I_m \otimes \left(X_Z^T X_Z\right)$. Combining with the update procedure in Algorithm 2, we can assume each row of B follows a normal distribution independently with the same covariance matrix. Thus, we can assume the variational posterior distribution $B = (B_1, \cdots, B_m)$ is $B_i \sim N(\mu_i, \Sigma)$ independently. Then we only need to update $\mu = (\mu_1, \cdots, \mu_s) \in \mathbb{R}^{s\times m}$ and $\Sigma \in \mathbb{R}^{s\times s}$.

Algorithm 7 Variational Algorithm for Multiple Linear Regression with Group Sparsity

1: **Input:** sparsity level s, initial support $z^{[0]} \in \bar{\mathcal{Z}}_s$, maximum iteration number M, tolerate level ϵ.

2: **Initialize:** Iteration number $t = 0$, objective function $L_0 = \infty$, change of objective function $\nabla L_0 = \infty$. The initial a, δ, μ and Σ are computed by followings:

$$a^{[0]} = m\sum_{j=1}^{s}\|\vec{X_{z_j}}\|^2, \qquad \delta^{[0]} = 1 + \sqrt{\frac{2\beta}{a^{[0]}}},$$

3:

$$\mu^{[0]} = \left(\delta^{[0]}\right)^{-1} \left(X_{Z^{[0]}}^T X_{Z^{[0]}}\right)^{-1} X_{Z^{[0]}}^T Y, \qquad \Sigma^{[0]} = \left(\delta^{[0]}\right)^{-1} \left(X_{Z^{[0]}}^T X_{Z^{[0]}}\right)^{-1}.$$

4: **while** $t < M$ and $\nabla L_t > \epsilon$ **do** $t \leftarrow t + 1$.

5: **Update Z:**

6: **for** $j = 1, \cdots, s$ in a randomized order **do**

7: Assume $X_{Z(-j)} \in \mathbb{R}^{n \times (s-1)}$ is the current design matrix without j-th column.

8: Compute the projection matrix $H_{Z(-j)} = X_{Z(-j)} \left(X_{Z(-j)}^T X_{Z(-j)}\right) X_{Z(-j)}^T$.

9: **for** $c = 1, \cdots, p$ **do**

10: Assume the j-th row vector for $\mu^{[t-1]}$ is $\mu_{j\cdot}^{[t-1]} \in \mathbb{R}^{1 \times m}$, other columns are

11: $\mu_{-j\cdot}^{[t-1]}$. $X_{Z(-j)}$ is the current design matrix without j-th column. Compute

12:

$$\begin{aligned} v_{jc} \;=\; & -\frac{m}{2} \log\left(\|X_c\|^2 - X_c^T H_{Z(-j)} X_c\right) - \mu_{j\cdot}^{[t-1]} Y^T X_c \\ & + \frac{1}{2}\delta^{[t-1]} \left[\left\|\mu_{j\cdot}^{[t-1]}\right\|^2 \|X_c\|^2 + 2\mu_{j\cdot}^{[t-1]} \left(X_{Z(-j)} \mu_{-j\cdot}\right)^T X_c\right] \\ & + \frac{1}{2}\delta^{[t-1]} \left[2\Sigma_{j,-j}^{[t-1]} X_{Z(-j)}^T X_c + \Sigma_{j,j}^{[t-1]} \|X_c\|^2\right] \end{aligned}$$

13: **end for**

$$z_j^{[t]} = \underset{1 \le c \le p}{\operatorname{argmin}} \{v_{jc}\}.$$

14: **end for**

15: **Update a and δ:**

$$a^{[t]} = \left\|X_{Z^{[t]}} \mu^{[t-1]}\right\|_F^2 + \mathsf{Tr}\left(\Sigma^{[t-1]} X_{Z^{[t]}}^T X_{Z^{[t]}}\right), \qquad \delta^{[t]} = 1 + \sqrt{\frac{2\beta}{a^{[t]}}}.$$

16: **Update μ and Σ:**

$$\mu^{[t]} = \left(\delta^{[t]}\right)^{-1} (X_{Z^{[t]}}^T X_{Z^{[t]}})^{-1} X_{Z^{[t]}}^T Y, \qquad \Sigma^{[t]} = \left(\delta^{[t]}\right)^{-1} (X_{Z^{[t]}}^T X_{Z^{[t]}})^{-1}.$$

17: **Compute L_t and ∇L_t:**

$$L_t = \frac{ms}{2}\log\frac{\delta^{[t]}}{4\beta e^2} + \sqrt{\frac{a^{[t]}\beta}{2}} - \frac{1}{2}Y^T X_{Z^{[t]}}\mu^{[t]} + (D+1)s\left[m + \log\left(\frac{ep}{s}\right)\right],$$

$$\nabla L_t = |L_{t-1} - L_t|.$$

18: **end while**

19: **Output:** $\widehat{z}^{(s)} = z^{[t]}$, $\widehat{a}^{(s)} = a^{[t]}$, $\widehat{\mu}^{(s)} = \mu^{[t]}$, $\widehat{\Sigma}^{(s)} = \Sigma^{[t]}$, $\widehat{L}^{(s)} = L_t$.

For each iteration, the computational complexity for Algorithm 7 is $O\left(s^4 + s^2 p(nm+s)\right)$ With the same arguments in Section 3.4.3, we can prove $z^{[t]} \in \bar{\mathcal{Z}}_s$ iteratively.

3.4.5 Multi-task Learning

In the multiple-task learning, we assume $Y = \Theta + W$, with $\Theta = X\Gamma$ and $X \in \mathbb{R}^{n\times p}$ and $\Gamma \in \mathbb{R}^{p\times m}$. However, in this model, the row vectors of the coefficient matrix $\Gamma = (\gamma_1, \cdots, \gamma_m)$ share a clustering structure, i.e. $\gamma_j = B_{*z_j}$. When the design matrix X is an identity matrix, it reduces to an ordinary multivariate clustering problem. We assume the design matrix has full column rank, i.e. $\det(X^T X) > 0$.

In this case, $Z = z$, $\tau = k$, $\mathcal{T} = m$, $B \in \mathbb{R}^{p\times k}$, $\mathcal{Z}_k = [k]^m$ and $\ell_k = pk$, so we can choose $\epsilon_k = pk + m\log k$ in this model. Recall the technique we applied for the previous cases, we can replace Y by its vectorization $\mathrm{Vec}(Y)$, Θ by $\mathrm{Vec}(\Theta)$ and B by $\mathrm{Vec}(B)$. Then the model is reformulated as $\mathrm{Vec}(\Theta) = (Z \otimes X)\mathrm{Vec}(B)$, where $Z = (\mathbb{I}\{z_i = j\})_{1\le i\le n, 1\le j\le k} \in [0,1]^{m\times k}$ is the membership matrix corresponding to z. Then the design matrix $\mathcal{X}_Z = Z \otimes X$. When $\det(X^T X) > 0$, $\det(\mathcal{X}_Z^T \mathcal{X}_Z) > 0$ is equivalent to $\det(Z^T Z) > 0$. Therefore, $\bar{\mathcal{Z}}_k = \{z \in \mathcal{Z}_k : \sum_{i=1}^n \mathbb{I}\{z_i = t\} > 0, \text{ for all } 1 \le t \le k\}$. The prior sampling procedure for a multi-task learning model is given as follows:

1. Sample $k \sim \pi$ from $[m]$, where $\pi(k) \propto \frac{\Gamma(pk)}{\Gamma(pk/2)}\exp\left(-D\left(pk + m\log k\right)\right)$;

2. Conditioning on k, sample z uniformly from $\bar{\mathcal{Z}}_k$;

3. Conditioning on k, sample λ from $\mathrm{IG}((pk+1)/2, \beta)$;

4. Conditioning on (s, z, λ), sample $B \in \mathbb{R}^{p\times k}$ from $f_{k,z,\lambda}$ with

$$f_{k,z,\lambda} \propto \exp\left(-\frac{\lambda}{2}\sum_{j=1}^{m}\|XB_{\cdot z_j}\|^2\right);$$

5. Set $\Gamma = (\gamma_1, \cdots, \gamma_m)$ with $\gamma_i = B_{\cdot z_j}$ for $1 \le i \le m$ and $\Theta = X\Gamma$.

Assume $(\widehat{k}, \widehat{Q}^{(\widehat{k})})$ is the solution of (3.11) to multi-task learning, then the convergence rate of the variational posterior distribution is given by the following corollary.

Corollary 3.4.5. *Assume $\Theta^* = X\Gamma^* \in \mathbb{R}^{n\times m}$ for $X \in \mathbb{R}^{n\times p}$, $\Gamma^* \in \mathbb{R}^{p\times m}$ with $\Gamma^*_{\cdot j} = B^*_{\cdot z_j^*}$ and $z^* \in \bar{\mathcal{Z}}_k$. $Y = \Theta^* + W$ with W satisfying the condition (3.13), then*

$$P_{\Theta^*}\widehat{Q}^{(\widehat{k})}\|\Theta - \Theta^*\|_F^2 \le M(pk^* + m\log k^*),$$

for any $D > D_{\beta,\rho}$ and some constant M only depending on β, ρ, D.

In this case, $\mathcal{X}_Z = Z \otimes X$. Therefore, $\mathcal{X}_Z^T\mathcal{X}_Z = (Z^TZ) \otimes (X^TX)$. As Z^TZ is diagonal matrix and X^TX does not rely on Z. We can assume the variational distribution for $B = (B_1, \cdots, B_k)$ is $B_c \sim N(\mu_c, \Sigma_c)$ independently. The VB algorithm for multi-task learning model is given in Algorithm 8.

Algorithm 8 Variational Algorithm for Multi-task Learning Model

1: **Input:** number of different tasks k, initial label $z^{[0]} \in \bar{\mathcal{Z}}_s$, maximum iteration number M, tolerate level ϵ.

2: **Initialize:** Iteration number $t = 0$, objective function $L_0 = \infty$, change of objective function $\nabla L_0 = \infty$. The initial a, δ, μ and Σ are computed by followings:

3:

$$a^{[0]} = m\sum_{j=1}^{s} \|\vec{X}_{z_j^{[0]}}\|^2, \qquad \delta^{[0]} = 1 + \sqrt{\frac{2\beta}{a^{[0]}}},$$

$$\mu_c^{[0]} = \left(\delta^{[0]}\right)^{-1} \left(n_c^{[0]}\right)^{-1} \sum_{i=1}^{n} \left(X^T X\right)^{-1} X^T Y_{\cdot i} \mathbb{I}\{z_i^{[0]} = c\},$$

$$\Sigma_c^{[0]} = \left(\delta^{[0]}\right)^{-1} \left(n_c^{[0]}\right)^{-1} (X^T X)^{-1}.$$

4: **while** $t < M$ and $\nabla L_t > \epsilon$ **do** $t \leftarrow t + 1$

5: **Update Z:**

6: **for** $i = 1, \cdots, m$ in a randomized order **do**

7: **for** $c = 1, \cdots, k$ **do**

8: Set $z_j(i)^{[t]}$ to be the current label for z_j and $n_c(i) = \sum_{j \neq i} \mathbb{I}\{z_j(i)^{[t]} = c\}$.

9: Compute

$$v_{ic} \;=\; -\frac{p}{2}\log(1 + n_c(i)^{-1}) - Y_{\cdot i}^T X\mu_c + \frac{1}{2}\delta^{[t-1]}\|X\mu_c\|^2 + \frac{p}{2n_c^{[t-1]}}.$$

10: **end for**

11: Set $z_i^{[t]} = \text{argmin}_{1 \leq c \leq k}\{v_{ic}\}$.

12: **end for**

13: Compute $n_c^{[t]} = \sum_{i=1}^{n} \mathbb{I}\{z_i^{[t]} = c\}$.

14: **Update a and δ:**

$$a^{[t]} = \sum_{i=1}^{m} \|X\mu_{z_i^{[t]}}\|^2 + \left(\delta^{[t-1]}\right)^{-1} p \sum_{c=1}^{k} \frac{n_c^{[t]}}{n_c^{[t-1]}}, \qquad \delta^{[t]} = 1 + \sqrt{\frac{2\beta}{a^{[t]}}}.$$

15: **Update μ and Σ:**

$$\mu_c^{[t]} = \left(\delta^{[t]}\right)^{-1} \left(n_c^{[t]}\right)^{-1} \sum_{i=1}^{n} \left(X^T X\right)^{-1} X^T Y_{\cdot i} \mathbb{I}\{z_i^{[t]} = c\},$$

$$\Sigma_c^{[t]} = \left(\delta^{[t]}\right)^{-1} \left(n_c^{[t]}\right)^{-1} (X^T X)^{-1}.$$

16: **Compute L_t and ∇L_t:**

$$L_t = \frac{kp}{2} \log \frac{\widehat{\delta}^{[t]}}{4\beta e^2} + \sqrt{\frac{\widehat{a}^{[t]}\beta}{2}} - \frac{1}{2} \mathsf{Tr}\left(Y^T X \mu_{z^{[t]}}^{[t]}\right) + (D+1)\left[pk + m \log k\right],$$

$$\nabla L_t = |L_{t-1} - L_t|.$$

17: **end while**

18: **Output:** $\widehat{z}^{(k)} = z^{[t]}$, $\widehat{a}^{(s)} = a^{[t]}$, $\widehat{\mu}^{(k)} = \mu^{[t]}$, $\widehat{\Sigma}^{(k)} = \Sigma^{[t]}$, $\widehat{L}^{(k)} = L_t$.

The computational complexity for each iteration is $O\left(mk(m+np)\right)$. As $\det\left(\mathcal{X}_Z^T \mathcal{X}_Z\right) > 0$ is equivalent to $\det\left(Z^T Z\right) > 0$, with the same discussion as in Section 3.4.1 we can also show that $z^{[t]} \in \bar{\mathcal{Z}}_k$ during the iterations.

3.4.6 Dictionary Learning

As discussed in [28], we will use the discrete version to set up the dictionary learning model. Assume the signal matrix $\theta \in \mathbb{R}^{n\times d}$ can be represented as $\theta = BZ$ for $B \in \mathbb{R}^{n\times p}$ and $Z \in \mathcal{Z}_{p,s} = \left\{Z \in \{-1,0,1\}^{p\times d} : \max_{j\in[d]} |\mathrm{supp}(Z_{\cdot j})| \le s\right\}$. In the dictionary learning, we have $\tau = (p,s)$, $\mathcal{T} = \{(p,s) \in [n \wedge d] \times [n] : s \le p\}$, $\ell_{p,s} = np$ and $|\mathcal{Z}_{p,s}| = \left(\sum_{t=1}^{s} \binom{p}{t} 2^t\right)^d \le (2p)^{(s+1)d}$. We can choose $\epsilon_{p,s} = np + 4ds \log p$. This model can also be reformulated as $\mathrm{Vec}(\theta^T) = (I_n \otimes Z^T)\mathrm{Vec}(B^T)$. Then $\mathcal{X}_Z = I_n \otimes Z^T$ in this case and $\bar{\mathcal{Z}}_{p,s} = \left\{Z \in \mathcal{Z}_{p,s} : \det(ZZ^T) > 0\right\}$. The general prior sampling procedure for dictionary learning is given by:

1. Sample $(p, s) \sim \pi$ from $\mathcal{T}$ with $\pi(p, s) \propto \frac{\Gamma(np)}{\Gamma(np/2)} \exp\left(-D\left(np + 4ds \log p\right)\right)$;

2. Given (p, s), sample Z uniformly from $\bar{\mathcal{Z}}_{p,s}$;

3. Given (p, s), sample λ from $\mathrm{IG}((np+1)/2, \beta)$;

4. Conditioning on (p, s, Z, λ), sample $B \sim f_{p,s,Z,\lambda}$ with

$$f_{p,s,Z,\lambda} \propto \exp\left(-\frac{\lambda}{2}\|BZ\|_F^2\right);$$

5. Set $\theta = BZ$.

Assume $(\widehat{p}, \widehat{s}, \widehat{Q}^{(\widehat{p},\widehat{s})})$ is the solution of (3.11) to dictionary learning model, then we have the following result.

Corollary 3.4.6. *Assume $\theta^* = B^* Z^*$ with $Z^* \in \bar{\mathcal{Z}}_{p^*,s^*}$ and $Y = \theta^* + W$ with W satisfying the condition (3.13), then*

$$P_{\theta^*} \widehat{Q}^{(\widehat{p},\widehat{s})} \|\theta - \theta^*\|_F^2 \leq M\left(np^* + ds^* \log p^*\right),$$

for any $D > D_{\beta,\rho}$ and some constant M only depending on β, ρ, D.

Note that for $\mathcal{X}_Z = I_n \otimes Z^T$, $\mathcal{X}_Z^T \mathcal{X}_Z = I_n \otimes (ZZ^T)$. Therefore, we can assume the variational distribution for $B = (b_1, \cdots, b_n)^T \in \mathbb{R}^{n \times p}$ is $b_i \sim N(\mu_i, \Sigma)$ independently. Then we only need to update $\mu = (\mu_1, \cdots, \mu_n)^T \in \mathbb{R}^{n \times p}$ and $\Sigma \in \mathbb{R}^{p \times p}$. The detailed algorithm is given in Algorithm 9.

Algorithm 9 Variational Algorithm for Dictionary Learning Model

1: **Input:** dimension parameter (p, s), initial state $Z^{[0]} \in \bar{\mathcal{Z}}_s$, maximum iteration number M, tolerate level ϵ.

2: **Initialize:** Iteration number $t = 0$, objective function $L_0 = \infty$, change of objective function $\nabla L_0 = \infty$. The initial a, δ, μ and Σ are computed by followings:

3:

$$a^{[0]} = n\,\mathsf{Tr}(ZZ^T), \qquad \delta^{[0]} = 1 + \sqrt{\frac{2\beta}{a^{[0]}}},$$

$$\mu^{[0]} = \left(\delta^{[0]}\right)^{-1} Y \left(Z^{[0]}\right)^T \left(Z^{[0]}\left(Z^{[0]}\right)^T\right)^{-1}, \qquad \Sigma^{[0]} = \left(\delta^{[0]}\right)^{-1} \left(Z^{[0]}\left(Z^{[0]}\right)^T\right)^{-1}.$$

4: **while** $t < M$ and $\nabla L_t > \epsilon$ **do** $t \leftarrow t + 1$

5: **Update** Z**:**

6: **for** $i = 1, \cdots, n$ in a randomized order **do**

7: **for** $j = 1, \cdots, d$ in a randomized order **do**

8: Denote $Z(i,j)$ denotes current matrix Z before its (i,j) entry being

9: updated.

10: **if** $\sum_{r \neq j} Z_{ir}(i,j) \geq s$ **then** Set $Z_{ij}^{[t]} = 0$.

11: **else**

12: Denote $Z(c;i,j)$ denotes current matrix Z with its (i,j) entry

13: replaced by c.

14: For $c = -1, 0, 1$, compute

$$v_{i,j,c} = -\frac{n}{2}\log\det\left(Z(c;i,j)^{[t]}\left(Z(c;i,j)^{[t]}\right)^T\right) - \mathsf{Tr}\left(Y^T\mu^{[t-1]}Z(c;i,j)^{[t]}\right)$$
$$+\frac{1}{2}\delta^{[t-1]}\left[\|\mu^{[t-1]}Z(c;i,j)\|_F^2 + n\,\mathsf{Tr}\left(\Sigma^{[t-1]}Z(c;i,j)^{[t]}\left(Z(c;i,j)^{[t]}\right)^T\right)\right].$$

15: Set $Z_{ij}^{[t]} = \mathrm{argmin}_{c\in\{-1,0,1\}}\left\{v_{i,j,c}\right\}$.

16: **end if**

17: **end for**

18: **end for**

19: **Update** a **and** δ**:**

$$a^{[t]} = \|\mu^{[t-1]}Z^{[t]}\|_F^2 + \left(\delta^{[t-1]}\right)^{-1} n\,\mathsf{Tr}\left(\Sigma^{[t-1]}Z^{[t]}\left(Z^{[t]}\right)^T\right), \qquad \delta^{[t]} = 1 + \sqrt{\frac{2\beta}{a^{[t]}}}.$$

20: **Update μ and Σ:**

$$\mu^{[t]} = \left(\delta^{[t]}\right)^{-1} Y \left(Z^{[t]}\right)^T \left(Z^{[t]} \left(Z^{[t]}\right)^T\right)^{-1}, \qquad \Sigma^{[t]} = \left(\delta^{[t]}\right)^{-1} \left(Z^{[t]} \left(Z^{[t]}\right)^T\right)^{-1}.$$

21: **Compute L_t and ∇L_t:**

$$L_t = \frac{np}{2} \log \frac{\widehat{\delta}^{[t]}}{4\beta e^2} + \sqrt{\frac{\widehat{a}^{[t]}\beta}{2}} - \frac{1}{2} \mathsf{Tr}\left(Y^T \mu^{[t]} Z^{[t]}\right) + (D+1)\left[np + 3ds \log \frac{ep}{s}\right],$$

$$\nabla L_t = |L_{t-1} - L_t|.$$

22: **end while**

23: **Output:** $\widehat{z}^{(k)} = z^{[t]}$, $\widehat{a}^{(k)} = a^{[t]}$, $\widehat{\mu}^{(k)} = \mu^{[t]}$, $\widehat{\Sigma}^{(k)} = \Sigma^{[t]}$, $\widehat{L}^{(k)} = L_t$.

The computational complexity of each iteration for this algorithm is $O(npd(pd+p^2+nd))$. However, in reality, this complexity can be smaller as when $\sum_{r\neq j} Z_{ir}(i,j) \geq s$, we don't need any calculations to update Z_{ij}. During the iteration, if $\det\left(Z(c;i,j)^{[t]} \left(Z(c;i,j)\right)^T\right) = 0$, then $v_{i,j,c} = \infty$ and c cannot be chosen as $Z_{ij}^{[t]}$. Thus, iteratively, we will have $Z^{[t]} \in \bar{\mathcal{Z}}_{p,s}$.

3.5 Simulations

In this section, we apply our novel VB algorithm in the stochastic block model and sparse linear regression. To apply this algorithm in practice, we need to choose β and D in the prior at first. In Theorem 3.2.1, D is required to be large and β can be arbitrary. However, in practice, we find small β and D can work better. When β and D are large, the selected model tends to be oversimplified, and the obtained estimator from the VB algorithm has a large bias to 0. Throughout this chapter, we choose $\beta = D = 0.1$ for all simulations.

We conduct two experiments for each model. In the first experiment for each model, we assume the true hyper-parameter τ^* is known, and we apply Algorithm 2. In the second experiment for each model, τ^* is assumed to be unknown. Then Algorithm 3 is applied to

select the optimal $\widehat{\tau}$ and compute its corresponding parameters in the variational posterior distribution.

As the procedures in Algorithm 2 are random. For each parameter setting, we run Algorithm 2 R times. The choice of replication time R is very important. If R is chosen to be small, the algorithm may be stuck in a sub-optimal point. If R is too large, the algorithm becomes slow. In experiments, we find the super large R does not improve the performance of the algorithm too much. We choose $R = 10$ in the stochastic block model and $R = 5$ in the sparse linear regression.

3.5.1 Stochastic Block Model

We first conduct a small scale of simulations for the stochastic block model. The number of nodes is set to be $n = 100$. The number of clusters is $k^* = 5$. For the first group of simulations, we assume $k^* = 5$ is known and only apply the Algorithm 4 without model selection. For the second group of simulations, we further apply the Algorithm 3 to select the best $\widehat{k}$.

We use the spectral method to initialize the label $Z^{[0]}$ for each k. It is given in Algorithm 10.

Algorithm 10 Spectral Method to Initialize $Z^{[0]}$ for SBM

1: **Input:** A symmetric adjacency matrix A, a number of clusters k.
2: Perform sparse eigenvalue decomposition on A. Assume its first k eigenvectors are $U_k = (u_1, \cdots, u_k)$.
3: Apply k-means algorithm on the row vectors of U_k to get the label $z^{[0]}$.
4: **Output:** Initial label $z^{[0]}$.

As the results of k-means for number of cluster k always contain k different labels, $Z^{[0]} \in \bar{\mathcal{Z}}_k$ is automatically satisfied in the beginning of the algorithm.

Experiment 1

In this experiment, assume $B_{cd} = p$ when $c = d$ and $B_{cd} = q$ when $c \neq d$. For (p, q), we consider the following scenarios:

- dense network with strong signals: $(p, q) = (0.8, 0.2)$;
- dense network with weak signals: $(p, q) = (0.6, 0.3)$.
- sparse network with weak signals: $(p, q) = (0.25, 0.05)$;
- sparse network with strong signals: $(p, q) = (0.5, 0.05)$;

Suppose the number of nodes in each cluster is $(n_1, \cdots, n_5)$. Then we consider the following two cases:

- balanced network: $(n_1, \cdots, n_5) = (20, 20, 20, 20, 20)$;
- unbalanced network: $(n_1, \cdots, n_5) = (5, 10, 40, 40, 5)$.

In simulations, we choose the repeat number $R = 10$ and maximum iteration number $M = 100$, and the tolerate level $\epsilon = 1 \times 10^{-8}$. Though the maximum iteration number is large, the algorithm usually converge within 10 iterations. For each group of parameters, we conduct the simulation 20 times. In all 20 simulations, the underlying parameters are the same, but the labels and adjacency matrix are generated randomly.

From the simulations, we compare two methods: the traditional spectral method and the VB algorithm initialized from spectral method. The label obtained traditional spectral method is simply $Z^{[0]}$, the estimator $\widehat{\theta}$ from the spectral method is calculated by $\widehat{\theta}_{ij} = \widehat{B}_{z_i z_j}$, where $\widehat{B}_{cd} = \frac{1}{n_c n_d} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{I}\{z_i = c, z_j = d\}$. The label of VB algorithm is $\widehat{Z}$ and the corresponding estimator $\widehat{\theta}$ is given by $\widehat{\theta}_{ij} = \widehat{\mu}_{z_i z_j}$.

Then consider the following two criteria to compare different methods:

- Hamming Distance: $h(\widehat{z}, z) = \min_{\pi \in S_k} \sum_{i=1}^{n} \mathbb{I}\{\widehat{z}_i \neq \pi(z_i)\}$, where S_k is the set of all permutations for k numbers and π refers to some permutation.

- ℓ_2 Distance: $\ell(\widehat{\theta}, \theta) = \|\theta - \widehat{\theta}\|_F^2$, where $\theta = \mathbb{E}(A)$.

The comparison results are given in Figures 3.1 and 3.2 respectively.

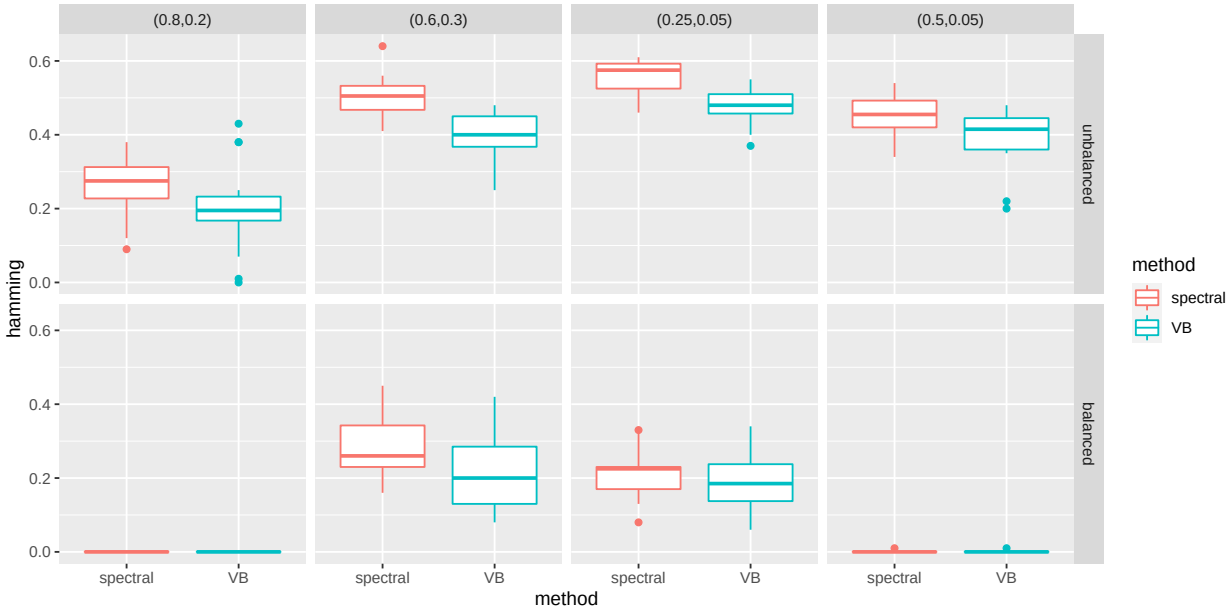

Figure 3.1: Compare Hamming Distance When k^* is Known

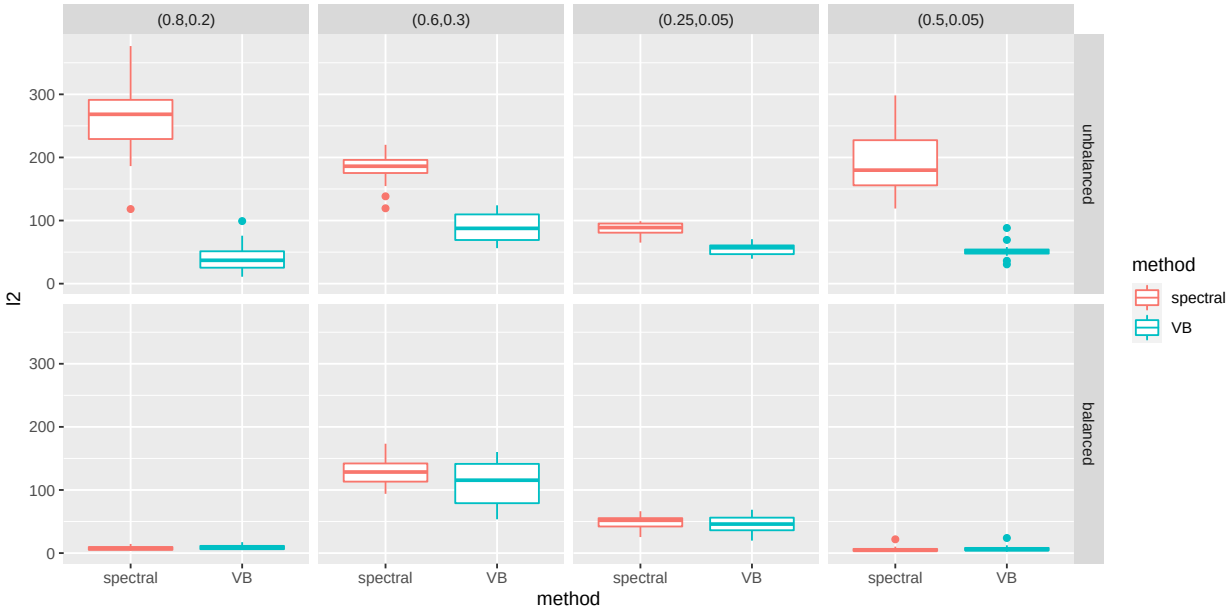

Figure 3.2: Compare ℓ_2 Distance When k^* is Known

From simulation results, the VB algorithm improves the clustering result and decreases the ℓ_2 loss from the output of the spectral method, especially when the network is unbalanced. The potential reason is that the design matrix $Z \otimes Z$ under unbalanced network has a large condition number $\frac{n_{\max}}{n_{\min}}$. Traditional approaches may fail when the condition number of the underlying design matrix is large. But according to our theorem, large condition number of the design matrix does not influence the convergence rate of the variational posterior distribution. Moreover, in practice, though the algorithm is not convex, it can take care of the singular structure automatically by having the term $-\log\det(\mathcal{X}_Z^T\mathcal{X}_Z)$ in the objective function.

Experiment 2

In this experiment, we adopt the same setting as above. Instead of assuming that we know the true number of communities, we run the algorithm with k from 2 to 10 and select $\widehat{k}$ by Algorithm 3. After 20 experiments for each parameter setting, the histograms of selected $\widehat{k}$ are given in Figure 3.3.

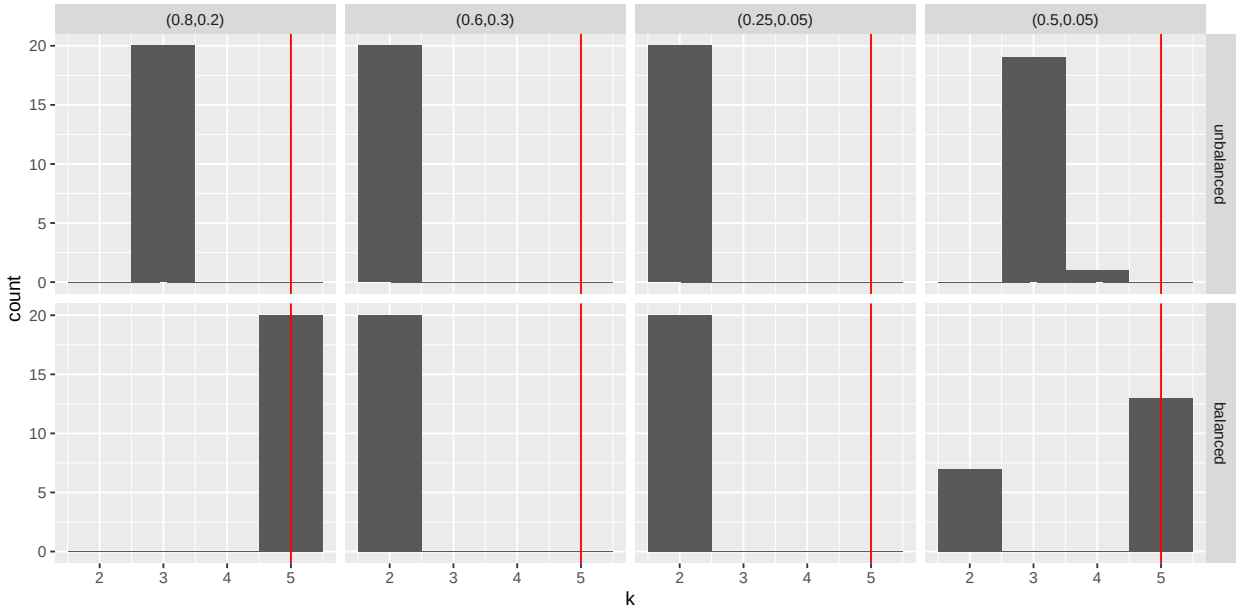

Figure 3.3: Histogram of $\widehat{k}$

To understand the result, we first introduce the signal-to-noise ratio for this special setting. According to [72], the signal-to-noise ratio for SBM in our setting can be defined as $I = \frac{n(p-q)^2}{p}$. Then the signal-to-noise ratios for 4 pairs of (p, q) are given by $I(0.8, 0.2) = 45$, $I(0.6, 0.3) = 15$, $I(0.25, 0.05) = 16$, $I(0.5, 0.05) = 40.5$. Even under balanced network cases, only when the signal-to-noise ratio is large, the true number of community number can be recovered. When the signal-to-noise ratio is small, the smallest k is usually selected. Under the unbalanced network cases, as three communities in the network are too small, they are combined as one community. Thus the equivalent number of community, in this case, is actually 3. That is why under a large signal-to-noise ratio, $\widehat{k} = 3$ is selected.

Now we compare the result from the VB algorithm with the automatically chosen $\widehat{k}$ and the result from spectral method assuming that the true k^* is known. However, as $\widehat{k}$ may not be the same as k^*, instead of using hamming distance, we adopt adjusted rand index(ARI) as the criterion. This criterion is proposed in [36] and is defined as follows:

Definition 3.5.1 (Adjusted Rand Index)**.** *Assume* $z_1 \in [k_1]^n$ *and* $z_2 \in [k_2]^n$, $n_{ij} = \sum_{t=1}^n \mathbb{I}\{z_{1t} = i, z_{2t} = j\}$, $a_i = \sum_{t=1}^n \mathbb{I}\{z_{1t} = i\}$, $b_j = \sum_{t=1}^n \mathbb{I}\{z_{2t} = j\}$, *then the adjusted rand index between* z_1 *and* z_2 *is defined as*

$$\mathrm{ARI}(z_1, z_2) = \frac{\sum_{i=1}^{k_1}\sum_{j=1}^{k_2}\binom{n_{ij}}{2} - \sum_{i=1}^{k_1}\binom{a_i}{2}\sum_{j=1}^{k_2}\binom{b_j}{2}/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i=1}^{k_1}\binom{a_i}{2} + \sum_{j=1}^{k_2}\binom{b_j}{2}\right] - \sum_{i=1}^{k_1}\binom{a_i}{2}\sum_{j=1}^{k_2}\binom{b_j}{2}/\binom{n}{2}}.$$

The range of ARI is usually from 0 to 1. The higher the ARI, the better the clustering result. Now the ARIs for results from VB algorithm and spectral method are shown in Figure 3.4.

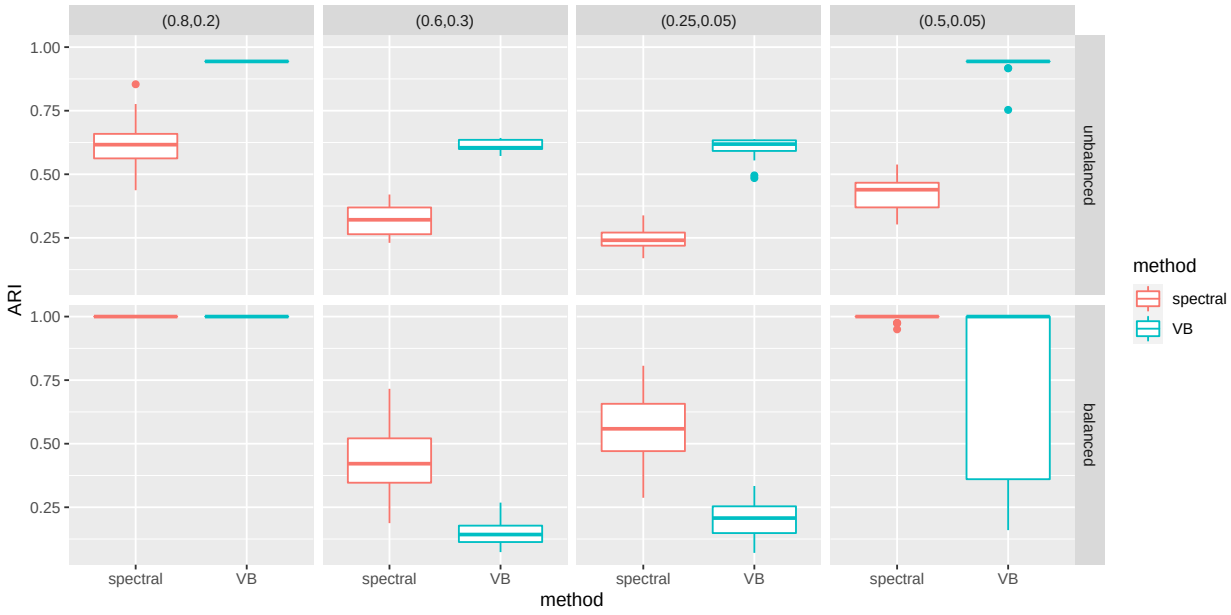

Figure 3.4: Compare ARI

In unbalanced cases, the VB algorithm outperforms the spectral method in all parameter settings, even though we assume that the number of clusters is known when we apply the spectral method. However, in balanced cases, only when SNR is large, the clustering performance of the VB algorithm can be good. The reason is that when SNR is small, the selected number of communities $\widehat{k}$ is not accurate. Therefore, it cannot be as good as the spectral method, which assumes the true number of communities is known.

Moreover, we can also compare ℓ_2 distances for these two methods in Figure 3.5. The results are likewise.

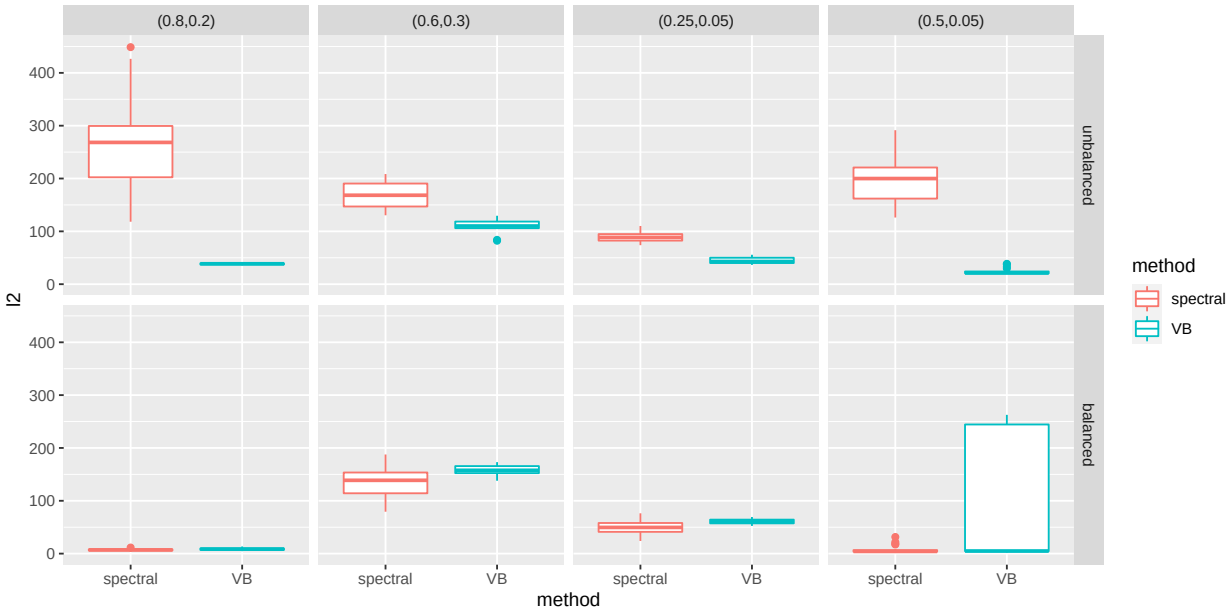

Figure 3.5: Compare ℓ_2 distance

3.5.2 Sparse Linear Regression

We conduct similar experiments for the sparse linear regression model. Assume the number of observations is $n = 100$, the number of predictors is $p = 200$ and the true sparsity is $s^* = 20$. We apply the results from the LASSO method as the initialization of our algorithm. According to [21], LASSO has a piecewise linear solution path in which the sparsity increases at most one at each change point. Then we can choose the support of the first change point in the LASSO solution path, at which the sparsity of the coefficient is s, as the initialization for the sparsity s in our VB algorithm.

In the first experiment, we assume $s^* = 20$ is known and apply Algorithm 6 to get the coefficient estimation. Then in the second experiment, we assume s^* is not known and use sparse linear regression version of Algorithm 3 to select an $\widehat{s}$.

Experiment1

In this experiment, we assume the design matrix is expressed as follows:

$$X = \alpha U_r^T V_r/\sqrt{r} + (1-\alpha)W,$$

where $U_r \in \mathbb{R}^{r\times n}$, $V_r \in \mathbb{R}^{r\times p}$, $W \in \mathbb{R}^{n\times p}$ are random Gaussian matrix with $r = 5$ and $\alpha \in [0,1]$ denotes the proportion of collinearity part. The true parameter is given by $\beta = (\beta_S, \beta_{S^c})$, where S is randomly chosen from $[p]$ with $|S| = 20$ and

$$\beta_S = \mathrm{SNR}\sqrt{\frac{\log p}{n}} \times (1, -1, \cdots, 1, -1).$$

In the simulation, we choose the repeat number $R = 5$, the maximum iteration number $M = 100$ and the tolerate threshold $\epsilon = 1 \times 10^{-8}$. To construct the design matrix, we choose $\alpha = (0, 0.6)$, which denote the independent random design and collinear random design, respectively. We choose the signal-to-noise ratio SNR from $(0.5, 1, 2, 4, 8, 16)$ so that we can see the performance of the VB algorithm when SNR is from sufficiently small to sufficiently large.

We compare the results from LASSO and the VB algorithm by the following two criteria:

- FDR: $h(\widehat{S}, S^*) = \frac{|\widehat{S}\cap S^*|}{|\widehat{S}|\vee 1}$, where $\widehat{S}$ is the estimated support and S^* is the model true support.
- ℓ_2 distance: $\ell(\widehat{\theta}, \theta) = \|\theta - \widehat{\theta}\|^2$, where $\widehat{\theta} = X\widehat{\beta}$ is the estimated signal and $\theta = X\beta^*$ is the true signal.

For each group of parameters, we conduct the simulation 20 times. The results are given in Figure 3.6 and 3.7 respectively.

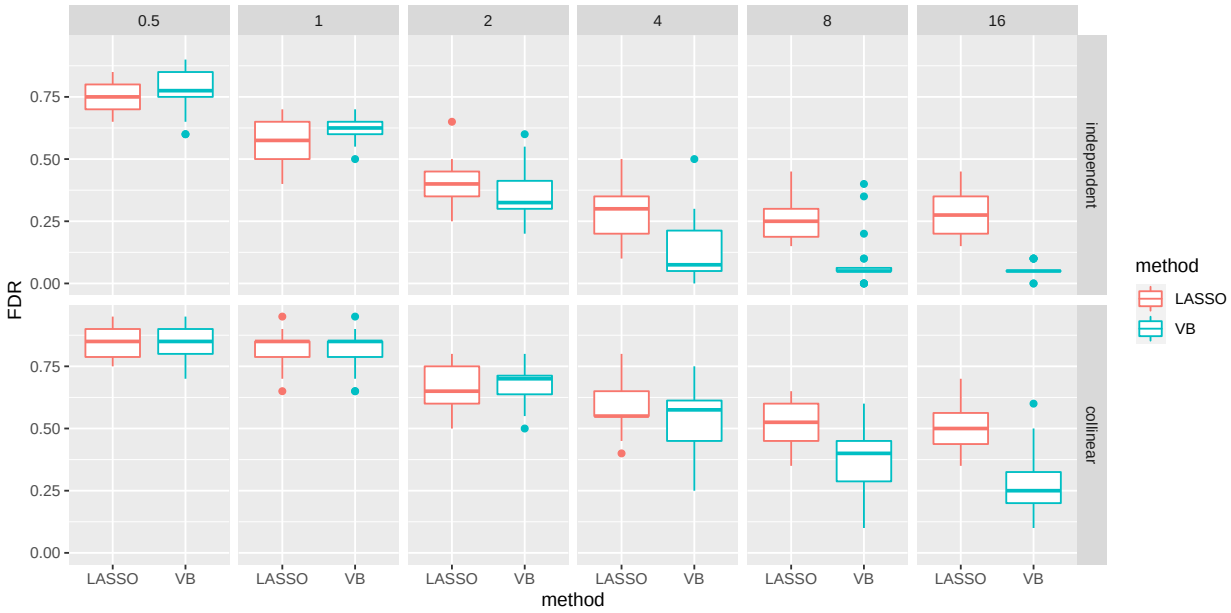

Figure 3.6: Compare FDR when s^* is known

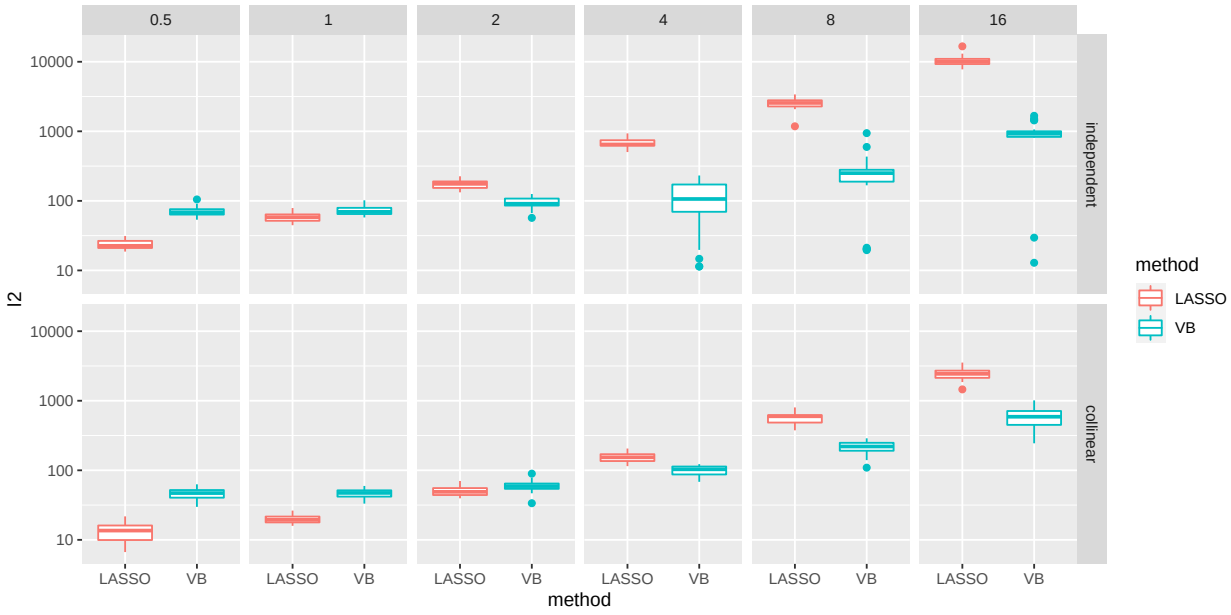

Figure 3.7: Compare ℓ_2 distance when s^* is known

The plots show that our method improves the LASSO result when SNR is large. A potential reason is that LASSO estimator has a bias when the true coefficients are large.

Moreover, [75] pointed out that the Strong Irrepresentable Condition, which is an important condition for the sparsity consistency for LASSO, only holds when $s^* < \sqrt{n}$. However, this condition is not satisfied in our settings, so the sparse consistency for LASSO does not hold. However, the concentration of variational posteriors does not have such constraint. Thus, when SNR is getting larger, both FDR and ℓ_2 loss decrease faster VB results than LASSO results under both the independent random design and the collinear random design.

Experiment2

In this experiment, we adopt the same settings as above. The true sparsity s^* is still 20. However, we run our VB algorithm with s from 1 to 40 to select the best $\widehat{s}$ by Algorithm 3. The histogram of $\widehat{s}$ is given in Figure 3.8.

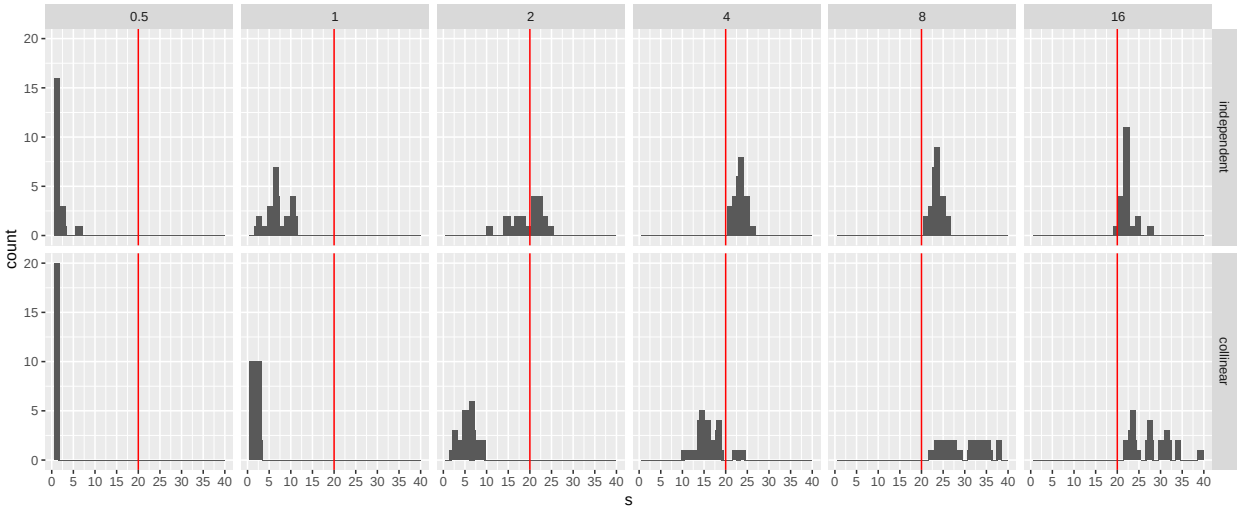

Figure 3.8: Histogram of $\widehat{s}$

These results are similar to those for variational model selection in SBM. When the signal-to-noise ratio is small, the VB algorithm fails to select the correct s^* and $\widehat{s}$ tends to be smaller. However, when the signal-to-noise ratio is large enough, $\widehat{s}$ will converge to s^*. It's worth mentioning that when the design matrix has collinearity, $\widehat{s}$ will first go over s^* and then slowly converge back to s^*. The reason is that when the design matrix has collinearity,

there are many columns highly correlated. In this case, when one column is significant, another column may also be significant. Thus, a larger s may be selected.

We can also compare FDR and ℓ_2 distance for the results from model selection by the variational algorithm and results from LASSO assuming the true sparsity s^* is known. The results are shown in Figure 3.9.

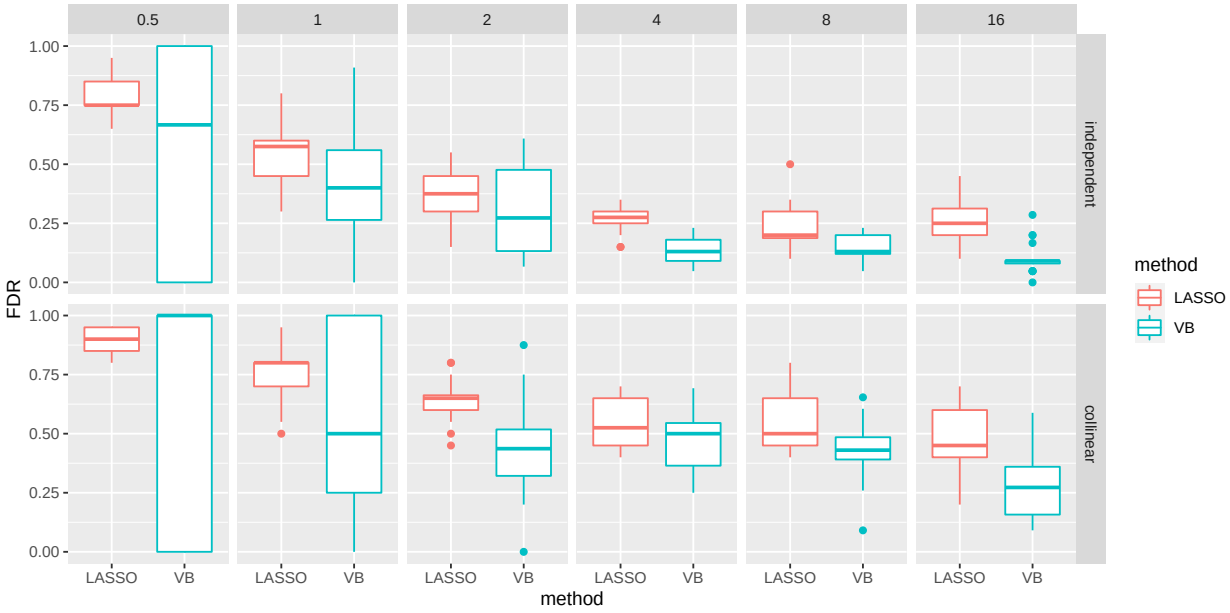

Figure 3.9: Compare FDR

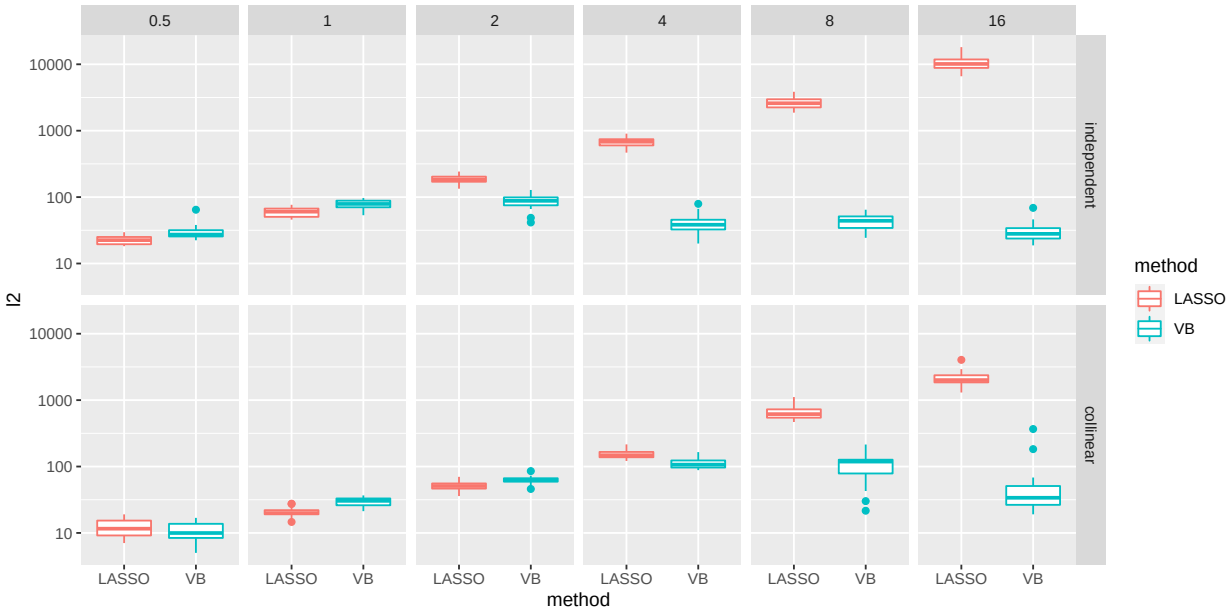

Figure 3.10: Compare ℓ_2 distance

Although when the signal-to-noise ratio is small, the VB method performs worse, the performance of VB improves much faster than LASSO under both FDR and ℓ_2 distance criteria when we increase signal-to-noise ratio. Specifically, when SNR > 2, the VB algorithm outperforms LASSO even though we apply true sparsity s^* in LASSO.

CHAPTER 4

CONVERGENCE RATES OF EMPIRICAL BAYES POSTERIOR DISTRIBUTIONS

4.1 Introduction

In the Bayes analysis, one of the most difficult tasks is to design a good prior, especially for complicated models, where the prior cannot be designed in a fully subjective way. Therefore, statisticians usually consider a prior family $\Pi(\theta|\lambda)$. If the hyper-parameter λ has a prior, this procedure is fully Bayes analysis. For λ chosen as a data-driven $\widehat{\lambda}$, it is known as an empirical Bayes approach.

More generally, the dimension of the parameter space Λ for λ can be either finite (parametric empirical Bayes) or infinite (nonparametric empirical Bayes). A special case is λ referring to the entire prior distribution. In this case, estimating λ is equivalent to estimating the prior itself.

The classical empirical Bayes method is conducted in a hierarchical sequence model, in which we assume the observation X_i comes from a measure P_{θ_i} independently and θ_i's are i.i.d generated from a prior Π_λ. Because the Bayes risk is minimized at the posterior mean, and the posterior mean relies on the prior Π_λ, our main goal is to estimate the prior Π_λ or the key parameter λ of the prior Π_λ from the data. The early work of empirical Bayes analysis can be traced back to [53] for a mixed Poisson model. Similar empirical Bayes analysis was also conducted in [62, 23] for studies of Shakespeare literatures, in [24, 22] for microarrays, in [20] for large scale prediction problems, in [40, 39] for sparse sequence model, in [37] for normal mean model and in [58, 7] for the analysis of FDR.

Recently, the empirical Bayes analysis is also applied for model selection, where different λ's refer to different models. In this case, θ_i's may not have i.i.d distribution. Instead of estimating the entire prior distribution Π, we want to estimate the hyper-parameter λ from the data to select a specific model. This type of works include [6] for smoothness testing, [41]

for inverse problems, [60, 61] for sequence models, and [3, 33] for Bayes model selection. In this section, we consider this type of empirical Bayes analysis. That is, we apply empirical Bayes procedure for model selection.

In practice, there are many ways to select λ based on data. One of the most popular estimators on $\widehat{\lambda}$ is the maximum marginal likelihood estimator (MMLE). MMLE chooses λ by maximizing the marginal likelihood corresponding to λ, say $\int p_\theta(X)d\Pi(\theta|\lambda)$. The MMLE in the sparse sequence model with spike and slab prior is studied in [40, 15, 17]. A generic anlaysis on the asymptotic behaviour of MMLE is conducted in [47] for parametric models and later generalized by [54] for nonparametric models. Besides, [54] also provides a way to determine the minimax rate for the convergence of MMLE posterior distribution. However, as discussed in Chapter 3, the "prior mass and testing" framework usually fails when the true parameter θ^* is assumed unbounded. Thus, in this chapter, we want to extend the result to remove the boundedness condition.

In MMLE, all λ's are treated equally, and the choice of λ only depends on the family of prior $\Pi(\theta|\lambda)$ and the data X. In our story, we add a weight measurement $w(\lambda)$ on different λ and choose $\widehat{\lambda}$ by maximizing the weighted marginal likelihood:

$$\widehat{\lambda} = \operatorname*{argmax}_{\lambda\in\Lambda} w(\lambda) \int p_\theta(X)d\Pi_\lambda(\theta).$$

We assume the data X are generated from a "true" distribution P_{θ^*}. Our goal is to analyze the concentration of the empirical Bayes posterior distribution $\Pi_{\widehat{\lambda}}(\cdot|X)$.

The first half part of this chapter discusses the relation between variational Bayes and empirical Bayes. We extend the examples in Section 2.3.1 and Section 2.3.2 to a more general setting and demonstrate that in the context of sieve priors, the variational Bayes tends to select a data-driven intrinsic dimension $\widetilde{k}$ automatically. As empirical Bayes does the same thing, this discovery motivates us to reveal the relation between VB and EB. We find variational Bayes posterior under a special variational set is exactly the empirical

Bayes posterior. Besides, this result can be extended to the generalized model selection setting in Section 2.4.1. With a slight modification on Theorem 2.4.1, we obtain a "prior mass and testing" type of conditions to verify the convergence of empirical Bayes posterior distributions when the hyper-parameter λ is discrete.

The second half of this chapter aims to develop a general convergence theorem that allows the true parameter θ^* to be unbounded. As the testing condition cannot be satisfied uniformly when the dimension of parameter spaces is infinite, we assume that the entire parameter space can be viewed as the union of small subspaces and assume the testing condition is satisfied for each subspace with an emtropy amplification term uniformly. In this way, all parameters in the parameter space can be testable over the true parameter θ^*. The prior mass condition is replaced by the prior ratio condition so that the true parameter θ^* does not need to be bounded. A summability condition is also proposed to avoid the overfitting problem. Then with the new testing condition, the prior ratio condition, and the summability condition, we provide a theorem to verify the convergence of the empirical Bayes posterior distributions. At the end, we apply this general theorem into the sparse sequence model, sparse linear regression, and the general linear structured model (3.1).

The rest of this chapter is organized as follows. Section 4.2 illustrates the relation between variational Bayes and empirical Bayes when a general type of sieve priors are used. In Section 4.3, we show that empirical Bayes can also be obtained from variational Bayes with model selection when the hyper-parameter space is discrete. Then a "prior mass and testing" framework of convergence theory is derived. In Section 4.4, we propose a general convergence result for the empirical Bayes posterior that allows the hyper-parameter to be continuous and the true parameter to be unbounded. In Section 4.5, we apply our general theory to sparse sequence model, sparse linear regression and the general linear structured model (3.1).

4.2 Variational Bayes and Empirical Bayes

In this section, we discuss an intriguing relation between variational Bayes and empirical Bayes in the context of sieve priors. We consider a nonparametric model $P_\theta^{(n)}$ with an infinite dimensional parameter $\theta = (\theta_j) \in \otimes_{j=1}^{\infty} \Theta_j \subset \mathbb{R}^\infty$. This includes the Gaussian sequence model and the infinite dimensional exponential family discussed in Section 2.3, as well as nonparametric regression and spectral density estimation. For each dimension, we assume $\Theta_j = \Theta_{j1} \cup \Theta_{j2}$ and $\Theta_{j1} \cap \Theta_{j2} = \emptyset$. Then, a sieve prior $\theta \sim \Pi$ is specified by the following sampling process.

1. Sample $k \sim \pi$;

2. Conditioning on k, sample $\theta_j \sim f_{j1}$ for all $j \in [k]$, and sample $\theta_j \sim f_{j2}$ for all $j > k$.

We assume that the densities f_{j1} and f_{j2} satisfy $\int_{\Theta_{j1}} f_{j1} = 1$ and $\int_{\Theta_{j2}} f_{j2} = 1$. A leading example of the sieve prior is case of $\Theta_{j1} = \mathbb{R}\backslash\{0\}$ and $\Theta_{j2} = \{0\}$, as is used in Section 2.3.1 and Section 2.3.2.

An empirical Bayes procedure maximizes $e^{m_k(X^{(n)})}\pi(k)$[1], where

$$m_k(X^{(n)}) = \log \int p(X^{(n)}|\theta) \prod_{j \le k} f_{j1}(\theta_j) \prod_{j>k} f_{j2}(\theta_j) d\theta$$

is the logarithm of marginal likelihood. With the maximizer $\widehat{k}$, the empirical Bayes posterior is defined as

$$d\widehat{Q}_{\mathrm{EB}}(\theta) \propto p(X^{(n)}|\theta) \prod_{j \le \widehat{k}} f_{j1}(\theta_j) \prod_{j>\widehat{k}} f_{j2}(\theta_j) d\theta. \tag{4.1}$$

Compared with a hierarchical Bayes approach, the empirical Bayes procedure does not need to evaluate the posterior distribution of k, and thus in many cases, it is easier to implement.

1. The canonical form of empirical Bayes has a flat prior on k.

We also study the mean-field approximation of the posterior distribution. In order to characterize its form, we need a few definitions. For any $g = (g_j)_{j=1}^{\infty}$, define

$$m_k(X^{(n)}; g) = \int \prod_{j=1}^{\infty} g_j(\theta_j) \log p(X^{(n)}|\theta) d\theta - \sum_{j \le k} D(g_j \| f_{j1}) - \sum_{j>k} D(g_j \| f_{j2}).$$

By Jensen's inequality, we observe that

$$m_k(X^{(n)}) \ge m_k(X^{(n)}, g), \tag{4.2}$$

for any g. We also define the density classes $\mathcal{G}_{j1} = \left\{ g \ge 0 : \int g = \int_{\Theta_{j1}} g = 1 \right\}$ and $\mathcal{G}_{j2} = \left\{ g \ge 0 : \int g = \int_{\Theta_{j2}} g = 1 \right\}$. The next theorem gives the exact form of the mean-field variational posterior.

Theorem 4.2.1. *Consider the variational posterior $\widehat{Q}_{\mathrm{VB}}$ induced by the sieve prior and the mean-field variational set $\mathcal{S}_{\mathrm{MF}}$. The distribution $\widehat{Q}_{\mathrm{VB}}$ is a product measure with the density of each coordinate specified by*

$$q_j = \begin{cases} \widetilde{g}_{j1}^{(\widetilde{k})}, & j < \widetilde{k}, \\ (1-\widetilde{p})\widetilde{g}_{j1}^{(\widetilde{k})} + \widetilde{p}\widetilde{g}_{j2}^{(\widetilde{k})}, & j = \widetilde{k}, \\ \widetilde{g}_{j2}^{(\widetilde{k})}, & j > \widetilde{k}, \end{cases}$$

where for each given k, $(\widetilde{g}_{j1}^{(k)})_{j=1}^{k}$ and $(\widetilde{g}_{j2}^{(k)})_{j=k}^{\infty}$ maximize the following objective function,

$$\pi(k-1) e^{m_{k-1}\left(X^{(n)}, (g_{j1})_{j=1}^{k-1} \cup (g_{j2})_{j=k}^{\infty}\right)} + \pi(k) e^{m_k\left(X^{(n)}, (g_{j1})_{j=1}^{k} \cup (g_{j2})_{j=k+1}^{\infty}\right)}, \tag{4.3}$$

under the constraints that $g_{j1} \in \mathcal{G}_{j1}$ and $g_{j2} \in \mathcal{G}_{j2}$ for all j, $\widetilde{k}$ maximizes

$$\pi(k-1) e^{m_{k-1}\left(X^{(n)}, (\widetilde{g}_{j1}^{(k)})_{j=1}^{k-1} \cup (\widetilde{g}_{j2}^{(k)})_{j=k}^{\infty}\right)} + \pi(k) e^{m_k\left(X^{(n)}, (\widetilde{g}_{j1}^{(k)})_{j=1}^{k} \cup (\widetilde{g}_{j2}^{(k)})_{j=k+1}^{\infty}\right)}, \tag{4.4}$$

and finally,

$$\widetilde{p} = \frac{\pi(\widetilde{k}-1)e^{m_{\widetilde{k}-1}\left(X^{(n)},(\widetilde{g}_{j1}^{(\widetilde{k})})_{j=1}^{\widetilde{k}-1}\cup(\widetilde{g}_{j2}^{(\widetilde{k})})_{j=\widetilde{k}}^{\infty}\right)}}{\pi(\widetilde{k}-1)e^{m_{\widetilde{k}-1}\left(X^{(n)},(\widetilde{g}_{j1}^{(\widetilde{k})})_{j=1}^{\widetilde{k}-1}\cup(\widetilde{g}_{j2}^{(\widetilde{k})})_{j=\widetilde{k}}^{\infty}\right)} + \pi(\widetilde{k})e^{m_{\widetilde{k}}\left(X^{(n)},(\widetilde{g}_{j1}^{(\widetilde{k})})_{j=1}^{\widetilde{k}}\cup(\widetilde{g}_{j2}^{(\widetilde{k})})_{j=\widetilde{k}+1}^{\infty}\right)}}.$$

The result of Theorem 4.2.1 also applies to the class $\mathcal{S}_{\mathrm{G}}$ discussed in Section 2.3.2 with $\mathcal{G}_{j1}$ replaced by the Gaussian class. We note that Theorem 4.2.1 can be viewed as an extension of Theorem 2.3.1. In fact, if the likelihood function can be factorized over each coordinate of θ, the form of $\widehat{Q}_{\mathrm{VB}}$ can be greatly simplified.

Corollary 4.2.1. *Under the same setting of Theorem 4.2.1, if we further assume that* $p(X^{(n)}|\theta) = \prod_{j=1}^{\infty} p(X_j^{(n)}|\theta_j)$*, then we will have*

$$\widetilde{g}_{j1}^{(\widetilde{k})}(\theta_j) \propto f_{j1}(\theta_j)p(X_j^{(n)}|\theta_j)\mathbf{1}_{\{\theta_j\in\Theta_{j1}\}},$$

$$\widetilde{g}_{j2}^{(\widetilde{k})}(\theta_j) \propto f_{j2}(\theta_j)p(X_j^{(n)}|\theta_j)\mathbf{1}_{\{\theta_j\in\Theta_{j2}\}},$$

$$\widetilde{k} = \operatorname*{argmax}_k \left(\pi(k-1|X^{(n)}) + \pi(k|X^{(n)})\right), \tag{4.5}$$

and

$$\widetilde{p} = \frac{\pi(k-1|X^{(n)})}{\pi(k-1|X^{(n)}) + \pi(k|X^{(n)})},$$

where

$$\pi(k|X^{(n)}) \propto \pi(k)\prod_{j=1}^{k}\int_{\Theta_{j1}} f_{j1}(\theta_j)p(X^{(n)})d\theta_j \prod_{j=k+1}^{\infty}\int_{\Theta_{j2}} f_{j2}(\theta_j)p(X^{(n)}|\theta_j)d\theta_j.$$

In light of Theorem 4.2.1, we can compare the variational Bayes approach and the empirical Bayes approach, especially the definitions of $\widetilde{k}$ and $\widehat{k}$. The empirical Bayes chooses the best model by maximizing $e^{m_k(X^{(n)})}\pi(k)$, or equivalently $\pi(k|X^{(n)})$, while the variational Bayes maximizes (4.4). There are two major differences. The first difference is that empiri-

cal Bayes uses the exact marginal likelihood function $m_k(X^{(n)})$ and variational Bayes uses a mean-field approximation of $m_k(X^{(n)})$. We remark that in the case of a likelihood that can be factorized, the mean-field approximation is exact, which leads to (4.5). The second difference is that empirical Bayes maximizes the posterior probability of the kth model, but the variational Bayes maximizes the sum of the posterior probabilities (or their mean-field approximations) of the $(k-1)$th and the kth models.

Despite the two differences, the empirical Bayes approach and the variational Bayes approach have a lot in common. Both are random probability distributions that summarize the information in data and prior. Both select a sub-model according to very similar criteria. To close this section, we show that with a special variational class, the induced variational posterior is exactly the empirical Bayes posterior.

Theorem 4.2.2. *Define the following set*

$$\mathcal{S}_{\mathrm{EB}} = \left\{ Q : Q\left(\left(\otimes_{j \le k} \Theta_{j1} \right) \bigotimes \left(\otimes_{j>k} \Theta_{j2} \right) \right) = 1 \text{ for some integer } k \right\}.$$

Then, the empirical Bayes posterior $\widehat{Q}_{\mathrm{EB}}$ defined by (4.1) is the variational posterior induced by the sieve prior and the variational class $\mathcal{S}_{\mathrm{EB}}$.

The result of Theorem 4.2.2 shows that for sieve priors, one can view the empirical Bayes approach as a variational Bayes approach, which suggests that it may be possible to unify the theoretical analysis in this section and the analysis of empirical Bayes procedures in [54].

4.3 Empirical Bayes for Model Selection

Actually, the relation between empirical Bayes and variational Bayes does not only appear in sieve prior cases. Let's recall the variational Bayes with the model selection procedure proposed in Section 2.4.1. We restate the model in Section 2.4.1 here.

$$\mathcal{M} = \left\{ P^{(n)}_{k,\theta^{(k)}} : k \in \mathcal{K}, \theta^{(k)} \in \Theta^{(k)} \right\},$$

where $\mathcal{K}$ is a countable set and $\theta^{(k)} = \left(\theta_1^{(k)}, \cdots, \theta_{m_k}^{(k)}\right)$.

Instead of assuming the sampling procedure for each k to be a product measure, we assume $\theta^{(k)}$ is sampled from a general probability measure $\Pi^{(k)}$. Then, the prior Π has the following hierarchical sampling procedure:

1. Sample $k \sim \pi(k)$ from $\mathcal{K}$;

2. Conditioning on k, sample $\theta^{(k)}$ from the probability measure $\Pi^{(k)}$.

This two step hierarchical prior has been studied by [2, 56] for generic models, by [52] for density estimation, by [42] for location-scale mixture model and by [50] for inverse problems.

We consider the same variational inference with model selection procedure introduced in Section 2.4.1. Then for a general variational set $\mathcal{S}$, the problem is converted to solve the optimization problem:

$$\max_{k\in\mathcal{K}} \max_{Q^{(k)}} \left\{ \int \log p(X^{(n)}|\theta^{(k)}) dQ^{(k)}(\theta^{(k)}) - D\left(Q^{(k)} \| \Pi^{(k)}\right) + \log \pi(k) \right\},$$

Now we assume $\mathcal{S}$ to be the set of all distributions rather than the mean field class $\mathcal{S}_{\mathrm{MF}}^{(k)}$. Then for a fixed $k \in \mathcal{K}$, the solution of $Q^{(k)}$ is given by $\Pi^{(k)}(\theta|X^{(n)})$. We plug this solution to (2.34), then the problem becomes

$$\max_{k\in\mathcal{K}} F\left(\Pi^{(k)}(\theta|X^{(n)}), k\right),$$

which is equivalent to

$$\max_{k\in\mathcal{K}} \int \pi(k) p(X^{(n)}|\theta^{(k)}) d\Pi^{(k)}(\theta^{(k)}). \tag{4.6}$$

Be aware that we do not assume the parameter spaces for different k's are disjointed. Therefore, if we set the variational class for each index k is the set of all distributions, then variational Bayes procedure with model selection is equivalent to empirical Bayes for a generic model when the hyper-parameter is discrete. Suppose $\widehat{k}$ is the solution to (4.6) and

$\widehat{\Pi}^{(\widehat{k})}$ is the empirical Bayes posterior distribution. Likewise, we can also propose a special variational class and show that with this special variational class, the induced variational posterior distribution is exactly the empirical Bayes posterior distribution.

Theorem 4.3.1. *Define the following set*

$$\mathcal{S}_{\mathrm{EB}} = \left\{ Q : Q\left(\left\{ k = k_0, \theta \in \Theta^{(k_0)} \right\} \right) = 1 \text{ for some } k_0 \in \mathcal{K} \right\},$$

Then, the empirical Bayes posterior $\widehat{\Pi}^{(\widehat{k})}$ *is the variational posterior induced by the hierarchical prior* Π *and the variational class* $\mathcal{S}_{\mathrm{EB}}$.

This theorem motivates us to develop a general convergence result for empirical Bayes posterior distribution, when the hyper-parameter is discrete. It is illustrated in the following theorem.

Theorem 4.3.2. *Suppose* ϵ_n *is a sequence that satisfies* $n\epsilon_n^2 \geq 1$. *Consider a loss function* $L(\cdot, \cdot)$, *such that for any two probability measures* P_1 *and* P_2, $L(P_1, P_2) \geq 0$. *Let* $C, C_1, C_2, C_3 > 0$ *be constants such that* $C > C_2 + C_3 + 2$. *We assume*

- *For any* $\epsilon > \epsilon_n$, *there exists a set* $\Theta_n(\epsilon)$ *and a testing function* ϕ_n, *such that*

$$P_0^{(n)}\phi_n + \sup_{\substack{\theta \in \Theta_n(\epsilon) \\ L(P_\theta^{(n)}, P_0^{(n)}) \geq C_1 n\epsilon^2}} P_\theta^{(n)}(1 - \phi_n) \leq \exp(-Cn\epsilon^2). \tag{4.7}$$

- *For any* $\epsilon > \epsilon_n$, *the set* $\Theta_n(\epsilon)$ *above satisfies*

$$\Pi(\Theta_n(\epsilon)^c) \leq \exp(-Cn\epsilon^2). \tag{4.8}$$

- *For some constant* $\rho > 1$, *there exists a* $k_0 \in \mathcal{K}$ *such that*

$$-\log \pi(k_0) - \log \Pi^{(k_0)}\left(\theta : D_\rho\left(P_0^{(n)} \| P_\theta^{(n)} \right) \leq C_3 n\epsilon_n^2 \right) \leq C_2 n\epsilon_n^2. \tag{4.9}$$

Then,

$$P_0^{(n)}\Pi^{(\widehat{k})}\left[L(P_\theta^{(n)}, P_0^{(n)})|X^{(n)}\right] \lesssim n\epsilon_n^2. \tag{4.10}$$

The proof of Theorem 4.3.2 largely follows the proof of Theorem 2.4.1, so we omit it in this thesis. Similar results are also proposed in [33] with the testing condition replaced by the entropy condition.

4.4 Convergence Rates for Empirical Bayes Posterior Distributions

In the previous section, we have derived a general convergence rate for empirical Bayes posterior distribution when the hyper-parameter space $\mathcal{K}$ is a discrete set. However, the empirical Bayes procedure is not limited to it. For example, in sparse linear regression, $Y = X_S\beta_S + \epsilon$. We can put an independent spike and slab prior $\beta_i \sim \lambda g + (1-\lambda)\delta_0$ with some density function g, then $\lambda \in (0,1)$ is the hyper-parameter in this case and the possible choice of λ is uncountable. For a generic λ, the general convergence rates for the empirical Bayes posterior distributions have been studied by [47] for parametric models and by [54] for nonparametric models. However, both of them require the prior mass condition. As we have discussed in Chapter 3, in order to satisfy the prior mass condition, the true parameter θ^* must be bounded, but this assumption is not necessary in practice. In sparse sequence model, this boundedness assumption can be removed by applying a heavy tail prior on the parameter [15]. However, in general settings, a new convergence result is necessary. In this section, our main goal is to derive a general convergence result that can not only be applied for the continuous hyper-parameter space, but allow the true parameter to be unbounded as well.

Assume the observation X is sampled from probability measure P_θ indexed by $\theta \in \Theta$. The density of P_θ is denoted as p_θ. Θ can be an infinite-dimensional space so that the non-parametric setting is also included in our theory. The prior for θ is Π, which is a mixture

probability measure given as

$$\Pi(\theta) = \sum_{\lambda \in \Lambda} w(\lambda) \Pi_\lambda(\theta), \tag{4.11}$$

when Λ is a discrete set or

$$\Pi(\theta) = \int_{\lambda \in \Lambda} w(\lambda) \Pi_\lambda(\theta) d\lambda, \tag{4.12}$$

when Λ is a continuous set.

Then the empirical Bayes procedure choose λ by

$$\widehat{\lambda} = \underset{\lambda \in \Lambda}{\operatorname{argmax}}\, w(\lambda) \int p_\theta(X) d\Pi_\lambda(\theta). \tag{4.13}$$

This procedure is different from the classical empirical Bayes procedure in that there is a weight probability measure $w(\lambda)$ on the hyper-parameter λ. Our goal is to analyze the posterior distribution from the empirical Bayes procedure:

$$\widehat{\Pi}(B) = \frac{\int_B p_\theta(X) d\Pi_{\widehat{\lambda}}(\theta)}{\int p_\theta(X) d\Pi_{\widehat{\lambda}}(\theta)}. \tag{4.14}$$

To further analyze the empirical Bayes posterior, we assume that the parameter space Θ can be decomposed as $\Theta = \cup_{Z \in \mathcal{Z}} \Theta_Z$, where Θ_Z is a linear subspace indexed by Z. Heuristically, we should expect that the testing condition over different Θ_Z's are different so that it can be satisfied in most cases. Thus, we write Π_λ as

$$\Pi_\lambda = \sum_{Z \in \mathcal{Z}} \nu_\lambda(Z) \Gamma_Z,$$

where Γ_Z is a probability measure supported on Θ_Z. Then we will have the following convergence result for the EB posterior distribution:

Theorem 4.4.1. *We denote $\gamma(Z) = \max_{\lambda \in \Lambda} \{w(\lambda)\nu_\lambda(Z)\}$. Suppose $L(\cdot, \cdot)$ to be a loss function such that $L(\theta_1, \theta_2) > 0$ for any $\theta_1, \theta_2 \in \Theta$, Assume $\theta^* \in \Theta_{Z^*}$ for $Z^* \in \mathcal{Z}$. Let $c, C, C_1, C_2, C_3, C_4, C_5, C_6 > 0$ be constants such that $C_3 > 3c$. We assume the following*

conditions are satisfied:

1. *(Testing). There exists a constant $M_0 > 0$, a testing function ϕ and a sequence $\{\epsilon(Z)\}_{Z\in\mathcal{Z}}$ satisfying $\epsilon(Z) \gtrsim 1$ such that*

$$P_{\theta^*}\phi \le \exp(-C\epsilon(Z^*)^2), \quad \sup_{\theta\in\Theta_Z: L(\theta,\theta^*)\ge\epsilon^2} P_\theta(1-\phi) \le \exp\left(C_2(\epsilon(Z)^2 + \epsilon(Z^*)^2) - C_1\epsilon^2\right), \tag{4.15}$$

   for any $\epsilon^2 > M_0\epsilon(Z^*)^2$ and $Z \in \mathcal{Z}$.

2. *(Prior Ratio). For any $Z \in \mathcal{Z}$ and $\epsilon^2 > M_0\epsilon(Z^*)^2$, denote*

$$R_Z(\epsilon) = \left\{\theta \in \Theta_Z : L(\theta,\theta^*) < \epsilon^2\right\} \tag{4.16}$$

   and

$$K = \left\{\theta \in \Theta_{Z^*} : D_{1+\rho}\left(P_{\theta^*}\|P_\theta\right) \le C_3\epsilon(Z^*)^2\right\}. \tag{4.17}$$

   for some $\rho > 0$. We assume that $\Gamma_{Z^*}(K) > 0$ and there exists a positive sequence $\{\delta(Z)\}_{Z\in\mathcal{Z}}$, such that

$$\frac{\Gamma_Z(R_Z(\epsilon))}{\Gamma_{Z^*}(K)} \le \frac{\delta(Z)}{\delta(Z^*)}\exp\left(c\epsilon^2 + C_4\epsilon(Z)^2 + C_5\epsilon(Z^*)^2\right). \tag{4.18}$$

3. *(Summability). For $\epsilon(Z)$, $\delta(Z)$ in (4.15) and (4.18), there exists a $\lambda^* \in \Lambda$ and a constant $M > 0$, such that*

$$\sum_{Z\in\mathcal{Z}} \frac{\gamma(Z)\delta(Z)}{w(\lambda^*)\nu_{\lambda^*}(Z^*)\delta(Z^*)}\exp\left((C_2+C_4)\epsilon(Z)^2\right) \lesssim \exp\left(C_6\epsilon(Z^*)^2\right), \tag{4.19}$$

Then there exist constants M_1, M_2 such that

$$P_{\theta^*}\widehat{\Pi}\left(L(\theta,\theta^*) > M_1\epsilon(Z^*)^2\right) \le 3\exp\left(-M_2\epsilon(Z^*)^2\right).$$

The testing condition (4.15) and summability condition (4.19) correspond to the testing condition (4.7) and (4.8) in Theorem 4.3.2. However, the prior mass condition (4.9) in Theorem 4.3.2 is replaced by a prior ratio condition (4.18) in Theorem 4.4.1 so that it can be satisfied when the true parameter θ^* is unbounded.

In Theorem 4.4.1, the sequence $\{\delta(Z)\}_{Z\in\mathcal{Z}}$ aims to calibrate the factor caused by the heterogeneity of dimensions of different Θ_Z. For example, in the general linear structured model (4.24), $\frac{\Gamma(\ell_\tau/2)}{\Gamma(\ell_\tau)}$ in the prior density function of B needed to be corrected by the prior of τ. Then we use $\delta(Z)$ to characterize this term while applying Theorem 4.4.1. More details are given in Section 4.5.3.

It's worth mentioning that in the proof of Theorem 4.4.1, we don't use the condition that $w(\lambda)$ is a probability measure for λ. Thus, we can set $w(\lambda) = 1$ for all $\lambda \in \Lambda$, then the empirical Bayes is reduced to ordinary empirical Bayes procedure without weight, and the $\widehat{\lambda}$ obtained from this procedure is exactly the maximum marginal likelihood estimator (MMLE) analyzed in [54].

4.5 Applications

In this section, we apply Theorem 4.4.1 to derive the convergence rates of empirical Bayes posterior distributions for the sparse sequence model, sparse linear regression, and the general linear structured model analyzed in Chapter 3.

4.5.1 Sparse Sequence Model

Consider the empirical Bayes procedure for the sparse sequence model. In the sparse sequence model, $p_\theta(X) = N(\theta, I_p)$. The true parameter θ^* is in a sparse parameter space with support S^*, i.e. $\Theta_{S^*} = \{\theta : \theta_i = 0, i \notin S^*\}$.

The prior $\theta \sim \Pi_\lambda$ is defined as

$$\theta_j \overset{i.i.d}{\sim} (1-\lambda)\delta_0 + \lambda g, \tag{4.20}$$

where $\lambda \in [0,1]$. The empirical Bayes procedure for spike and slab prior is studied by [40] for the generic sparse sequence model and by [39] for the wavelet shrinkage method. The minimax results for the generic sparse sequence model is established in [15]. In [15], they prove that the empirical Bayes posterior distribution by using Laplace slab prior can only converge in a suboptimal rate. However, in their setting, they use the classical MMLE estimator for $\widehat{\lambda}$, where $w(\lambda)$ is assumed to be 1. In our setting, we point out that as long as $w(\lambda)$ satisfies some conditions, the empirical Bayes posterior distribution with Laplace slab prior can also converge at a optimal rate.

We put a beta prior on the hyperparameter index λ, i.e. $\lambda \sim \text{Beta}(\alpha, \beta)$, then

$$w(\lambda) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \lambda^{\alpha-1}(1-\lambda)^{\beta-1}. \tag{4.21}$$

Obviously, the parameter space consists of subspaces with different dimensions. The structure Z in the sparse sequence model refers to the sparsity pattern $S \subset [p]$ and Γ_S refers to a product measure on $\left\{\theta_j\right\}_{j\in S}$ with each coordinate independently following a distribution g. We choose g to be a Laplace distribution with the density:

$$g(\theta) = \frac{\rho}{2} e^{-\rho|\theta|}. \tag{4.22}$$

For this model, the posterior distribution for a specific $\lambda \in [0,1]$ is given by

$$\Pi_\lambda\left(B|X\right) = \frac{\int_B p_\theta(X) d\Pi_\lambda(\theta)}{\int p_\theta(X) d\Pi_\lambda(\theta)}.$$

With the empirical Bayes procedure, λ is selected by

$$\widehat{\lambda} = \underset{\lambda\in(0,1)}{\text{argmax}}\, w(\lambda) \int p_\theta(X) d\Pi_\lambda(\theta).$$

The empirical Bayes posterior distribution $\widehat{\Pi}$ is given by $\widehat{\Pi} = \Pi_{\widehat{\lambda}}(\cdot|X)$. The following theorem

illustrate that with some assumptions on the weight $w(\lambda)$, we can show the convergence of EB posterior distribution:

Theorem 4.5.1. *For the sparse sequence model $p_\theta(X) = N(\theta, I_p)$, assume $\theta^* \in \Theta_{S^*}$ with $|S^*| = s^* < p^a$ with some $a < 1$. Assume the prior in (4.12) is specifically defined by (4.20), (4.21) and (4.22) and the corresponding EB posterior distribution is $\widehat{\Pi}$, then if there exist two constants $\nu_1 > \nu_2 > D$ for a sufficient large constant $D > 0$ such that*

$$p^{-\nu_1}(\beta + p) \leq \alpha \leq p^{-\nu_2}\beta - p, \tag{4.23}$$

Then there exists two constants $M_1, M_2 > 0$, such that

$$P_{\theta^*}\widehat{\Pi}\left(\|\theta - \theta^*\|^2 > M_1 s^* \log p\right) \leq 3\exp\left(-M_2 s^* \log p\right).$$

The rate in Theorem 4.5.1 is minimax [65] and the boundedness condition for the true parameter θ^* is not required. Similar results are also proposed in [18] for the convergence rate of the posterior distribution in sparse sequence model.

4.5.2 Sparse Linear Regression

Sparse linear regression model with spike and slab prior is well studied recently. The convergence results true posterior distribution are given in [16] and the corresponding results for variational posterior distribution are proposed in [51]. In this section, we consider the convergence rate for empirical Bayes posterior distribution in sparse linear regression model with spike and slab prior.

Consider the likelihood $p_\theta(Y) = N(X\theta, I_n)$ for some fixed design matrix $X \in \mathbb{R}^{n\times p}$. We apply the same prior as in the last section. The prior sampling procedure is given as follows:

1. Sample λ from Beta(α, β) with the density:

$$w(\lambda) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\lambda^{\alpha-1}(1-\lambda)^{\beta-1},$$

2. Conditioning on λ, sample $\theta_i \overset{i.i.d}{\sim} (1-\lambda)\delta_0 + \lambda g$, with

$$g(\theta) = \frac{\rho}{2}\exp\left(-\rho|\theta|\right).$$

The empirical Bayes posterior distribution is also defined as $\widehat{\Pi} = \Pi_{\widehat{\lambda}}(\cdot|X)$ with $\widehat{\lambda}$ defined as

$$\widehat{\lambda} = \underset{\lambda\in(0,1)}{\operatorname{argmax}}\, w(\lambda) \int p_\theta(X) d\Pi_\lambda(\theta).$$

Then we have the following concentration result for its EB posterior distribution with the same condition (4.23) in the following theorem.

Theorem 4.5.2. *For the sparse linear regression model $p_\theta(Y) = N(X\theta, I_n)$, assume $\theta^* \in \Theta_{S^*}$ with $|S^*| = s^* < p^a$ for some $a < 1$. Suppose κ is the restricted eigenvalue defined by*

$$\kappa = \inf\left\{\frac{\|Xu\|\sqrt{s^*}}{\|u\|_1} : \|u_{S^{*c}}\|_1 \le 3\|u_{S^*}\|_1, u \neq 0\right\},$$

If $\max_{1\le j\le p}\|X_{\cdot,j}\| \le p^C$ for some constant $C > 0$ and $s^* < p^a$ for some $a < 1$, then with the same spike and Laplace slab prior and the condition (4.23), we have

$$P_{\theta^*}\widehat{\Pi}\left(\|X\theta - X\theta^*\|^2 > M_1\frac{s^*\log p}{\kappa^2\wedge 1}\right) \le 3\exp\left(-M_2\frac{s^*\log p}{\kappa^2\wedge 1}\right).$$

We mention that the condition $\max_{1\le j\le p}\|X_{\cdot,j}\| \le p^C$ is natural and also assumed in [16, 51]. This convergence rate in Theorem 4.5.2 is actually the same as the convergence rate of LASSO estimator [9].

4.5.3 General Linear Structured Model

Finally, we consider the general linear structured model (3.1). Suppose

$$Y = X_Z B + W, \qquad W \sim N(0, I_n), \tag{4.24}$$

where $Z \in \bar{\mathcal{Z}}_\tau$ for some $\tau \in \mathcal{T}$ and $B \in \mathbb{R}^{\ell_\tau}$. In this case, the parameter space of the hyper-parameter τ is discrete and there are natural structured parameter space $\Theta_Z = \left\{\theta = X_Z B : B \in \mathbb{R}^{\ell_\tau}\right\}$ for $Z \in \bar{\mathcal{Z}}_\tau$. With the same prior in [28], we have the hierarchical prior

$$\Pi(\theta) = \sum_{\tau \in \mathcal{T}} w(\tau) \Pi_\tau(\theta),$$

in which

$$w(\tau) \propto \frac{\Gamma(\ell_\tau)}{\Gamma(\ell_\tau/2)} \exp\left(-D\epsilon_\tau\right),$$

and

$$\Pi_\tau(\theta) = \sum_{Z \in \mathcal{Z}_\tau} \nu_\tau(Z) \Gamma_Z,$$

with

$$\nu_\tau(Z) = \frac{1}{|\bar{\mathcal{Z}}_\tau|}$$

and

$$\frac{d\Gamma_Z(B)}{dB} = \frac{\sqrt{\det(X_Z^T X_Z)}}{2} \left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell_\tau} \frac{\Gamma(\ell_\tau/2)}{\Gamma(\ell_\tau)} \exp\left(-\lambda \|X_Z B\|\right),$$

where $\theta = X_Z B$. Then we apply the empirical Bayes procedure to select the hyper index τ:

$$\hat{\tau} = \underset{\tau \in \mathcal{T}}{\operatorname{argmax}}\, w(\tau) \int p_\theta(Y) d\Pi_\tau(\theta),$$

and the corresponding posterior distribution $\widehat{\Pi}$ is given by

$$\widehat{\Pi}(B) = \frac{\int_B p_\theta(Y) d\Pi_{\hat{\tau}}(\theta)}{\int p_\theta(Y) d\Pi_{\hat{\tau}}(\theta)}.$$

We put the same conditions as in [28]:

$$\epsilon_\tau \geq \ell_\tau + \log|\bar{\mathcal{Z}}_\tau|, \qquad |\{\tau \in \mathcal{T} : t-1 < \epsilon_\tau \leq t\}| \leq t \text{ for all } t \in \mathbb{N}. \tag{4.25}$$

Assume data Y is also generated from a linear structured model with Guassian noise: $Y = \theta^* + W$, where $\theta^* = X_{Z^*}B^*$ and $W \sim N(0, I_N)$. By applying Theorem 4.4.1, we have the following result for the convergence rate of empirical Bayes posterior distribution.

Theorem 4.5.3. *For the general linear structured model (4.24), we put the same prior as in [28] for (τ, Z, B). If (4.25) is satisfied and $\Theta_Z \cap \Theta_{Z'} = \emptyset$ for $Z \neq Z'$, then for $Z^* \in \bar{\mathcal{Z}}_{\tau^*}$ then there exists a constant $D_0 > 0$ only depending on λ, such that*

$$P_{X_{Z^*}B^*}\widehat{\Pi}\left(\|X_Z B - X_{Z^*}B^*\|^2 \geq M_1\epsilon_{\tau^*}\right) \leq \exp\left(-M_2\epsilon_{\tau^*}\right),$$

for all $D > D_0$ and some constants $M_1, M_2 > 0$.

The condition $\Theta_Z \cap \Theta_{Z'} = \emptyset$ for $Z \neq Z'$ can usually be satisfied. Because the distribution for B given Z is continuous, it's not harm to remove some degenerated spaces from it. For example, in the sparse linear regression, we can assume $\Theta_S = (\cup_{i\in S}\mathbb{R}_0) \cup \left(\cup_{i\notin S}\{0\}\right)$, where $\mathbb{R}_0 = \mathbb{R}\backslash\{0\}$. In this way, for $S \neq S'$, $\Theta_S \neq \Theta_{S'}$.

In the previous two examples, we set $\delta(Z) = 1$ in Theorem 4.4.1. However, in this example, we set $\delta(Z) = \frac{\Gamma(\ell_\tau/2)}{\Gamma(\ell_\tau)}$ so that this factor can be cancelled by the corresponding factor in $w(\tau)$.

CHAPTER 5

PROOFS

5.1 Proofs in Chapter 2

5.1.1 Proof of Theorem 2.2.1

This section provides the proof of Theorem 2.2.1, which is divided into several lemmas. We first give an inequality that uses the basic property of the KL-divergence.

Lemma 5.1.1. *For any function $f \geq 0$ and two probability measure P and Q, we have*

$$\int f(x)dQ(x) \leq D(Q\|P) + \log \int \exp(f(x))dP(x).$$

Proof. By the definition of KL-divergence, we have

$$\begin{aligned}
& D(Q\|P) + \log \int \exp(f(x))dP(x) \\
= & \int \log\left(\frac{dQ(x)\int \exp(f(y))dP(y)}{dP(x)}\right) dQ(x) \\
= & \int \log\left(\frac{dQ(x)\int \exp(f(y))dP(y)}{\exp(f(x))dP(x)}\right) dQ(x) + \int f(x)dQ(x) \\
= & D(Q\|\widetilde{P}) + \int f(x)dQ(x) \\
\geq & \int f(x)dQ(x),
\end{aligned}$$

where $\widetilde{P}$ is a probability measure given by

$$d\widetilde{P}(x) = \frac{\exp(f(x))dP(x)}{\int \exp(f(y))dP(y)}.$$

□

Then, we can use Lemma 5.1.1 to derive a useful bound for $P_0^{(n)}\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)})$.

Lemma 5.1.2. *For the $\widehat{Q}$ defined in (2.3), we have*

$$\begin{aligned} & P_0^{(n)}\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) \\ \leq & \inf_{a>0}\frac{1}{a}\left(\inf_{Q\in\mathcal{S}} P_0^{(n)}D(Q\|\Pi(\cdot|X^{(n)})) + \log P_0^{(n)}\Pi(\exp(aL(P_\theta^{(n)}, P_0^{(n)}))|X^{(n)})\right). \end{aligned}$$

Proof. By Lemma 5.1.1, we have

$$a\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) \leq D(\widehat{Q}\|\Pi(\cdot|X^{(n)})) + \log\Pi(\exp(aL(P_\theta^{(n)}, P_0^{(n)}))|X^{(n)}),$$

for all $a > 0$. By the definition of $\widehat{Q}$, we have

$$D(\widehat{Q}\|\Pi(\cdot|X^{(n)})) \leq D(Q\|\Pi(\cdot|X^{(n)})),$$

for all $Q \in \mathcal{S}$. Taking expectation on both sides, we have

$$aP_0^{(n)}\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) \leq P_0^{(n)}D(Q\|\Pi(\cdot|X^{(n)})) + P_0^{(n)}\log\Pi(\exp(aL(P_\theta^{(n)}, P_0^{(n)}))|X^{(n)}).$$

Using Jensen's inequality, we get

$$P_0^{(n)}\log\Pi(\exp(aL(P_\theta^{(n)}, P_0^{(n)}))|X^{(n)}) \leq \log P_0^{(n)}\Pi(\exp(aL(P_\theta^{(n)}, P_0^{(n)}))|X^{(n)}).$$

Therefore,

$$P_0^{(n)}\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) \leq \frac{1}{a}\left(P_0^{(n)}D(Q\|\Pi(\cdot|X^{(n)})) + \log P_0^{(n)}\Pi(\exp(aL(P_\theta^{(n)}, P_0^{(n)}))|X^{(n)})\right).$$

The proof is complete by taking minimum over $a > 0$ and $Q \in \mathcal{S}$. □

In order to bound $P_0^{(n)}\Pi(\exp(aL(P_\theta^{(n)}, P_0^{(n)}))|X^{(n)})$, we need the following lemma on the posterior tail probability. Its proof is similar to the one used in [31].

Lemma 5.1.3. *Under the conditions of Theorem 2.2.1, we have*

$$P_0^{(n)}\Pi\left(L(P_\theta^{(n)}, P_0^{(n)}) > C_1 n\epsilon^2 | X^{(n)}\right) \leq \exp(-Cn\epsilon^2) + \exp(-\lambda n\epsilon^2) + 2\exp(-n\epsilon^2),$$

for all $\epsilon \geq \epsilon_n$, where $\lambda = \rho - 1$ for ρ in (C3).

Proof. We first define the sets

$$U_n = \left\{\theta : L(P_\theta^{(n)}, P_0^{(n)}) > C_1 n\epsilon^2\right\}, \qquad K_n = \left\{\theta : D_{1+\lambda}(P_0^{(n)} \| P_\theta^{(n)}) \leq C_3 n\epsilon_n^2\right\}.$$

We also define the event

$$A_n = \left\{X^{(n)} : \int \frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X^{(n)}) d\widetilde{\Pi}(\theta) \leq \exp(-(C_3+1)n\epsilon^2)\right\},$$

where the probability measure $\widetilde{\Pi}$ is defined as $\widetilde{\Pi}(B) = \frac{\Pi(B \cap K_n)}{\Pi(K_n)}$. Let $\Theta_n(\epsilon)$ and ϕ_n be the set and the testing function in (C1). Then, we bound $P_0^{(n)}\Pi(U_n|X^{(n)})$ by

$$\begin{aligned}
& P_0^{(n)}\Pi(U_n|X^{(n)}) \\
\leq\ & P_0^{(n)}\phi_n + P_0^{(n)}(A_n) + P_0^{(n)}(1-\phi_n)\Pi(U_n|X^{(n)})\mathbb{I}_{A_n^c} \\
=\ & P_0^{(n)}\phi_n + P_0^{(n)}(A_n) + P_0^{(n)}\frac{\int_{U_n}\frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X^{(n)})d\Pi(\theta)}{\int \frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X^{(n)})d\Pi(\theta)}(1-\phi_n)\mathbb{I}_{A_n^c}.
\end{aligned}$$

We will give bounds for the three terms above respectively. By (C1),

$$P_0^{(n)}\phi_n \leq \exp(-Cn\epsilon^2). \tag{5.1}$$

Using the definitions of A_n, we have

$$\begin{aligned}
P_0^{(n)}(A_n) &= P_0^{(n)}\left(\left(\int \frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X^{(n)})d\widetilde{\Pi}(\theta)\right)^{-\lambda} > \exp(\lambda(C_3+1)n\epsilon^2)\right) \\
&\leq \exp(-\lambda(C_3+1)n\epsilon^2)P_0^{(n)}\left(\int \frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X^{(n)})d\widetilde{\Pi}(\theta)\right)^{-\lambda} \\
&\leq \exp(-\lambda(C_3+1)n\epsilon^2)\int\left(\int \frac{(dP_0^{(n)})^{1+\lambda}}{(dP_\theta^{(n)})^{\lambda}}\right)d\widetilde{\Pi}(\theta) \\
&= \exp(-\lambda(C_3+1)n\epsilon^2)\int \exp(\lambda D_{1+\lambda}(P_0^{(n)}\|P_\theta^{(n)}))d\widetilde{\Pi}(\theta) \\
&\leq \exp(-\lambda(C_3+1)n\epsilon^2+\lambda C_3 n\epsilon_n^2) \\
&\leq \exp(-\lambda n\epsilon^2). \qquad (5.2)
\end{aligned}$$

Now we analyze the third term. On the event A_n^c, we have

$$\int \frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X^{(n)})d\Pi(\theta) \geq \Pi(K_n)\int \frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X^{(n)})d\widetilde{\Pi}(\theta) \geq \exp(-(C_2+C_3+1)n\epsilon^2),$$

where the last inequality is by (C3). Then, it follows that

$$\begin{aligned}
& P_0^{(n)}\frac{\int_{U_n}\frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X^{(n)})d\Pi(\theta)}{\int \frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X^{(n)})d\Pi(\theta)}(1-\phi_n)\mathbb{I}_{A_n^c} \\
\leq\ & \exp((C_3+C_2+1)n\epsilon^2)P_0^{(n)}\int_{U_n}\frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X^{(n)})(1-\phi_n)d\Pi(\theta) \\
\leq\ & \exp((C_3+C_2+1)n\epsilon^2)\left[\int_{U_n\cap\Theta_n(\epsilon)}P_\theta^{(n)}(1-\phi_n)d\Pi(\theta)+\Pi(\Theta_n(\epsilon)^c)\right] \\
\leq\ & \exp((C_3+C_2+1)n\epsilon^2)(\exp(-Cn\epsilon^2)+\exp(-Cn\epsilon^2)),
\end{aligned}$$

where the last inequality is by (C1) and (C2). Since $C > C_3 + C_2 + 2$, we obtain the bound

$$P_0^{(n)} \frac{\int_{U_n} \frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X_i) d\Pi(\theta)}{\int \frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X_i) d\Pi(\theta)} (1 - \phi_n) \mathbb{I}_{A_n^c} \leq 2 \exp(-n\epsilon^2). \tag{5.3}$$

Combining the bounds (5.1), (5.2) and (5.3), we have

$$P_0^{(n)} \Pi(U_n | X^{(n)}) \leq \exp(-Cn\epsilon^2) + \exp(-\lambda n\epsilon^2) + 2\exp(-n\epsilon^2).$$

□

Next, we derive a moment generating function bound for a sub-exponential random variable.

Lemma 5.1.4. *Suppose the random variable X satisfies*

$$\mathbb{P}(X \geq t) \leq c_1 \exp(-c_2 t),$$

for all $t \geq t_0 > 0$. Then, for any $0 < a \leq \frac{1}{2} c_2$,

$$\mathbb{E} \exp(aX) \leq \exp(at_0) + c_1.$$

Proof. Set $Y = \exp(aX)$ for some $0 < a \leq \frac{1}{2} c_2$. Then, for any $M_0 > 0$.

$$\begin{aligned} \mathbb{E}Y &\leq M_0 + \int_{M_0}^{\infty} \mathbb{P}(Y \geq y) dy \\ &= M_0 + \int_{M_0}^{\infty} \mathbb{P}\left(X \geq \frac{1}{a} \log y \right) dy \leq M_0 + c_1 \int_{M_0}^{\infty} y^{-c_2/a} dy. \end{aligned}$$

Choose $M_0 = \exp(at_0)$, and then since $a \leq \frac{1}{2} c_2$, we have

$$\mathbb{E}Y \leq \exp(at_0) + \frac{c_1 a}{c_2 - a} \exp((a - c_2)t_0) \leq \exp(at_0) + c_1 \exp(-at_0) \leq \exp(at_0) + c_1.$$

□

Now we are ready to prove Theorem 2.2.1.

Proof of Theorem 2.2.1. By Lemma 5.1.3, we have

$$P_0^{(n)}\Pi\left(L(P_\theta^{(n)}, P_0^{(n)}) > t|X^{(n)}\right) \leq c_1 \exp(-c_2 t),$$

for all $t \geq t_0$. Here, $c_1 = 4$, $c_2 = \min\{\lambda, 1\}/C_1$ as $C > C_1 + C_2 + 2 > 1$ and $t_0 = C_1 n\epsilon_n^2$. Then, by Lemma 5.1.4, we have

$$P_0^{(n)}\Pi\left(\exp\left(aL(P_\theta^{(n)}, P_0^{(n)})\right)|X^{(n)}\right) \leq \exp\left(aC_1 n\epsilon_n^2\right) + 4,$$

for all $a \leq \min\{\lambda, 1\}/(2C_1)$. Taking $a = \min\{\lambda, 1\}/(2C_1)$ and using Lemma 5.1.2, we get

$$\begin{aligned} P_0^{(n)}\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) &\leq \frac{n\gamma_n^2 + \log(4 + e^{aC_1 n\epsilon_n^2})}{a} \leq \frac{n\gamma_n^2}{a} + C_1 n\epsilon_n^2 + \frac{4e^{-aC_1 n\epsilon_n^2}}{a} \\ &\leq Mn(\gamma_n^2 + \epsilon_n^2), \end{aligned}$$

with some $M > 0$ that only depends on C, C_1, λ. □

5.1.2 Proofs of Theorem 2.2.3, Theorem 2.2.4 and Theorem 2.4.1

Proof of Theorem 2.2.3. For any $Q \in \mathcal{S} \cap \mathcal{E}$, we have $\mathrm{supp}(Q) \subset \mathcal{C}$, and thus $QD(P_0^{(n)}\|P_\theta^{(n)}) \leq C_2 n\epsilon_n^2$. By (C4*), we have $D(Q\|\Pi) \leq C_1 n\epsilon_n^2$. Therefore, $R(Q) \leq (C_1 + C_2)n\epsilon_n^2$, and the proof is complete. □

Proof of Theorem 2.2.4. It is sufficient to find a $Q \in \mathcal{S}_{\mathrm{MF}}$ and bound

$$R(Q) = \frac{1}{n}\left(D(Q\|\Pi) + QD(P_0^{(n)}\|P_\theta^{(n)})\right).$$

We choose Q to be the product measure $dQ(\theta) = \prod_{j=1}^m dQ_j(\theta_j)$, with

$$Q_j(B_j) = \frac{\widetilde{Q}_j(B_j \cap \widetilde{\Theta}_j)}{\widetilde{Q}_j(\widetilde{\Theta}_j)}.$$

Then, it is easy to see that $Q \in \mathcal{S}_{\mathrm{MC}}$ and $\mathrm{supp}(Q) \subset \otimes_{j=1}^m \widetilde{\Theta}_j$. By (2.8), we have

$$QD(P_0^{(n)} \| P_\theta^{(n)}) \leq C_1 n \epsilon_n^2.$$

Moreover, we can write $D(Q\|\Pi)$ as below

$$D(Q\|\Pi) = Q \log \frac{dQ}{d\widetilde{Q}} + Q \log \frac{d\widetilde{Q}}{d\Pi},$$

where

$$Q \log \frac{dQ}{d\widetilde{Q}} = -\sum_{j=1}^m \log \widetilde{Q}_j(\widetilde{\Theta}_j) \leq C_3 n \epsilon_n^2,$$

by (2.9), and

$$Q \log \frac{d\widetilde{Q}}{d\Pi} \leq C_2 n \epsilon_n^2,$$

by (2.8). Hence, we obtain the desired bound. □

To show Theorem 2.4.1, we need a model selection version of Lemma 5.1.2:

Lemma 5.1.5. *For $\widehat{Q}^{(\widehat{k})}$ defined as the solution of (2.34),*

$$\begin{aligned} & P_0^{(n)} \left[\widehat{Q}^{(\widehat{k})} L\left(P^{(n)}_{\widehat{k},\theta^{(\widehat{k})}}, P_0^{(n)} \right) \right] \\ \leq & \inf_{a>0} \frac{1}{a} \left[\min_{k \in \mathcal{K}} \min_{Q^{(k)} \in \mathcal{S}^{(k)}_{\mathrm{MF}}} \left\{ D\left(Q^{(k)} \| \Pi^{(k)} \right) + Q^{(k)} D\left(P_0^{(n)} \| P^{(n)}_{k,\theta^{(k)}} \right) - \log \pi(k) \right\} \right. \\ & \left. + P_0^{(n)} \log \Pi\left(\exp\left(a L(P^{(n)}_{k,\theta^{(k)}}, P_0^{(n)}) \right) \middle| X^{(n)} \right) \right], \end{aligned}$$

where Π is the prior distribution on $P_{k,\theta^{(k)}}$ induced by the sampling process of $(k, \theta^{(k)})$.

Proof. We use $p_0^{(n)}$, $p_{k,\theta^{(k)}}^{(n)}$ to denote the densities of $P_0^{(n)}$, $P_{k,\theta^{(k)}}^{(n)}$. A lower bound can be directly derived from the right hand side minus the left hand side. For any $a > 0$, any $k \in \mathcal{K}$, and any $Q^{(k)} \in \mathcal{S}_{\mathrm{MF}}^{(k)}$, we have

$$
\begin{aligned}
& D\left(Q^{(k)} \| \Pi^{(k)}\right) + Q^{(k)} D\left(P_0^{(n)} \| P_{k,\theta^{(k)}}^{(n)}\right) - \log \pi(k) \\
& -aP_0^{(n)}\left[\widehat{Q}^{(\widehat{k})} L\left(P_{\widehat{k},\theta^{(\widehat{k})}}^{(n)}, P_0^{(n)}\right)\right] \\
= \; & P_0^{(n)}\left(-F(Q^{(k)}, k) + \log p_0^{(n)}(X^{(n)})\right) - aP_0^{(n)}\left[\widehat{Q}^{(\widehat{k})} L\left(P_{\widehat{k},\theta^{(\widehat{k})}}^{(n)}, P_0^{(n)}\right)\right] \\
\geq \; & P_0^{(n)}\left(-F(\widehat{Q}^{(\widehat{k})}, \widehat{k}) + \log p_0^{(n)}(X^{(n)})\right) - aP_0^{(n)}\left[\widehat{Q}^{(\widehat{k})} L\left(P_{\widehat{k},\theta^{(\widehat{k})}}^{(n)}, P_0^{(n)}\right)\right] \\
= \; & P_0^{(n)} D\left(\widehat{Q}^{(\widehat{k})} \| \Pi^{(\widehat{k})}\right) + P_0^{(n)} \widehat{Q}^{(\widehat{k})} \log \frac{p_0^{(n)}(X^{(n)})}{p_{\widehat{k},\theta^{(\widehat{k})}}^{(n)}(X^{(n)})} - P_0^{(n)} \log \pi(\widehat{k}) \\
& -aP_0^{(n)}\left[\widehat{Q}^{(\widehat{k})} L\left(P_{\widehat{k},\theta^{(\widehat{k})}}^{(n)}, P_0^{(n)}\right)\right] \\
= \; & P_0^{(n)}\left[\widehat{Q}^{(\widehat{k})} \log \frac{d\widehat{Q}^{(\widehat{k})}(\theta^{(\widehat{k})}) p_0^{(n)}(X^{(n)})}{\pi(\widehat{k}) d\Pi^{(\widehat{k})}(\theta^{(\widehat{k})}) p_{\widehat{k},\theta^{(\widehat{k})}}^{(n)}(X^{(n)}) \exp\left(aL\left(P_{\widehat{k},\theta^{(\widehat{k})}}^{(n)}, P_0^{(n)}\right)\right)}\right] \\
= \; & D\left(P_0^{(n)} \| P_\Pi^{(n)}\right) \\
& + P_0^{(n)}\left[\widehat{Q}^{(\widehat{k})} \log \frac{d\widehat{Q}^{(\widehat{k})}(\theta^{(\widehat{k})}) p_\Pi^{(n)}(X^{(n)})}{\pi(\widehat{k}) d\Pi^{(\widehat{k})}(\theta^{(\widehat{k})}) p_{\widehat{k},\theta^{(\widehat{k})}}^{(n)}(X^{(n)}) \exp\left(aL\left(P_{\widehat{k},\theta^{(\widehat{k})}}^{(n)}, P_0^{(n)}\right)\right)}\right] \\
= \; & D\left(P_0^{(n)} \| P_\Pi^{(n)}\right) + P_0^{(n)} D\left(\widehat{Q}^{(\widehat{k})} \| \widetilde{\Pi}^{(\widehat{k})}\right) \\
& - P_0^{(n)} \log \frac{\int \pi(\widehat{k}) p_{\widehat{k},\theta^{(\widehat{k})}}^{(n)}(X^{(n)}) \exp\left(aL\left(P_{\widehat{k},\theta^{(\widehat{k})}}^{(n)}, P_0^{(n)}\right)\right) d\Pi^{(\widehat{k})}(\theta^{(\widehat{k})})}{p_\Pi^{(n)}(X^{(n)})} \\
\geq \; & - P_0^{(n)} \log \frac{\sum_{k \in \mathcal{K}} \int \pi(k) p_{k,\theta^{(k)}}^{(n)}(X^{(n)}) \exp\left(aL\left(P_{k,\theta^{(k)}}^{(n)}, P_0^{(n)}\right)\right) d\Pi^{(k)}(\theta^{(k)})}{p_\Pi^{(n)}(X^{(n)})} \\
= \; & - P_0^{(n)} \log \Pi\left(\exp\left(aL\left(P_{k,\theta^{(k)}}^{(n)}, P_0^{(n)}\right)\right) \middle| X^{(n)}\right),
\end{aligned}
$$

where $P_{\Pi}^{(n)}$ is the probability measure with the density $p_{\Pi}^{(n)}$ with

$$p_{\Pi}^{(n)}(X^{(n)}) = \sum_{k\in\mathcal{K}} \pi(k) \int p_{k,\theta^{(k)}}^{(n)}(X^{(n)}) d\Pi^{(k)}(\theta^{(k)}) = \int p_{k,\theta^{(k)}}^{(n)} d\Pi \left(P_{k,\theta^{(k)}}^{(n)} \right),$$

and

$$d\widetilde{\Pi}^{(k)}(\theta^{(k)}) = \frac{d\Pi^{(k)}(\theta^{(k)}) p_{k,\theta^{(k)}}^{(n)}(X^{(n)}) \exp\left(aL\left(P_{k,\theta^{(k)}}^{(n)}, P_0^{(n)} \right) \right)}{\int p_{k,\theta^{(k)}}^{(n)}(X^{(n)}) \exp\left(aL\left(P_{k,\theta^{(k)}}^{(n)}, P_0^{(n)} \right) \right) d\Pi^{(k)}(\theta^{(k)})}.$$

The proof is complete. □

Proof of Theorem 2.4.1. By Lemma 5.2.2, we have

$$\begin{aligned}
& P_0^{(n)} \left[\widehat{Q}^{(\widehat{k})} L\left(P_{\widehat{k},\theta^{(\widehat{k})}}^{(n)}, P_0^{(n)} \right) \right] \\
\leq & \inf_{a>0} \frac{1}{a} \left[\min_{k\in\mathcal{K}} \min_{Q^{(k)}\in\mathcal{S}_{\mathrm{MF}}^{(k)}} \left\{ D\left(Q^{(k)} \| \Pi^{(k)} \right) + Q^{(k)} D\left(P_0^{(n)} \| P_{k,\theta^{(k)}}^{(n)} \right) - \log \pi(k) \right\} \right. \\
& \left. + P_0^{(n)} \log \Pi \left(\exp\left(aL\left(P_{k,\theta^{(k)}}^{(n)}, P_0^{(n)} \right) \right) \middle| X^{(n)} \right) \right].
\end{aligned}$$

Now we analyze each term on the right hand side. By Jensen's Inequality together with Lemma 5.1.3 and Lemma 5.1.4, we have

$$\begin{aligned}
& P_0^{(n)} \log \Pi \left(\exp\left(aL\left(P_{k,\theta^{(k)}}^{(n)}, P_0^{(n)} \right) \right) \middle| X^{(n)} \right) \\
\leq & \log P_0^{(n)} \Pi \left(\exp\left(aL\left(P_{k,\theta^{(k)}}^{(n)}, P_0^{(n)} \right) \right) \middle| X^{(n)} \right) \lesssim n\epsilon_n^2,
\end{aligned}$$

with some small constant $a > 0$. This is because the conditions (C1) and (C2) with respect to prior Π hold by assumption, and (C3) is implied by (C3*) with the argument

$$\begin{aligned}
& \Pi\left(\left\{ P_{k,\theta^{(k)}}^{(n)} : D_\rho\left(P_0^{(n)} \| P_{k,\theta^{(k)}}^{(n)} \right) \leq C_3 n\epsilon_n^2 \right\} \right) \\
\geq & \Pi\left(\left\{ P_{k,\theta^{(k)}} : k = k_0, \theta^{(k_0)} \in \Theta^{(k_0)} \right\} \right) \\
\geq & \pi(k_0) \Pi^{(k_0)}(\Theta^{(k_0)}) \geq \exp\left(-C_2 n\epsilon_n^2 \right).
\end{aligned}$$

For the remaining terms, we choose $k = k_0$ and $dQ^{(k_0)} = \frac{d\Pi^{(k_0)}\mathbb{I}_{\Theta^{(k_0)}}}{\Pi^{(k_0)}(\Theta^{(k_0)})}$. According to prior structure, $Q^{(k_0)} \in \mathcal{S}_{\mathrm{MF}}^{(k_0)}$, and

$$\begin{aligned} & Q^{(k_0)} D\left(P_0^{(n)} \| P_{k_0,\theta^{(k_0)}}^{(n)}\right) \leq \max_{\theta^{(k)} \in \Theta^{(k_0)}} D\left(P_0^{(n)} \| P_{k_0,\theta^{(k_0)}}^{(n)}\right) \\ \leq & \max_{\theta^{(k)} \in \Theta^{(k_0)}} D_\rho\left(P_0^{(n)} \| P_{k_0,\theta^{(k_0)}}^{(n)}\right) \lesssim n\epsilon_n^2. \end{aligned}$$

We also have

$$D\left(Q^{(k_0)} \| \Pi^{(k_0)}\right) - \log \pi(k_0) = -\sum_{j=1}^{m_{k_0}} \Pi_j^{(k_0)}(\Theta_j^{(k_0)}) - \log \pi(k_0) \lesssim n\epsilon_n^2.$$

Hence, we obtain the desired result. □

5.1.3 Proofs of Theorem 2.3.1, Proposition 2.3.1, Theorem 2.3.2

Proof of Theorem 2.3.1. Theorem 2.3.1 can be regarded as a simple application of Corollary 4.2.1 with $\Theta_{j1} = \mathbb{R}\backslash\{0\}$, $\Theta_{j2} = \{0\}$ and $p(X_j^{(n)}|\theta_j) \propto \exp\left(-\frac{n}{2}(X_j - \theta_j)^2\right)$. We defer the proof of Corollary 4.2.1 to Section 5.3.1. □

To show Proposition 2.3.1, the following lemma is needed.

Lemma 5.1.6. *For the prior distribution* Π *defined in (2.12), we assume that* $\max_j \|f_j\|_\infty \leq a$ *and* $\pi(k)$ *is nonincreasing over* k*. Then, we have*

$$P_{\theta^*}^{(n)} \widetilde{k} \lesssim \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}},$$

for any $\theta^* \in \Theta_\alpha(B)$*, where* $\mathbb{P}_\theta = \otimes_{j=1}^\infty N(\theta_j, n^{-1/2})$*.*

Proof. We use the notation

$$W_j = \int f_j(\theta_j) \exp\left(-\frac{n(\theta_j - Y_j)^2}{2}\right) d\theta_j.$$

By the condition $\|f_j\|_\infty \le a$, we have $W_j \le a\sqrt{\frac{2\pi}{n}} \le 1$. Define the objective function

$$L(k) = \sum_{j<k} \log \frac{1}{W_j} + \sum_{j>k} \frac{nY_j^2}{2} - \log\left(\pi(k-1)\exp\left(-\frac{nY_k^2}{2}\right) + \pi(k)Z_k\right).$$

It is easy to check that

$$\widetilde{k} = \underset{k}{\operatorname{argmax}}\,(\pi(k-1|Y) + \pi(k|Y)) = \underset{k}{\operatorname{argmin}}\, L(k).$$

To give a bound for $\widetilde{k}$, we first study the difference $L(k_1) - L(k_2)$ for any $k_1 < k_2$. We use the inequalities

$$\log\left(\frac{\pi(k-1)\exp(-\frac{n}{2}Y_k^2) + \pi(k)W_k}{\pi(k-1)+\pi(k)}\right) \le \max\left\{-\frac{n}{2}Y_k^2, \log W_k\right\} \le 0,$$

and

$$\log\left(\frac{\pi(k-1)\exp(-\frac{n}{2}Y_k^2) + \pi(k)W_k}{\pi(k-1)+\pi(k)}\right) \ge \min\left\{-\frac{n}{2}Y_k^2, \log W_k\right\} \ge -\frac{n}{2}Y_k^2 + \log W_k.$$

Then, we have

$$\begin{aligned}
L(k_1) - L(k_2) &\le \sum_{j=k_1}^{k_2} \frac{nY_j^2}{2} + \sum_{j=k_1+1}^{k_2-1} \log W_j + \log\left(\frac{\pi(k_2-1)+\pi(k_2)}{\pi(k_1-1)+\pi(k_1)}\right) \\
&\le \sum_{j=k_1}^{k_2} \frac{nY_j^2}{2} - (k_2-k_1-1)\left(\frac{1}{2}\log n - \log(a\sqrt{2\pi})\right) \\
&\le n\sum_{j=k_1}^{k_2} \theta_j^{*2} + \sum_{j=k_1}^{k_2} Z_j^2 - (k_2-k_1-1)\left(\frac{1}{2}\log n - \log(a\sqrt{2\pi})\right) \\
&\le nB^2 k_1^{-2\alpha} + \sum_{j=k_1}^{k_2} Z_j^2 - (k_2-k_1-1)\left(\frac{1}{2}\log n - \log(a\sqrt{2\pi})\right),
\end{aligned}$$

where $Z_j \sim N(0,1)$. Now we bound $P_{\theta^*}^{(n)}\widetilde{k}$ by

$$P_{\theta^*}^{(n)}\widetilde{k} \leq Ck_0 + \sum_{l>Ck_0} lP_{\theta^*}^{(n)}(\widetilde{k}=l), \tag{5.4}$$

where $k_0 = \lceil \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}} \rceil$, and C is some large constant. For each $l > Ck_0$,

$$\begin{aligned}
P_{\theta^*}^{(n)}(\widetilde{k}=l) &\leq P_{\theta^*}^{(n)}\left(L(l) \leq L(k_0)\right) \\
&\leq \mathbb{P}\left(nB^2k_0^{-2\alpha} + \sum_{j=k_0}^{l} Z_j^2 - (l-k_0-1)\left(\frac{1}{2}\log n - \log(a\sqrt{2\pi})\right) \geq 0\right) \\
&\leq \mathbb{P}\left(\sum_{j=k_0}^{l} Z_j^2 \geq (l-k_0-1)\left(\frac{1}{2}\log n - \log(a\sqrt{2\pi})\right) - C_1\left(\frac{n}{\log n}\right)^{\frac{2\alpha}{2\alpha+1}}\right) \\
&\leq \mathbb{P}\left(\sum_{j=k_0}^{l} Z_j^2 \geq c(l-k_0-1)\log n\right),
\end{aligned}$$

where the last inequality is by the fact that $C_1\left(\frac{n}{\log n}\right)^{\frac{2\alpha}{2\alpha+1}}$ is of a smaller order than $(l-k_0-1)\log n$. Finally, a standard chi-squared tail bound gives

$$P_{\theta^*}^{(n)}(\widetilde{k}=l) \lesssim \exp\left(-C'(l-k_0)\log n\right).$$

Using (5.4) and summing over l, we get $P_{\theta^*}^{(n)}\widetilde{k} \lesssim k_0$, and the proof is complete. □

Proof of Proposition 2.3.1. According to Theorem 2.3.1, the variational posterior $\widehat{Q}$ is a product measure, and for any coordinate after a $\widetilde{k}$, the component is δ_0. By Theorem 5.1.6, we know that $P_{\theta^*}^{(n)}\widetilde{k} \leq C\left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}}$. Use the notation $\bar{k} = C\left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}}$. Then, we have $P_{\theta^*}^{(n)}\left(\widetilde{k} > 2\bar{k}\right) \leq 1/2$ by Markov inequality. Consider a θ^* with every entry zero except that $\theta^*_{\lceil 2\bar{k}\rceil} = B\lceil 2\bar{k}\rceil^{-\alpha}$. It is easy to check that $\theta^* \in \Theta_\alpha(B)$. For this θ^*, we have

$$P_{\theta^*}^{(n)}\widehat{Q}\|\theta-\theta^*\|^2 \geq P_{\theta^*}^{(n)}\widehat{Q}(\theta_{\lceil 2\bar{k}\rceil} - \theta^*_{\lceil 2\bar{k}\rceil})^2\mathbb{I}\{\widetilde{k} \leq 2\bar{k}\}$$

$$
\begin{aligned}
&= \theta^{*2}_{\lceil 2\bar{k} \rceil} P^{(n)}_{\theta^*} \left(\widetilde{k} \leq 2\bar{k} \right) \\
&\geq \frac{1}{2} \theta^{*2}_{\lceil 2\bar{k} \rceil} \\
&\asymp n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^{\frac{2\alpha}{2\alpha+1}}.
\end{aligned}
$$

Thus, the proof is complete. □

The proofs of Theorem 2.3.2 will be split into the following three lemmas. Recall that we use the loss $L(P^{(n)}_{\theta}, P^{(n)}_{\theta^*}) = n\|\theta - \theta^*\|^2$ for this model.

Lemma 5.1.7. *For the prior* Π *that satisfies (2.15), the conditions (C1) and (C2) hold for all* $\epsilon \geq n^{-1/2}$.

Proof. Given any $\epsilon \geq n^{-1/2}$ and any $C > 0$, we define

$$
\Theta_n(\epsilon) = \left\{ \theta = (\theta_j) : \theta_j = 0, \text{ for all } j > Cn\epsilon^2/C_2 \right\}.
$$

Then, by (2.15), we have

$$
\Pi(\Theta_n(\epsilon)^c) \leq \Pi(k > Cn\epsilon^2/C_2) \lesssim \exp\left(-Cn\epsilon^2 \right).
$$

This proves (C2). To show (C1), we consider the following testing problem,

$$
H_0 : \theta = \theta^*, \quad H_1 : \theta \in \Theta_n(\epsilon) \text{ and } \|\theta - \theta^*\|^2 \geq \widetilde{C}\epsilon^2.
$$

Define $N(\delta, S, d)$ as the δ-covering number of a set S under a metric d. Then, according to Lemma 5 in [32] and Theorem 7.1 in [31], it is sufficient to establish the bound

$$
\log N\left(\epsilon/8, \{\theta \in \Theta_n(\epsilon) : \|\theta - \theta^*\| \leq \epsilon\}, \|\cdot\|\right) \lesssim n\epsilon^2.
$$

This is obviously true given a standard volume ratio calculation in a Euclidean space of dimension $\lceil Cn\epsilon^2/C_2 \rceil$. Then, by Theorem 7.1 in [31], there exists a testing procedure ϕ_n

such that (C1) holds. Note that the testing error can be arbitrarily small given a sufficiently large $\widetilde{C} > 0$. □

Lemma 5.1.8. *Assume $\theta^* \in \Theta_\alpha(B)$. For the prior Π that satisfies (2.16) and (2.17), the conditions (C3) and (C4) hold for $\epsilon_n = n^{-\frac{\alpha}{2\alpha+1}}(\log n)^{\frac{\alpha}{2\alpha+1}}$.*

Proof. We first show (C4). We will apply Theorem 2.2.4 by constructing a $\widetilde{Q} \in \mathcal{S}_{\mathrm{MF}}$ and $\otimes_j \widetilde{\Theta}_j$ that satisfy the conditions (2.8) and (2.9). Define $\widetilde{\Theta}_j = [\theta_j^* - n^{-1/2}, \theta_j^* + n^{-1/2}]$ for all $j \le k_0$ and $\widetilde{\Theta}_j = \{0\}$ for all $j > k_0$, where $k_0 = \left\lceil \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}} \right\rceil$ is the same as defined in 2.16. We also define the measure $\widetilde{Q}$ by

$$d\widetilde{Q}(\theta) = \prod_{j=1}^{k_0} f_j(\theta_j) \prod_{j>k_0} \delta_0(\theta_j) d\theta.$$

It is easy to see that $\widetilde{Q} \in \mathcal{S}_{\mathrm{MF}}$. For any $\theta \in \otimes_j \widetilde{\Theta}_j$, we have

$$D_2(P_{\theta^*}^{(n)} \| P_\theta^{(n)}) = 2D(P_{\theta^*}^{(n)} \| P_\theta^{(n)}) = n\|\theta - \theta^*\|^2 \le k_0 \lesssim n\epsilon_n^2, \tag{5.5}$$

and

$$\log \frac{d\widetilde{Q}(\theta)}{d\Pi(\theta)} \le \log \frac{1}{\pi(k_0)} \le -\log C_3 + C_4 k_0 \log k_0 \lesssim n\epsilon_n^2. \tag{5.6}$$

Therefore, the condition (2.8) holds. To check the condition (2.9), we use the bound

$$\begin{aligned} -\sum_{j=1}^{\infty} \log \widetilde{Q}_j(\widetilde{\Theta}_j) &= -\sum_{j=1}^{k_0} \log \widetilde{Q}_j(\widetilde{\Theta}_j) = -\sum_{j=1}^{k_0} \log \int_{\theta_j^* - n^{-1/2}}^{\theta_j^* + n^{-1/2}} f_j(x) dx \\ \le \; & -k_0 \log(2n^{-1/2}) - \frac{1}{2n^{-1/2}} \sum_{j=1}^{k_0} \int_{\theta_j^* - n^{-1/2}}^{\theta_j^* + n^{-1/2}} \log f_j(x) dx, \end{aligned}$$

where we have used Jensen's inequality above. We are going to bound each of the integral

above using (2.17). For any $j \le k_0$, we have

$$\begin{aligned} & -\frac{1}{2n^{-1/2}} \int_{\theta_j^*-n^{-1/2}}^{\theta_j^*+n^{-1/2}} \log f_j(x)dx \le c_0 + c_1 j^{2\alpha+1}(3\theta_j^{*2} + n^{-1}) \\ \le \;& c_0 + 3c_1 k_0 j^{2\alpha}\theta_j^{*2} + c_1 k_0^{2\alpha+1} n^{-1} \le c_0 + c_1 + 3c_1 k_0 j^{2\alpha}\theta_j^{*2}. \end{aligned}$$

Hence, we get

$$-\sum_{j=1}^{\infty} \log \widetilde{Q}_j(\widetilde{\Theta}_j) \le \frac{1}{2} k_0 \log n + (c_0 + c_1 - \log 2)k_0 + 3c_1 k_0 \sum_j j^{2\alpha}\theta_j^{*2} \lesssim n\epsilon_n^2, \tag{5.7}$$

which implies that (2.9) holds. The condition (C4) is thus proved by applying Theorem 2.2.4.

Finally, we derive the condition (C3). In view of (5.5), there is a constant $C > 0$, such that

$$\begin{aligned} & -\log \Pi\left(D_2(P_{\theta^*}^{(n)} \| P_\theta^{(n)}) \le Cn\epsilon_n^2\right) \\ \le \;& -\log \pi(k_0) - \log \widetilde{Q}\left(D_2(P_{\theta^*}^{(n)} \| P_\theta^{(n)}) \le Cn\epsilon_n^2\right) \\ \le \;& -\log \pi(k_0) - \sum_{j=1}^{\infty} \log \widetilde{Q}_j(\widetilde{\Theta}_j) \lesssim n\epsilon_n^2. \end{aligned}$$

The last inequality above is by (5.6) and (5.7). Hence, the proof is complete. □

Proofs of Theorem 2.3.2. The results are directly implied by Lemma 5.1.7, Lemma 5.1.8. □

5.1.4 Proof of Theorem 2.3.3

For Theorem 2.3.3, the loss function is $L(P_\theta^n, P_{\theta^*}^n) = nH^2(P_\theta, P_{\theta^*})$. We split the proof of Theorem 2.3.3 into following two lemmas.

Lemma 5.1.9. *Assume $\theta^* \in \Theta_\alpha(B)$ for $\alpha > 1/2$. For the prior Π that satisfies (2.20) and (2.22), the conditions (C1) and (C2) hold for all $\epsilon \ge \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}$.*

Lemma 5.1.10. *Assume $\theta^* \in \Theta_\alpha(B)$ for $\alpha > 1/2$. For the prior Π that satisfies (2.21) and (2.23), the conditions (C3) and (C4) hold for $\epsilon_n^2 = \left(\frac{\log n}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$.*

Before proving these two lemmas, we need the following two results that establish relations between different divergence functions for the exponential family model.

Lemma 5.1.11. *If $\|\theta - \theta'\|_1 \leq \frac{1}{\sqrt{2}}$, then*

$$H(P_\theta, P_{\theta'}) \leq 2\sqrt{2}\|\theta - \theta'\|_1.$$

Proof. We first give some uniform bounds that are well known for exponential family density functions (see [52]). For any θ, θ', we have

$$\left\| \log \frac{dP_\theta}{dP_{\theta'}} \right\|_\infty \leq 2\sqrt{2}\|\theta - \theta'\|_1. \tag{5.8}$$

We start from the left hand side of the inequality:

$$\begin{aligned}
& H^2(P_\theta, P_{\theta'})^2 = \frac{1}{2}\int \left(\sqrt{\frac{dP_\theta}{dP_{\theta'}}} - 1\right)^2 dP_{\theta'} \\
\leq\ & \frac{1}{2}\int \left(\exp(\sqrt{2}\|\theta - \theta'\|_1) - 1\right)^2 dP_{\theta'} + \frac{1}{2}\int \left(\exp(-\sqrt{2}\|\theta - \theta'\|_1) - 1\right)^2 dP_{\theta'} \\
\leq\ & \frac{1}{2}\int 8\|\theta - \theta'\|_1^2 dP_{\theta'} + \frac{1}{2}\int 8\|\theta - \theta'\|_1^2 dP_{\theta'} \\
=\ & 8\|\theta - \theta'\|_1^2,
\end{aligned}$$

where we have applied the property that $\frac{e^x - 1}{x}$ is monotonically increasing for all x. Then it follows that $H(P_\theta, P_{\theta'}) \leq 2\sqrt{2}\|\theta - \theta'\|_1$. □

Lemma 5.1.12. *For any θ and any $\theta^* \in \Theta_\alpha(B)$ with $\alpha > 1/2$, we have*

$$\begin{aligned}
& C_0^{-1} \exp\left(-3\sqrt{2}\|\theta^* - \theta\|_1\right) \|\theta^* - \theta\|^2 \leq 2H^2(P_{\theta^*}, P_\theta) \leq D(P_{\theta^*}\|P_\theta) \\
& \leq D_2(P_{\theta^*}\|P_\theta) \leq C_0 \exp\left(3\sqrt{2}\|\theta^* - \theta\|_1\right) \|\theta^* - \theta\|^2,
\end{aligned}$$

where the constant $C_0 > 0$ only depends on α and B.

Proof. For any $\theta^* \in \Theta_\alpha(B)$, we have $\left\|\log \frac{dP_{\theta^*}}{d\ell}\right\|_\infty \leq 2\sqrt{2}\|\theta^*\|_1$. Since

$$\|\theta^*\|_1^2 \leq \left(\sum_{j=1}^{\infty} j^{-2\alpha}\right)\left(\sum_{j=1}^{\infty} j^{2\alpha}\theta_j^{*2}\right) \leq B^2\gamma_\alpha, \tag{5.9}$$

where $\gamma_\alpha = \sum_{j=1}^{\infty} j^{-2\alpha} = O(1)$ for $\alpha > 1/2$. This gives

$$\left\|\log \frac{dP_{\theta^*}}{d\ell}\right\|_\infty \leq 2\sqrt{2}\gamma_\alpha^{1/2}B. \tag{5.10}$$

Now we proceed to show Lemma 5.1.12. Given the result of Proposition 2.2.1, it is sufficient to prove the first and the last inequalities. Define

$$V(P_{\theta^*}, P_\theta) = \int \left(\log \frac{dP_{\theta^*}}{dP_\theta} - D(P_{\theta^*}\|P_\theta)\right)^2 dP_{\theta^*}.$$

Following the argument in the proof of Lemma 3.2 in [29], we have

$$e^{-\left\|\log \frac{dP_{\theta^*}}{d\ell}\right\|_\infty}\|\theta^* - \theta\|^2 \leq V(P_{\theta^*}, P_\theta) \leq 4H^2(P_{\theta^*}, P_\theta)e^{3/2\|\log \frac{dP_\theta}{dP_{\theta^*}}\|_\infty}.$$

By (5.8) and (5.10), we have

$$C_0^{-1}\|\theta - \theta^*\|^2 \leq 2H^2(P_{\theta^*}, P_\theta)\exp\left(3\sqrt{2}\|\theta - \theta^*\|_1\right),$$

for $C_0 = 2\exp(2\sqrt{2}\gamma_\alpha^{1/2}B)$, which implies the first inequality.

For the last inequality, we have

$$\begin{aligned} D_2(P_{\theta^*}\|P_\theta) &= \log\left(\int dP_{\theta^*}\exp\left(\log\frac{dP_{\theta^*}}{dP_\theta}\right)\right) \\ &= \log\left(1 + \sum_{l=1}^{\infty}\frac{1}{l!}\int dP_{\theta^*}\left(\log\frac{dP_{\theta^*}}{dP_\theta}\right)^l\right) \end{aligned}$$

$$
\begin{aligned}
&\leq \log\left(1 + D(P_{\theta^*}\|P_\theta)\sum_{l=1}^{\infty}\frac{1}{l!}\left\|\log\frac{dP_{\theta^*}}{dP_\theta}\right\|_\infty^{l-1}\right)\\
&\leq D(P_{\theta^*}\|P_\theta)\exp\left(\left\|\log\frac{dP_{\theta^*}}{dP_\theta}\right\|_\infty\right)\\
&\leq D(P_{\theta^*}\|P_\theta)e^{2\sqrt{2}\|\theta-\theta^*\|_1},
\end{aligned}
$$

where we have used the inequality that $\frac{e^x-1}{x} \leq e^x$ for all $x > 0$ and the last inequality is by (5.8). By the same argument in the proof of Lemma 3.2 in [29], we have

$$
D(P_{\theta^*}\|P_\theta) \leq e^{\sqrt{2}\|\theta-\theta^*\|_1+2\sqrt{2}\|\theta^*\|_1}\|\theta-\theta^*\|^2.
$$

Therefore, we obtain the bound

$$
D_2(P_{\theta^*}\|P_\theta) \leq e^{3\sqrt{2}\|\theta-\theta^*\|_1+2\sqrt{2}\|\theta^*\|_1}\|\theta-\theta^*\|^2,
$$

which implies the desired result by (5.9). □

Now we are ready to prove Lemma 5.1.9 and Lemma 5.1.10.

Proof of Lemma 5.1.9. Given any $\epsilon \geq \left(\frac{\log n}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$, we define the set

$$
\Theta_n(\epsilon) = \left\{\theta = (\theta_j) : \theta_j \in [-w_n, w_n] \text{ for } 1 \leq j \leq k_n, \theta_j = 0 \text{ for } j > k_n\right\},
$$

where $w_n = (\widetilde{C}n\epsilon^2)^{1/\beta}$ and $k_n = \left\lceil \frac{\widetilde{C}n\epsilon^2}{\log(n\epsilon^2)} \right\rceil$. We bound $\Pi(\Theta_n(\epsilon)^c)$ by

$$
\begin{aligned}
\Pi(\Theta_n(\epsilon)^c) &\leq \Pi(k > k_n) + \sum_{j=1}^{k_n}\Pi(k=j)\sum_{i=1}^{j}\Pi(|\theta_i| > w_n | k = j)\\
&\leq \Pi(k > k_n) + \sum_{j=1}^{k_n}\Pi(k=j)\sum_{i=1}^{j}\int_{|x|>w_n} f_i(x)dx\\
&\leq \Pi(k > k_n) + \sum_{j=1}^{k_n}\int_{|x|>w_n} f_j(x)dx
\end{aligned}
$$

$$
\begin{aligned}
&\leq \Pi(k > k_n) + \sum_{j=1}^{k_n} e^{-c_1 w_n^\beta/2} \int e^{c_1|x|^\beta/2} f_j(x)dx \\
&\leq \Pi(k > k_n) + \sum_{j=1}^{k_n} e^{-c_1 w_n^\beta/2} \int e^{c_1|x|^\beta/2 - c_0 - c_1|x|^\beta} dx \\
&\lesssim \exp(-C_2 k_n \log k_n) + k_n \exp\left(-c_1 \widetilde{C} n\epsilon^2/2\right),
\end{aligned}
$$

where we have used the conditions (2.20) and (2.22). Therefore, for any $C > 0$, we can choose a sufficiently large $\widetilde{C}$, such that $\Pi(\Theta_n(\epsilon)^c) \lesssim \exp(-Cn\epsilon^2)$, which proves (C2).

To prove (C1), we consider the following testing problem,

$$
H_0 : \theta = \theta^*, \quad H_1 : \theta \in \Theta_n(\epsilon) \text{ and } H(P_\theta, P_{\theta^*}) \geq C'\epsilon.
$$

By Theorem 7.1 in [31], it is sufficient to establish the bound

$$
\log N(\epsilon, \{P_\theta : \theta \in \Theta_n(\epsilon)\}, H) \lesssim n\epsilon^2.
$$

Note that for any $\theta, \theta' \in \Theta_n(\epsilon)$, we have $\|\theta - \theta'\|_1 \leq \sqrt{k_n}\|\theta - \theta'\|$. Therefore, by Lemma 5.1.11,

$$
H(P_\theta, P_{\theta'}) \lesssim \|\theta - \theta'\|_1 \leq \sqrt{k_n}\|\theta - \theta'\|,
$$

when $\|\theta - \theta'\|_1 \leq \frac{1}{\sqrt{2}}$. This means as long as $\|\theta - \theta'\| \leq k_n^{-1/2}(\epsilon \wedge 2^{-1/2})$, we have $H(P_\theta, P_{\theta'}) \lesssim \epsilon$. Thus, there exists a constant c', such that

$$
\begin{aligned}
&\log N(\epsilon, \{P_\theta : \theta \in \Theta_n(\epsilon)\}, H) \\
\leq\ &\log N\left(c' k_n^{-1/2}(\epsilon \wedge 2^{-1/2}), \{\theta \in \mathbb{R}^{k_n} : \|\theta\|^2 \leq k_n w_n^2\}, \|\cdot\|\right) \\
\lesssim\ &k_n \log\left(\frac{k_n w_n}{c'(\epsilon \wedge 2^{-1/2})}\right) \\
\lesssim\ &k_n \log(n\epsilon^2) \asymp n\epsilon^2,
\end{aligned}
$$

where we have used the condition $\epsilon \geq \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}$ in the last two steps above.

It implies the existence of a testing function that satisfies (C1). The testing error can be made arbitrarily small by choosing a sufficiently large C'. Hence, the proof is complete. $\square$

Proof of Lemma 5.1.10. In the first part of the proof, we derive (C3). We take $k_0 = \lceil (n/\log n)^{\frac{1}{2\alpha+1}} \rceil$. Define $\widetilde{\Theta} = \otimes_j \widetilde{\Theta}_j$, where $\widetilde{\Theta}_j = [\theta_j^* - n^{-1/2}, \theta_j^* + n^{-1/2}]$ for all $j \leq k_0$ and $\widetilde{\Theta}_j = \{0\}$ for all $j > k_0$. Then, by Lemma 5.1.11, for all $\theta \in \widetilde{\Theta}$,

$$\begin{aligned} D_2(P_{\theta^*} \| P_\theta) &\leq C_0 \exp(3\sqrt{2}\|\theta^* - \theta\|_1)\|\theta - \theta^*\|^2 \\ &= C_0 \exp\left(3\sqrt{2}\left(\frac{k_0}{\sqrt{n}} + \sum_{j>k_0} |\theta_j^*|\right)\right)\left(\frac{k_0}{n} + \sum_{j>k_0} \theta_j^{*2}\right) \\ &\leq C_0 \exp\left(3\sqrt{2}\left(n^{\frac{1-2\alpha}{2+4\alpha}} + B\gamma_\alpha^{1/2}\right)\right)\left(\frac{k_0}{n} + k_0^{-2\alpha}B^2\right) \\ &\lesssim n\epsilon_n^2. \end{aligned}$$

where we have use the condition $\alpha > 1/2$.

Therefore, it is sufficient to lower bound $\Pi(\widetilde{\Theta})$, which has been done in the proof of Lemma 5.1.8.

Now we will derive (C4). Rather than using the results of Theorem 2.2.3 or Theorem 2.2.4, we will construct a $Q \in \mathcal{S}_{\mathrm{G}}$ and bound $R(Q)$ directly. Note that in the current setting, we have

$$R(Q) = \frac{1}{n} D(Q\|\Pi) + Q D(P_{\theta^*} \| P_\theta).$$

For $k_0 = \lceil (n/\log n)^{\frac{1}{2\alpha+1}} \rceil$, define $Q = \otimes_j Q_j$, where $Q_j = N(\theta_j^*, n^{-1})$ for $j \leq k_0$ and $Q_j = N(0,0)$ for $j > k_0$. Then, it is easy to see that $Q \in \mathcal{S}_{\mathrm{G}}$.

We first give a bound for $D(Q\|\Pi)$. Let F_j denote the probability distribution with density function f_j. Then, we have

$$D(Q\|\Pi) \leq \log \frac{1}{\pi(k_0)} + \sum_{j=1}^{k_0} D\left(N(\theta_j^*, n^{-1}) \| F_j\right),$$

where the first term on the right hand side above can be bounded as

$$\log \frac{1}{\pi(k_0)} \lesssim k_0 \log k_0 \lesssim n\epsilon_n^2,$$

according to the condition (2.21). For any $j \leq k_0$, we use ψ_j to denote the density function of $N(\theta_j^*, n^{-1})$. Then, by (2.23), we have

$$\begin{aligned} D\left(N(\theta_j^*, n^{-1}) \| F_j\right) &= \int \psi_j \log \psi_j - \int \psi_j \log f_j \\ &\leq \int \psi_j \log \psi_j + c_0' + c_1' j^{2\alpha+1} \int \phi_j(x) x^2 dx \\ &= \frac{1}{2} \log\left(\frac{n}{2\pi e}\right) + c_0' + c_1' j^{2\alpha+1}(n^{-1} + \theta_j^{*2}). \end{aligned}$$

Since $\theta^* \in \Theta_\alpha(B)$, we have

$$\sum_{j=1}^{k_0} D\left(N(\theta_j^*, n^{-1}) \| F_j\right) \lesssim k_0 \log n \lesssim n\epsilon_n^2.$$

Therefore, we have obtained $D(Q\|\Pi) \lesssim n\epsilon_n^2$.

We then derive a bound for $QD(P_{\theta^*}\|P_\theta)$. For $j \leq k_0$, we write $\theta_j = \theta_j^* + \frac{1}{\sqrt{n}} Z_j$ where $Z_j \sim N(0,1)$. Then according to Lemma 5.1.12, it follows that

$$\begin{aligned} QD(P_{\theta^*}\|P_\theta) &\lesssim Q \exp\left(3\sqrt{2}\|\theta - \theta^*\|_1\right) \|\theta - \theta^*\|^2 \\ &= Q\left[e^{3\sqrt{2}\sum_{j=1}^{k_0} |\theta_j - \theta_j^*|} \left(\sum_{j=1}^{k_0} (\theta_j - \theta_j^*)^2 + \sum_{j>k_0} \theta_j^{*2}\right)\right] \\ &= \mathbb{E} e^{3\sqrt{2}\sum_{j=1}^{k_0} |Z_j|/\sqrt{n}} \sum_{j=1}^{k_0} Z_j^2/n + \left(\sum_{j>k_0} \theta_j^{*2}\right) \mathbb{E} e^{3\sqrt{2}\sum_{j=1}^{k_0} |Z_j|/\sqrt{n}}, \end{aligned} \tag{5.11}$$

where the last inequality is by (5.9). Suppose we can show

$$\mathbb{E} e^{3\sqrt{2}\sum_{j=1}^{k_0} |Z_j|/\sqrt{n}} = O(1), \tag{5.12}$$

and

$$\mathbb{E}Z_1^2 e^{3\sqrt{2}\sum_{j=1}^{k_0}|Z_j|/\sqrt{n}} = O(1). \tag{5.13}$$

Then, up to a constant, (5.11) can be bounded by

$$\frac{k_0}{n} + \sum_{j>k_0} \theta_j^{*2} \lesssim \epsilon_n^2,$$

which further implies $QD(P_{\theta^*} \| P_\theta) \lesssim \epsilon_n^2$.

To complete the proof, we show (5.12). We have

$$\begin{aligned}
\mathbb{E}e^{3\sqrt{2}\sum_{j=1}^{k_0}|Z_j|/\sqrt{n}} &\leq \mathbb{E}\exp\left(\frac{3\sqrt{2}}{\sqrt{n}}\sum_{j=1}^{k_0}(1+Z_j^2)\right) \\
&= \exp\left(\frac{3\sqrt{2}k_0}{\sqrt{n}}\right)\mathbb{E}\exp\left(\frac{3\sqrt{2}}{\sqrt{n}}\chi_{k_0}^2\right) \\
&= \exp\left(\frac{3\sqrt{2}k_0}{\sqrt{n}}\right)\left(1-\frac{6\sqrt{2}}{\sqrt{n}}\right)^{-\frac{k_0}{2}}.
\end{aligned}$$

Since $\alpha > 1/2$, we have $k_0/\sqrt{n} = O(1)$, and thus (5.12) holds. For (5.13), we have

$$\mathbb{E}Z_1^2 e^{3\sqrt{2}\sum_{j=1}^{k_0}|Z_j|/\sqrt{n}} = \left(\mathbb{E}Z_1^2 e^{3\sqrt{2}|Z_1|/\sqrt{n}}\right)\left(\mathbb{E}e^{3\sqrt{2}\sum_{j=2}^{k_0}|Z_j|/\sqrt{n}}\right).$$

Note that $\mathbb{E}Z_1^2 e^{3\sqrt{2}|Z_1|/\sqrt{n}} = O(1)$, and $\mathbb{E}e^{3\sqrt{2}\sum_{j=2}^{k_0}|Z_j|/\sqrt{n}}$ shares the same bound for (5.12). This implies (5.13) also holds. □

Proof of Theorem 2.3.3. The result is immediately implied by Lemma 5.1.9 and Lemma 5.1.10 in view of Theorem 2.2.2. □

5.1.5 *Proofs of Theorem 2.3.4, Theorem 2.3.5 and Theorem 2.3.6*

Proof of Theorem 2.3.4. Recall that Θ_k is the space of piecewise constant vectors with at most k pieces. Then, we have the partition

$$\mathbb{R}^n = \Theta_{n-1} \cup (\Theta_n \backslash \Theta_{n-1}).$$

First of all, we consider $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}$. Suppose the measure $Q \in \mathcal{S}_{\mathrm{MF}}$ and $D(Q\|\Pi) < \infty$, then the support of Q must be a subset of the support of Π. Note that the distributions g_i's are all absolutely continuous. That is, for any singleton x, $\Pi(\theta_j = x) = 0$, which indicates that $Q(\theta_j = x) = 0$ for any singleton x. Thus, Q is continuous in each coordinate and for any $j \in [n-1]$, $Q(\theta_j = \theta_{j+1}) = \int Q(\theta_j = \theta_{j+1} = x)dx = \int Q(\theta_j = x)Q(\theta_{j+1} = x)dx = 0$. Therefore,

$$Q\left(\Theta_{n-1}\right) = Q\left(\text{there exists a } j \in [n-1], \text{ such that } \theta_j = \theta_{j+1}\right) = 0,$$

because otherwise the independent structure of Q would imply a delta measure for some coordinate, which leads to $D(Q\|\Pi) = \infty$. This implies that Q is supported on $\Theta_n \backslash \Theta_{n-1}$. Therefore,

$$D(Q\|\Pi) = \int \log \frac{\prod_{i=1}^n q_i(\theta_i)}{\Pi(\Theta_n \backslash \Theta_{n-1}) \prod_{i=1}^n g(\theta_i)} dQ(\theta),$$

where $q_i(\theta_i) = \frac{dQ_i(\theta_i)}{d\theta_i}$. Then, by the definition of $\mathcal{S}_{\mathrm{MF}}$ and the independent structure of $P_\theta^{(n)}$, we have

$$\begin{aligned}
\widehat{Q}_{\mathrm{MF}} &= \operatorname*{argmin}_{Q \in \mathcal{S}_{\mathrm{MF}}} \left\{ D(Q\|\Pi) + QD(P_{\theta^*}^{(n)} \| P_\theta^{(n)}) \right\} \\
&= \operatorname*{argmin}_{Q: \frac{dQ(\theta)}{d\theta} = \prod_{i=1}^n q_i(\theta_i)} \left\{ Q \sum_{i=1}^n \left(\log \frac{q_i(\theta_i)}{g(\theta_i)} + D\left(N(\theta_i^*, \sigma^2) \| N(\theta_i, \sigma^2) \right) \right) \right\}.
\end{aligned}$$

This gives

$$\frac{d\widehat{Q}_{\mathrm{MF}}(\theta)}{d\theta} \propto \prod_{i=1}^{n} g(\theta_i) \exp\left(-\frac{(\theta_i - X_i)^2}{2\sigma^2}\right).$$

In other words, the mean-field variational posterior $\widehat{Q}_{\mathrm{MF}}$ is a product measure, and on each coordinate, it equals the posterior distribution induced by the prior g_i. Now we give a lower bound for $P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MF}} \|\theta - \theta^*\|^2$. Since $\|\theta - \theta^*\|^2 = \sum_{i=1}^{n} (\theta_i - \theta_i^*)^2$, we have

$$P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MF}} \|\theta - \theta^*\|^2 = \sum_{i=1}^{n} P_{\theta_i^*} \mathbb{E}\left((\theta_i - \theta_i^*)^2 | X_i\right),$$

where we use $\mathbb{E}(\cdot|X_i)$ to stand for the posterior expectation of θ_i with the prior $\theta_i \sim g_i$. By Jensen's inequality,

$$\mathbb{E}\left((\theta_i - \theta_i^*)^2 | X_i\right) \geq \left(\mathbb{E}(\theta_i|X_i) - \theta^*\right)^2.$$

Therefore,

$$\begin{aligned}
& \sup_{\theta^* \in \Theta_k(B)} P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MF}} \|\theta - \theta^*\|^2 \\
\geq & \sup_{\theta^* \in \Theta_1(B)} \sum_{i=1}^{n} P_{\theta_i^*} \left(\mathbb{E}(\theta_i|X_i) - \theta_i^*\right)^2 \\
\geq & \frac{1}{2} \sum_{i=1}^{n} P_{\theta_i^* = -B} \left(\mathbb{E}(\theta_i|X_i) - \theta_i^*\right)^2 + \frac{1}{2} \sum_{i=1}^{n} P_{\theta_i^* = B} \left(\mathbb{E}(\theta_i|X_i) - \theta_i^*\right)^2 \\
= & \frac{1}{2} \sum_{i=1}^{n} \left(P_{\theta_i^* = -B} \left(\mathbb{E}(\theta_i|X_i) - \theta_i^*\right)^2 + P_{\theta_i^* = B} \left(\mathbb{E}(\theta_i|X_i) - \theta_i^*\right)^2\right) \\
\geq & \sum_{i=1}^{n} B^2 \int \min\left(dN(B, \sigma^2), dN(-B, \sigma^2)\right) \\
\gtrsim & n.
\end{aligned}$$

Next, we consider $\mathcal{S} = \mathcal{S}_{\mathrm{MF}}^{\mathrm{joint}}$. As $\widehat{Q}_{\mathrm{MF}}^{\mathrm{joint}} \in \mathcal{S}_{\mathrm{MF}}^{\mathrm{joint}}$, we can assume

$$d\widehat{Q}_{\mathrm{MF}}^{\mathrm{joint}}(w, z, \theta) = d\widehat{Q}^{(w)}(w) \prod_{i=1}^{n} d\widehat{Q}_i^{(z)}(z) \prod_{i=1}^{n} d\widehat{Q}_i^{(\theta)}(\theta_i).$$

For the same reason, $\prod_{i=1}^n d\widehat{Q}^{(\theta)}(\theta_i)$ is supported on $\Theta_n \backslash \Theta_{n-1}$. The joint distribution of prior is written as

$$\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} w^{\alpha_0+\sum_{i=2}^n z_i - 1}(1-w)^{\beta_0+n-2-\sum_{i=2}^n z_i} g(\theta_1) \prod_{i=2}^n g(\theta_i)^{z_i} \delta_{\theta_{i-1}}^{1-z_i}.$$

Thus, conditioning on $\theta \in \Theta_n \backslash \Theta_{n-1}$, $z_i = 1$ for all $2 \le i \le n$. In other words, $\widehat{Q}_i^{(z)}(z_i = 1) = 1$ for all $2 \le i \le n$. Plug it in the definition of $\widehat{Q}_{\mathrm{MF}}^{\mathrm{joint}}$, we have

$$\begin{aligned}
&\left(\widehat{Q}^{(\theta)}, \widehat{Q}^{(w)}\right) \\
= &\underset{\substack{(Q^{(\theta)}, Q^{(w)}) \\ dQ^{(\theta)} = \prod_{i=1}^n q_i^{(\theta)}(\theta_i) d\theta \\ dQ^{(w)} = q^{(w)}(w) dw}}{\operatorname{argmin}} \left\{ Q^{(w)} \log \frac{q^{(w)}}{\pi(w) w^{n-1}} \right. \\
&\left. + Q^{(\theta)} \sum_{i=1}^n \left(\log \frac{q_i^{(\theta)}(\theta_i)}{g(\theta_i)} + D\left(N(\theta_i^*, \sigma^2) \| N(\theta_i, \sigma^2) \right) \right) \right\}.
\end{aligned}$$

This gives $\frac{d\widehat{Q}^{(w)}(w)}{dw} \propto \pi(w) w^{n-1}$ and $\widehat{Q}^{(\theta)}(\theta) = \widehat{Q}_{\mathrm{MF}}(\theta)$. It implies that

$$\begin{aligned}
&\sup_{\theta^* \in \Theta_k(B)} P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MF}}^{\mathrm{joint}} \|\theta - \theta^*\|^2 = \sup_{\theta^* \in \Theta_k(B)} P_{\theta^*}^{(n)} \widehat{Q}^{(\theta)} \|\theta - \theta^*\|^2 \\
= &\sup_{\theta^* \in \Theta_k(B)} P_{\theta^*}^{(n)} \widehat{Q}_{\mathrm{MF}} \|\theta - \theta^*\|^2 \gtrsim n
\end{aligned}$$

The proof is complete. □

Proof of Theorem 2.3.5. This theorem is a special case of Theorem 2.5.6, whose proof is given in Section 5.1.9. □

Proof of Theorem 2.3.6. If $D\left(Q(w,z,\theta) \| \Pi(w,z,\theta|Y)\right) < \infty$, we will have

$$\mathrm{supp}(Q) \subseteq \mathrm{supp}(\Pi(\cdot|Y)) \subseteq \mathrm{supp}(\Pi).$$

For $Q \in \mathcal{S}_{\text{MF}}^{\text{joint}}$, as $\Pi(z_i = 0, \theta_i \neq \theta_{i-1}) = \Pi(z_i = 1, \theta_i = \theta_{i-1}) = 0$, we can conclude that $Q_i^{(z)}(z_i = 0)Q_i^{(\theta)}(\theta_i \neq \theta_{i-1}|\theta_{i-1}) = 0$ and $Q_i^{(z)}(z_i = 1)Q_i^{(\theta)}(\theta_i = \theta_{i-1}|\theta_{i-1}) = 0$. In other words, the conclusion leads to $Q_i^{(z)}(z_i = 1) = 0$, $Q_i^{(\theta)}(\theta_i \neq \theta_{i-1}|\theta_{i-1}) = 0$ or $Q_i^{(z)}(z_i = 1) = 1$, $Q_i^{(\theta)}(\theta_i = \theta_{i-1}|\theta_{i-1}) = 0$.

Thus, we can define a set $S \subseteq \{2, 3, \cdots, n\}$, such that for $i \notin S$, $Q_i^{(z)}(z_i = 1) = 0$ and $dQ_i^{(\theta)}(\theta_i|\theta_{i-1}) = \delta_{\theta_{i-1}}(\theta_i)d\theta_i$, whereas for $i \in S$, $Q_i^{(z)}(z_i = 1) = 1$ and $dQ_i^{(\theta)}(\theta_i|\theta_{i-1}) = q_i^{(\theta)}(\theta_i|\theta_{i-1})d\theta_i$, a continuous density function. Then we can write

$$\frac{dQ(w, z, \theta)}{dwd\theta} = q^{(w)}(w)q_1^{(\theta)}(\theta_1) \prod_{i \in S} \mathbb{I}_{z_i=1}(z_i)q_i^{(\theta)}(\theta_i|\theta_{i-1}) \prod_{i \notin S} \mathbb{I}_{z_i=0}(z_i)\delta_{\theta_{i-1}}(\theta_i).$$

Plug it into $D(Q(w, z, \theta)\|\Pi(w, z, \theta|Y))$, and we get

$$\begin{aligned}
& D\left(Q(w, z, \theta)\|\Pi(w, z, \theta|Y)\right) \\
= & \int q^{(w)}(w) \log \frac{q^{(w)}(w)}{\pi(w)w^{|S|}(1-w)^{n-1-|S|}} dw \\
& + \int_{\Theta(S)} q_1^{(\theta)}(\theta_1) \prod_{i \in S} q_i^{(\theta)}(\theta_i|\theta_{i-1}) \prod_{i \notin S} \delta_{\theta_{i-1}}(\theta_i) \\
& \times \log \frac{q_1^{(\theta)}(\theta_1) \prod_{i \in S} q_i^{(\theta)}(\theta_i|\theta_{i-1}) \prod_{i \notin S} \delta_{\theta_{i-1}}(\theta_i)}{g(\theta_1) \prod_{i \in S} g(\theta_i) \prod_{i \notin S} \delta_{\theta_{i-1}}(\theta_i) \exp\left(-\frac{1}{2}\sum_{i=1}^n (Y_i - \theta_i)^2\right)} d\theta,
\end{aligned}$$

where

$$\Theta(S) = \{\theta|\theta_i = \theta_{i-1} \text{ for } i \notin S \text{ and } \theta_i \neq \theta_{i-1} \text{ for } i \in S\}.$$

Then

$$\begin{aligned}
& \min_{Q \in \mathcal{S}_{\text{MC}}^{\text{joint}}} D\left(Q(w, z, \theta)\|\Pi(w, z, \theta|Y)\right) \\
\Leftrightarrow \quad & \min_S \left\{ \min_{Q^{(\theta)} \in \mathcal{S}_{\text{MC}}, Q^{(w)}} \left\{ \int q^{(w)}(w) \log \frac{q^{(w)}(w)}{\pi(w)w^{|S|}(1-w)^{n-1-|S|}} dw \right.\right.
\end{aligned}$$

$$+\int_{\Theta(S)} q^{(\theta)}(\theta)\log\frac{q^{(\theta)}(\theta)}{g(\theta_1)\prod_{i\in S}g(\theta_i)\prod_{i\notin S}\delta_{\theta_{i-1}}(\theta_i)\exp\left(-\frac{1}{2}\sum_{i=1}^n(Y_i-\theta_i)^2\right)}d\theta\Bigg\}\Bigg\}, \tag{5.14}$$

For a given set $S=\{a_1+1,a_2+1,\cdots,a_{k-1}+1\}$ with $0=a_0<a_1<\cdots<a_{k-1}<a_k=n$, we first solve the minimization over $q^{(w)}$ and $q^{(\theta)}$. The solutions without constraint that $Q^{(\theta)}\in\mathcal{S}_{\mathrm{MC}}$ are given by

$$\widehat{q}^{(w)}(w)=\frac{\Gamma(n-1+\alpha_0+\beta_0)}{\Gamma(k-1+\alpha_0)\Gamma(n-k+\beta_0)}w^{k+\alpha_0-2}(1-w)^{n-k+\beta_0-1},$$

and

$$\begin{aligned}\widehat{q}^{(\theta)}(\theta) &= \frac{g(\theta_1)\prod_{i\in S}g(\theta_i)\prod_{i\notin S}\delta_{\theta_{i-1}}(\theta_i)\exp\left(-\frac{1}{2}\sum_{i=1}^n(Y_i-\theta_i)^2\right)}{\int g(\theta_1)\prod_{i\in S}g(\theta_i)\prod_{i\notin S}\delta_{\theta_{i-1}}(\theta_i)\exp\left(-\frac{1}{2}\sum_{i=1}^n(Y_i-\theta_i)^2\right)d\theta}\\ &= \prod_{j=1}^k\frac{g(\theta_{a_{j-1}+1})\exp\left(-\frac{1}{2}\sum_{i=a_{j-1}+1}^{a_j}(Y_i-\theta_{a_{j-1}+1})^2\right)}{\int g(\theta_{a_{j-1}+1})\exp\left(-\frac{1}{2}\sum_{i=a_{j-1}+1}^{a_j}(Y_i-\theta_{a_{j-1}+1})^2\right)d\theta_{a_{j-1}+1}}\prod_{i\notin S}\delta_{\theta_{i-1}}(\theta_i).\end{aligned}$$

As $\widehat{Q}^{(\theta)}$ obtained above is still in the variational set $\mathcal{S}_{\mathrm{MC}}$, this is a valid solution to (5.14) for a specific set S, which implies that

$$\widehat{Q}^{(w)}=\mathrm{Beta}(k-1+\alpha_0,n-k+\beta_0),$$

and

$$\begin{cases} d\widehat{Q}_1^{(\theta)}(\theta_1)\propto g(\theta_1)\exp\left(-\frac{1}{2}\sum_{i\in(a_0:a_1]}(X_i-\theta_1)^2\right)d\theta_1, & \\ d\widehat{Q}_i^{(\theta)}(\theta_i|\theta_{i-1})\propto g(\theta_i)\exp\left(-\frac{1}{2}\sum_{l\in(a_{j-1}:a_j]}(X_l-\theta_i)^2\right)d\theta_i, & i=a_{j-1}+1, j>1,\\ d\widehat{Q}_i^{(\theta)}(\theta_i|\theta_{i-1})=\delta_{\theta_{i-1}}(\theta_i)d\theta_i, & \text{otherwise.}\end{cases}$$

Now the only thing is to show that $\widehat{k}$ and $\widehat{a}_1,\cdots,\widehat{a}_{k-1}$ are the solution of (2.28). Plug $\widehat{Q}^{(w)}$

and $\widehat{Q}^{(\theta)}$ into (5.14), and then

$$\begin{aligned} & \int \widehat{q}^{(w)}(w) \log \frac{\widehat{q}^{(w)}(w)}{\pi(w) w^{|S|} (1-w)^{n-1-|S|}} dw \\ = \; & \log \frac{\Gamma(n-1+\alpha_0+\beta_0)}{\Gamma(k-1+\alpha_0)\Gamma(n-k+\beta_0)} - \log \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \end{aligned} \tag{5.15}$$

and

$$\begin{aligned} & \int_{\Theta(S)} q^{(\theta)}(\theta) \log \frac{q^{(\theta)}(\theta)}{g(\theta_1) \prod_{i \in S} g(\theta_i) \prod_{i \notin S} \delta_{\theta_{i-1}}(\theta_i) \exp\left(-\frac{1}{2}\sum_{i=1}^n (Y_i - \theta_i)^2\right)} d\theta \\ = \; & \int_{\Theta(S)} \widehat{q}_1^{(\theta)}(\theta_1) \prod_{i \in S} \widehat{q}_i^{(\theta)}(\theta_i | \theta_{i-1}) \prod_{i \notin S} \delta_{\theta_{i-1}}(\theta_i) \\ & \times \log \frac{\widehat{q}_1^{(\theta)}(\theta_1) \prod_{i \in S} \widehat{q}_i^{(\theta)}(\theta_i | \theta_{i-1}) \prod_{i \notin S} \delta_{\theta_{i-1}}(\theta_i)}{g(\theta_1) \prod_{i \in S} g(\theta_i) \prod_{i \notin S} \delta_{\theta_{i-1}}(\theta_i) \exp\left(-\frac{1}{2}\sum_{i=1}^n (Y_i - \theta_i)^2\right)} d\theta \\ = \; & \sum_{j=1}^k \log \left(\int g(\theta_{a_j+1}) \exp\left(-\frac{1}{2} \sum_{i=a_{j-1}+1}^{a_j} (Y_i - \theta_{a_{j-1}+1})^2 \right) d\theta_{a_{j-1}+1} \right) \\ = \; & \sum_{j=1}^k \log \left(\int g(\theta) \exp\left(-\frac{1}{2} \sum_{i=a_{j-1}+1}^{a_j} (Y_i - \theta)^2 \right) d\theta \right). \end{aligned} \tag{5.16}$$

Plug (5.15) and (5.16) into (5.14), and the optimization problem becomes (2.28). The proof is complete. □

5.1.6 Proofs of Theorem 2.4.2 and 2.4.3

To prove Theorem 2.4.2, we first establish an upper bound of $P_{\widetilde{f}_0}^n \widehat{Q} H^2(P_{\widehat{k},\theta^{(\widehat{k})}}, P_{\widetilde{f}_0})$ by applying Theorem 2.4.1 for a $\widetilde{f}_0$ that is constructed to be close to f_0. Then, with a change-of-measure argument, we derive a bound for $P_{f_0}^n \widehat{Q} H^2(P_{\widehat{k},\theta^{(\widehat{k})}}, P_{f_0})$. The construction of the surrogate density function $\widetilde{f}_0$ is given by the following lemma.

Lemma 5.1.13. *Suppose that the true density f_0 satisfies conditions (B1)-(B3). For a*

constant $H_1 > 2\alpha$, *we define* $\widetilde{f}_0(x) = \frac{f_0(x)\mathbb{I}_{E_{\sigma_0}}(x)}{\int_{E_{\sigma_0}} f_0(x)dx}$ *with* $E_{\sigma_0} = \{x : f_0(x) \geq \sigma_0^{H_1}\}$. *For a constant* $\xi_4 \leq \min\{\xi_3, p\}$ *and a sufficiently small* $\sigma_0 > 0$, *there exists a finite mixture* $p(x|k_{\sigma_0}, \theta_{\sigma_0})$ *with* $k_{\sigma_0} = O(\sigma_0^{-1}|\log\sigma_0|^{p/\xi_4})$ *and* $\theta_{\sigma_0} = (\mu_{\sigma_0}, w_{\sigma_0}, \sigma_0)$, *such that*

$$D_2\left(P_{\widetilde{f}_0} \| P_{k_{\sigma_0},\theta_{\sigma_0}}\right) = O(\sigma_0^{2\alpha}). \tag{5.17}$$

Moreover, (5.17) holds for all mixtures $p(x|k_{\sigma_0}, (\mu, w, \sigma))$ *such that* $\sigma \in [\sigma_0, \sigma_0 + \sigma_0^{H_1+2\alpha+2}]$, $\|\mu - \mu_{\sigma_0}\|_1 \leq \sigma_0^{H_1+2\alpha+2}$ *and* $w \in \Delta_{k_{\sigma_0}}(w_{\sigma_0}, \sigma_0^{H_1+2\alpha+1})$.

With the definition of $\widetilde{f}_0$ and its property given by Lemma 5.1.13, we can derive an upper bound for $P^n_{\widetilde{f}_0}\widehat{Q}H^2(P_{\widehat{k},\theta^{(\widehat{k})}}, P_{\widetilde{f}_0})$ by checking the conditions (C1), (C2) and (C3*) in Theorem 2.4.1. This argument is split into the next two lemmas.

Lemma 5.1.14. *For the prior* Π *that satisfies conditions (2.40), (2.42) and (2.45), the conditions (C1) and (C2) hold for* $L(P^{(n)}, P_0^{(n)}) = nH^2(P, P_0)$ *and all* $\epsilon > n^{\delta}$ *with some constant* $\delta > -1/2$ *with respect to* $P_0^{(n)} = P^n_{\widetilde{f}_0}$ *for any* $\sigma_0 \to 0$ *and* $P^{(n)} = P^n_{k,\theta^{(k)}}$.

Lemma 5.1.15. *Suppose that the true density* f_0 *satisfies conditions (B1)-(B3), and the prior* Π *satisfies conditions (2.41), (2.43), (2.44) and (2.46). Then the condition (C3*) holds for Theorem 2.4.2 with respect to* $P_0^{(n)} = P^n_{\widetilde{f}_0}$. *Here, the density* $\widetilde{f}_0$ *is defined in Lemma 5.1.13 with* σ_0 *chosen as* $n^{-\frac{1}{2\alpha+1}}(\log n)^{\frac{r}{2\alpha+1}}$ *and the rate is* $\epsilon_n = n^{-\frac{\alpha}{2\alpha+1}}(\log n)^{\frac{\alpha r}{2\alpha+1}}$ *with* r *given in Theorem 2.4.2.*

We first prove Lemma 5.1.13, and then prove Lemma 5.1.14 and Lemma 5.1.15. To facilitate the proof of Lemma 5.1.13, we introduce the following lemma, which is analogous to Theorem 1 in in [42].

Lemma 5.1.16. *Let* f_0 *be a density satisfying conditions (B1)-(B3), and let* K_{σ_0} *denote the convolution operator induced by the kernel* ψ_{σ_0}. *Then there exists a density* h_α *such that for a small enough* $\sigma_0 > 0$,

$$\int \frac{f_0^2}{K_{\sigma_0}h_\alpha} = 1 + O(\sigma_0^{2\alpha}).$$

Proof. We set $G_{\sigma_0} = \{x : f_0(x) \geq \sigma_0^{H_0}\}$ and

$$A_{\sigma_0} = \{x : |l_j(x)| \leq B\sigma_0^{-j}|\log\sigma_0|^{-j/p}, j-1, \cdots, \lfloor\alpha\rfloor, |L(x)| \leq B\sigma_0^{-\alpha}|\log\sigma_0|^{-\alpha/p}\}.$$

This is the same definition that appears in Lemma 1 of [42]. Note that $\int \frac{f_0(x)^2}{K_{\sigma_0}h_\alpha(x)}dx - 1 \geq 0$, and we only need to derive an upper bound for this integral. We first have the following decomposition

$$\int \frac{f_0(x)^2}{K_{\sigma_0}h_\alpha(x)}dx = \int_{A_{\sigma_0}\cap G_{\sigma_0}} \frac{(f_0(x) - K_{\sigma_0}h_\alpha(x))^2}{K_{\sigma_0}h_\alpha(x)}dx$$
$$+ \int_{A^c_{\sigma_0}\cup G^c_{\sigma_0}} \frac{f_0(x)^2}{K_{\sigma_0}h_\alpha(x)}dx + \int_{A^c_{\sigma_0}\cup G^c_{\sigma_0}} (K_{\sigma_0}h_\alpha(x) - f_0(x))dx + \int_{A_{\sigma_0}\cap G_{\sigma_0}} f_0(x)dx.$$

The first and third terms can be bounded by $O(\sigma_0^{2\alpha})$ according to the same argument in the proof of Theorem 1 in [42] when H_0 is chosen to be large enough. For the second term, according to Remark 1 in [42], we have $\frac{f_0(x)}{K_{\sigma_0}h_\alpha(x)} \leq M_0$ with some constant $M_0 > 0$ for all x. Then Lemma 2 in [42] implies

$$\int_{A^c_{\sigma_0}\cup G^c_{\sigma_0}} \frac{f_0(x)^2}{K_{\sigma_0}h_\alpha(x)}dx \leq M_0 \int_{A^c_{\sigma_0}\cup G^c_{\sigma_0}} f_0(x)dx = O(\sigma_0^{2\alpha}).$$

The last term can be upper bounded by 1. Summing up all the terms, we obtain the desired conclusion. □

Proof of Lemma 5.1.13. The proof uses a slightly modified argument in the proof of Lemma 4 in [42]. First of all, according to Lemma 5.1.16, there exists a density h_α such that $\int \frac{f_0(x)^2}{K_{\sigma_0}h_\alpha(x)}dx = 1 + O(\sigma_0^{2\alpha})$. Define $E'_{\sigma_0} = \{x : f_0(x) \geq \sigma_0^{H_2}\}$, where $H_2 > H_1$ is chosen to be large enough. Set $\widetilde{h}_\alpha(x) = \frac{h_\alpha(x)\mathbb{I}_{E'_{\sigma_0}}(x)}{\int_{E'_{\sigma_0}} h_\alpha(x)dx}$. Define the number $a_{\sigma_0} = C_0|\log\sigma_0|^{1/\xi_4}$, with $\xi_4 \leq \min\{\xi_3, p\}$ and some constant $C_0 > 0$. We choose $p(x|k_{\sigma_0}, \theta_{\sigma_0})$ with $\theta_{\sigma_0} =$

$(\mu_{\sigma_0}, w_{\sigma_0}, \sigma_0)$ to be the finite mixture given by Lemma 12 in [42] that satisfies

$$\|K_{\sigma_0}\widetilde{h}_\alpha - p_{k_{\sigma_0},\theta_{\sigma_0}}\|_\infty \leq \sigma_0^{-1}\exp(-C_0|\log\sigma_0|^{p/\xi_4}),$$

for $x \in [-a_{\sigma_0}, a_{\sigma_0}]$, where $p_{k_{\sigma_0},\theta_{\sigma_0}}$ is the density of $P_{k_{\sigma_0},\theta_{\sigma_0}}$. We will show that this mixture density satisfies (5.17). We write

$$D_2(P_{\widetilde{f}_0}\|P_{k_{\sigma_0},\theta_{\sigma_0}}) = \int \frac{\widetilde{f}_0(x)^2}{p_{k_{\sigma_0},\theta_{\sigma_0}}(x)}dx = \int_{E_{\sigma_0}} \frac{\widetilde{f}_0^2}{f_0^2}\frac{f_0^2}{K_{\sigma_0}h_\alpha}\frac{K_{\sigma_0}h_\alpha}{K_{\sigma_0}\widetilde{h}_\alpha}\frac{K_{\sigma_0}\widetilde{h}_\alpha}{p_{k_{\sigma_0},\theta_{\sigma_0}}}.$$

The four ratios will be bounded separately.

1. According to (B2), we know that $\int f_0(x)^b dx = O(1)$, for any constant $b > 0$. Since $H_1 > 2\alpha$,

   $$\int_{E_{\sigma_0}^c} f_0(x)dx \leq (\sigma_0^{H_1})^{\frac{2\alpha}{H_1}} \int_{E_{\sigma_0}^c} f_0(x)^{1-\frac{2\alpha}{H_1}}dx = O(\sigma_0^{2\alpha}).$$

   This leads to

   $$\left|\frac{\widetilde{f}_0^2(x)}{f_0^2(x)} - 1\right| = \left|\frac{1}{(1-\int_{E_{\sigma_0}^c} f_0(x)dx)^2} - 1\right| \leq C_1\sigma_0^{2\alpha},$$

   for a constant $C_1 > 0$ and all $x \in E_{\sigma_0}$.

2. For the second term, we have

   $$\int_{E_{\sigma_0}} \frac{f_0^2}{K_{\sigma_0}h_\alpha}dx = \int \frac{f_0^2}{K_{\sigma_0}h_\alpha}dx - \int_{E_{\sigma_0}^c} \frac{f_0^2}{K_{\sigma_0}h_\alpha}dx.$$

   Since $\frac{f_0(x)}{K_{\sigma_0}h_\alpha(x)} \leq M_0$ for a constant M_0 uniformly over x,

   $$\int_{E_{\sigma_0}^c} \frac{f_0^2}{K_{\sigma_0}h_\alpha}dx \leq M_0 \int_{E_{\sigma_0}^c} f_0(x)dx = O(\sigma_0^{2\alpha}).$$

Combining with Lemma 5.1.16, we conclude that

$$\left| \int_{E_{\sigma_0}} \frac{f_0^2}{K_{\sigma_0} h_\alpha} dx - 1 \right| \leq C_2 \sigma_0^{2\alpha},$$

for a constant $C_2 > 0$.

3. By the same argument in the proof of Lemma 4 in [42], we get

$$\left| \frac{K_{\sigma_0} h_\alpha(x)}{K_{\sigma_0} \widetilde{h}_\alpha(x)} - 1 \right| \leq C_3 \sigma_0^{2\alpha},$$

for a constant $C_3 > 0$ and all $x \in E_{\sigma_0}$.

4. According to the proof of Lemma 4 in [42], we have $E'_{\sigma_0} \subset \{x : f_0(x) \geq c_0 \sigma_0^{H_2}\}$ for some constant c_0. Because $\xi_4 \leq \xi_3$, $E'_{\sigma_0} \subset [-a_{\sigma_0}, a_{\sigma_0}]$. This leads to the inequality $\|K_{\sigma_0} \widetilde{h}_\alpha - p_{k_{\sigma_0}, \theta_{\sigma_0}}\|_\infty \leq \sigma_0^{-1} \exp(-C_0 |\log \sigma_0|^{p/\xi_4})$. Note that for any $x \in E_{\sigma_0}$, we have $K_{\sigma_0}(x) \widetilde{h}_\alpha(x) \gtrsim K_{\sigma_0} h_\alpha(x) \gtrsim f_0(x) \gtrsim \sigma_0^{H_1}$ uniformly over $x \in E_{\sigma_0}$. Thus, for a sufficiently large C_0,

$$\sigma_0^{-1} \exp(-C_0 |\log \sigma_0|^{p/\xi_4}) = \sigma^{C_0 |\log \sigma_0|^{(p-\xi_4)/\xi_4} - 1} = O(\sigma_0^{H_1 + 2\alpha}),$$

where we have used the condition $\xi_4 \leq p$. Then we have

$$\left| \frac{K_{\sigma_0} \widetilde{h}_\alpha(x)}{p(x | k_{\sigma_0}, \theta_{\sigma_0})} - 1 \right| \leq \frac{\|K_{\sigma_0} \widetilde{h}_\alpha - p_{k_{\sigma_0}, \theta_{\sigma_0}}\|_\infty}{K_{\sigma_0} \widetilde{h}_\alpha(x) - \|K_{\sigma_0} \widetilde{h}_\alpha - p_{k_{\sigma_0}, \theta_{\sigma_0}}\|_\infty} \leq C_4 \sigma_0^{2\alpha}.$$

for all $x \in E_{\sigma_0}$ with some constant $C_4 > 0$.

Combining the bounds of all terms above, we get

$$\int \frac{\widetilde{f}_0^2(x)}{p(x | k_{\sigma_0}, \theta_{\sigma_0})} dx = 1 + O(\sigma_0^{2\alpha}),$$

which indicates that (5.17) holds. When $\sigma \in [\sigma_0, \sigma_0^{H_1 + 2\alpha + 2}]$, $\|\mu - \mu_{\sigma_0}\|_1 \leq \sigma_0^{H_1 + 2\alpha + 2}$ and

$w \in \Delta_{k_{\sigma_0}}(w_{\sigma_0}, \sigma_0^{H_1+2\alpha+1})$, according to Lemma 3 in [42], we have

$$\|p_{k_{\sigma_0},(\mu,w,\sigma)} - p_{k_{\sigma_0},(\mu_{\sigma_0},w_{\sigma_0},\sigma_0)}\|_\infty = O(\sigma_0^{H_1+2\alpha}).$$

Then the four points listed above also hold, which means that (5.17) is also satisfied for these $(k_{\sigma_0}, (\mu, w, \sigma))$. The proof is complete. □

Now we prove Lemma 5.1.14 and Lemma 5.1.15.

Proof of Lemma 5.1.14. We consider the set

$$\Theta_n(\epsilon) = \cup_{k=1}^{k_n} \Theta^{(k)}(\epsilon), \tag{5.18}$$

where

$$\Theta^{(k)}(\epsilon) = \left\{ P_{k,\theta^{(k)}} : \theta^{(k)} = (\mu, w, \sigma), \mu \in \otimes_{j=1}^{k}[-b_n, b_n], \sigma \in (m_\sigma, M_\sigma] \right\},$$

with $k_n = \left\lceil \frac{n\epsilon^2}{\log(n\epsilon^2)} \right\rceil$, $b_n = (n\epsilon^2)^{\frac{1}{c_3}}$, $m_\sigma = (n\epsilon^2)^{-\frac{1}{2b_3}}$ and $M_\sigma = \exp\left(\frac{1}{2}n\epsilon^2\right)$. It's easy to see that

$$\Theta_n(\epsilon)^c \subseteq \left\{ P_{k,\theta^{(k)}} : k > k_n \right\} \bigcup \left(\cup_{k=1}^{k_n} \widetilde{\Theta}^{(k)}(\epsilon) \right),$$

where $\widetilde{\Theta}^{(k)}(\epsilon) = \left\{ P_{k,\theta^{(k)}} : \theta^{(k)} = (\mu, w, \sigma), \max_j |\mu_j| > B_n \text{ or } \sigma \notin (m_\sigma, M_\sigma] \right\}$. Thus

$$\Pi(\Theta_n(\epsilon)^c) \leq \sum_{k>k_n} \pi(k) + \sum_{k=1}^{k_n} \pi(k) \left[\Pi^{(k)} \left(\sigma \notin (m_\sigma, M_\sigma] \right) + \Pi^{(k)} \left(\max_{1\leq j\leq k} |\mu_j| > b_n \right) \right]$$

Now we derive an upper bound for each term.

1. Set $\tau = \sigma^{-2}$, and then for all $k \in \mathbb{N}_+$,

$$\begin{aligned} \Pi^{(k)}(\sigma \notin (m_\sigma, M_\sigma]) &\leq \int_0^{\exp(-n\epsilon^2)} p_\tau(\tau) d\tau + \int_{(n\epsilon^2)^{1/b_3}}^{\infty} p_\tau(\tau) d\tau \\ &\leq b_0 \exp(-n\epsilon^2) + b_1 \exp(-b_2 n\epsilon^2), \end{aligned}$$

where we have used the condition (2.45).

2. By the condition (2.40), we have

$$\sum_{k>k_n} \pi(k) \leq C_1 \exp(-C_2 k_n \log(k_n)) \leq C_1 \exp(-\widetilde{C}_2 n\epsilon^2).$$

3. According to the conditions (2.40) and (2.42),

$$\begin{aligned}
& \sum_{k=1}^{k_n} \pi(k)\Pi^{(k)}(\max_j |\mu_j| > b_n) \leq \sum_{k=1}^{\infty} \pi(k)\Pi^{(k)}(\max_j |\mu_j| > b_n) \\
\leq & \sum_{k=1}^{\infty} \pi(k) k \left(\int_{-\infty}^{-b_n} p_\mu(x)dx + \int_{b_n}^{\infty} p_\mu(x)dx \right) \\
\leq & c_1 \exp(-c_2 n\epsilon^2) \sum_{m=1}^{\infty} \sum_{k=m}^{\infty} \pi(k) \\
\leq & c_1 \exp(-c_2 n\epsilon^2) \sum_{m=1}^{\infty} C_1 \exp(-C_2 m \log m) \\
\leq & \widetilde{c}_1 \exp(-c_2 n\epsilon^2).
\end{aligned}$$

Summing up the three bounds above, we have $\Pi(\Theta_n(\epsilon)^c) \lesssim \exp(-C_0 n\epsilon^2)$ for some constant $C_0 > 0$. In order that the constant C_0 can be arbitrarily large, one can replace ϵ by $\widetilde{C}\epsilon$ for a sufficiently large $\widetilde{C}$ and use the same argument above. We therefore obtain (C2).

Now we start to show (C1). By Theorem 7.1 in [31], it is sufficient to bound the metric entropy

$$\log N(\epsilon, \Theta_n(\epsilon), H) \lesssim n\epsilon^2.$$

Since $H^2(P_1, P_2) \leq \mathsf{TV}(P_1, P_2)$, we have $N(\epsilon, \Theta_n(\epsilon), H) \leq N(\epsilon^2, \Theta_n(\epsilon), \mathsf{TV})$. According to (5.18),

$$N(\epsilon^2, \Theta_n(\epsilon), \mathsf{TV}) \leq \sum_{k=1}^{k_n} N(\epsilon^2, \Theta^{(k)}(\epsilon), \mathsf{TV}),$$

and thus it is sufficient to bound $N(\epsilon^2, \Theta^{(k)}(\epsilon), \mathsf{TV})$ for each $k \in [k_n]$.

We use ψ to denote ψ_σ with $\sigma = 1$ in short. According to Lemma 3 in [42], for any $P_{k,\theta}$ with $\theta = (\mu, w, \sigma)$ and $P_{k,\widetilde{\theta}}$ with $\widetilde{\theta} = (\widetilde{\mu}, \widetilde{w}, \widetilde{\sigma})$ such that $P_{k,\theta}, P_{k,\widetilde{\theta}} \in \Theta^{(k)}(\epsilon)$, we have

$$\begin{aligned}
\mathsf{TV}(P_{k,\theta}, P_{k,\widetilde{\theta}}) &\leq \|w - \widetilde{w}\|_1 + 2\|\psi\|_\infty \sum_{i=1}^{k} \frac{w_i \wedge \widetilde{w}_i}{\sigma \wedge \widetilde{\sigma}} |\mu_i - \widetilde{\mu}_i| + \frac{|\sigma - \widetilde{\sigma}|}{\sigma \wedge \widetilde{\sigma}} \\
&\leq \|w - \widetilde{w}\|_1 + 2\frac{\|\psi\|_\infty}{m_\sigma} \|\mu - \widetilde{\mu}\|_1 + \frac{|\sigma - \widetilde{\sigma}|}{m_\sigma}.
\end{aligned}$$

Based on the fact that $N(\epsilon, A \times B, d_1 + d_2) \leq N(t\epsilon, A, d_1) \times N((1-t)\epsilon, B, d_2)$, we have

$$\begin{aligned}
& N(\epsilon^2, \Theta^{(k)}(\epsilon), \mathsf{TV}) \\
\leq\ & N\left(\frac{\epsilon^2}{3}, \Delta_k, \|\cdot\|_1\right) N\left(\frac{m_\sigma \epsilon^2}{6\|\psi\|_\infty}, [-b_n, b_n]^k, \|\cdot\|_1\right) N\left(\frac{m_\sigma \epsilon^2}{3}, (m_\sigma, M_\sigma], |\cdot|\right).
\end{aligned}$$

Then, we use Lemma 5 in [42], and obtain

$$N\left(\frac{\epsilon^2}{3}, \Delta_k, \|\cdot\|_1\right) \leq \exp\left((k-1)\log\frac{15}{\epsilon^2}\right) \leq \exp(C_1 k \log(n\epsilon^2)),$$

$$N\left(\frac{m_\sigma \epsilon^2}{6\|\psi\|_\infty}, [-b_n, b_n]^k, \|\cdot\|_1\right) \leq \frac{k!(b_n + \widetilde{\epsilon})^k}{\widetilde{\epsilon}^k} \leq \exp(C_2 k(\log k + \log n\epsilon^2)),$$

where $\widetilde{\epsilon} = \frac{m_\sigma \epsilon^2}{6\|\psi\|_\infty}$, and

$$N\left(\frac{m_\sigma \epsilon^2}{3}, (m_\sigma, M_\sigma], |\cdot|\right) \leq \frac{M_\sigma}{m_\sigma \epsilon^2/3} \leq \exp(C_3 n\epsilon^2),$$

for some constants $C_1, C_2, C_3 > 0$. Note that we have used the condition $\epsilon > n^\delta$ for some constant $\delta > -1/2$ to derive the above bounds. Finally, we have

$$N(\epsilon^2, \Theta_n(\epsilon), \mathsf{TV}) \leq k_n \exp\left(C_1 k_n \log n\epsilon^2 + C_2 k_n(\log k_n + \log n\epsilon^2) + C_3 n\epsilon^2\right),$$

which leads to

$$\log N(\epsilon^2, \Theta_n(\epsilon), \mathsf{TV}) \lesssim k_n \log(n\epsilon^2) \lesssim n\epsilon^2.$$

The proof is complete. □

Proof of Lemma 5.1.15. According to Lemma 5.1.13, there exist k_{σ_0}, $\theta_{\sigma_0} = (\mu_{\sigma_0}, w_{\sigma_0}, \sigma_0)$ such that (5.17) holds. Then we set $k_0 = k_{\sigma_0}$ in Theorem 2.4.1 and $\Theta^{(k_{\sigma_0})} = \Theta_{\mu}^{(k_{\sigma_0})} \otimes \Theta_{w}^{(k_{\sigma_0})} \otimes \Theta_{\sigma}^{(k_{\sigma_0})}$, where $\Theta_{\mu}^{(k_{\sigma_0})} = \otimes_{j=1}^{k_{\sigma_0}} \Theta_{\mu_j}^{(k_{\sigma_0})}$. To be specific, let H_1 be any fixed constant such that $H_1 > 2\alpha$, and then we define

$$\Theta_{\mu_j}^{(k_{\sigma_0})} = [\mu_{\sigma_0,j} - k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}, \mu_{\sigma_0,j} + k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}],$$

$$\Theta_{w}^{(k_{\sigma_0})} = \Delta_{k_{\sigma_0}}(w_{\sigma_0}, \sigma_0^{H_1+2\alpha+1}),$$

and

$$\Theta_{\sigma}^{(k_{\sigma_0})} = [\sigma_0, \sigma_0 + \sigma_0^{H_1+2\alpha+2}].$$

The conclusion of Lemma 5.1.13 implies

$$\Theta^{(k_{\sigma_0})} \subset \left\{(\mu, w, \sigma) : nD_2\left(P_{\widetilde{f}_0} \| P_{k_{\sigma_0},\theta_{\sigma_0}}\right) \leq C_2 n\sigma_0^{2\alpha}\right\}, \tag{5.19}$$

for a constant $C_2 > 0$. Choose $\sigma_0 = n^{-\frac{1}{2\alpha+1}}(\log n)^{\frac{r}{2\alpha+1}}$, then $n^{\frac{1}{2\alpha+1}} \leq k_{\sigma_0} \leq n^{\frac{1}{2\alpha+1}+t}$ for any $t > 0$ as $n \to \infty$. Then the condition (2.41) implies

$$-\log \pi(k_{\sigma_0}) \lesssim k_{\sigma_0} \log k_{\sigma_0}. \tag{5.20}$$

We also have

$$\Pi_{\mu_j}^{(k_{\sigma_0})}(\Theta_{\mu_j}^{(k_{\sigma_0})}) \geq \int_{\mu_{\sigma_0,j}-k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}}^{\mu_{\sigma_0,j}+k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}} p_\mu(x)dx.$$

According to the condition (2.43) and Lemma 5.1.13, we have $|\mu_j| \lesssim |\log \sigma_0|^{1/\xi_4}$ with $\xi_4 \leq \min\{\xi_3, p\}$ as in Lemma 5.1.13. Then,

$$-\log \Pi_{\mu_j}^{(k_{\sigma_0})}(\Theta_{\mu_j}^{(k_{\sigma_0})}) \lesssim |\log \sigma_0| + |\log \sigma_0|^{c_6/\xi_4} \lesssim |\log \sigma_0|^{\max\{1, c_6/\xi_4\}}.$$

By (2.44), we have

$$-\log \Pi_w^{(k_{\sigma_0})}(\Theta_w^{(k_{\sigma_0})}) \lesssim k_{\sigma_0}(\log k_{\sigma_0})^{d_3}|\log \sigma_0|.$$

Finally, the condition (2.46) leads to

$$-\log \Pi_\sigma^{(k_{\sigma_0})}(\Theta_\sigma^{(k_{\sigma_0})}) \leq -\log\left(\int_{(\sigma_0+\sigma_0^{H_1+2\alpha+2})^{-1}}^{\sigma_0^{-1}} p_\tau(x)dx\right) \lesssim |\log \sigma_0| + \sigma_0^{-b_6} \lesssim \sigma_0^{-1}.$$

With the choice $\xi_4 = \min\{p, \xi_3\}$ and $k_{\sigma_0} = O(\sigma_0^{-1}|\log \sigma_0|^{p/\xi_4})$, we have

$$-\log \pi(k_{\sigma_0}) - \sum_{j=1}^{k_{\sigma_0}} \Pi_{\mu_j}(\Theta_{\mu_j}^{(k_{\sigma_0})}) - \log \Pi_w^{(k_{\sigma_0})}(\Theta_w^{(k_{\sigma_0})}) - \log \Pi_\sigma(\Theta_\sigma^{(k_{\sigma_0})}) \leq C_3\sigma_0^{-1}(\log \sigma_0)^r.$$

where $r = \frac{p}{\min\{p,\xi_3\}} + \max\{d_3+1, \frac{c_6}{\min\{p,\xi_3\}}\}$. Plug in $\sigma_0 = n^{-\frac{1}{2\alpha+1}}(\log n)^{\frac{r}{2\alpha+1}}$,we obtain (C3*) with respect to $\widetilde{f}_0$. □

Finally we prove Theorem 2.4.2.

Proof of Theorem 2.4.2. We bound $P_{f_0}^n \widehat{Q} H^2(P_{\widehat{k},\theta^{(\widehat{k})}}, P_{f_0})$ by

$$\begin{aligned} P_{f_0}^n \widehat{Q} H^2(P_{\widehat{k},\theta^{(\widehat{k})}}, P_{f_0}) &\leq P_{\widetilde{f}_0}^n \widehat{Q} H^2(P_{\widehat{k},\theta^{(\widehat{k})}}, P_{f_0}) + \mathsf{TV}(P_{f_0}^n, P_{\widetilde{f}_0}^n) \\ &\leq 2P_{\widetilde{f}_0}^n \widehat{Q} H^2(P_{\widehat{k},\theta^{(\widehat{k})}}, P_{\widetilde{f}_0}) + 2H^2(P_{\widetilde{f}_0}, P_{f_0}) + \mathsf{TV}(P_{f_0}^n, P_{\widetilde{f}_0}^n). \end{aligned}$$

By Lemma 5.1.14, Lemma 5.1.15 and Theorem 2.4.1, we have

$$P_{\widetilde{f}_0}^n \widehat{Q} H^2(P_{\widehat{k},\theta^{(\widehat{k})}}, P_{\widetilde{f}_0}) \lesssim \sigma_0^{2\alpha},$$

for $\sigma_0 = n^{-\frac{1}{2\alpha+1}}(\log n)^{\frac{r}{2\alpha+1}}$. Note that $\widetilde{f}_0(x) = \frac{f_0(x)\mathbb{I}_{E_{\sigma_0}}(x)}{\int_{E_{\sigma_0}} f_0(x)dx}$ with $E_{\sigma_0} = \{x : f_0(x) \geq \sigma_0^{H_1}\}$,

and $R = \int_{E_{\sigma_0}^c} f_0(x)dx \leq \sigma_0^{H_1/2} \int_{E_{\sigma_0}^c} \sqrt{f_0(x)}dx = O(\sigma_0^{H_1/2})$. Then,

$$H^2(P_{\widetilde{f}_0}, P_{f_0}) = 1 - \int \sqrt{f_0(x)\widetilde{f}_0(x)}dx = 1 - \sqrt{1-R} = O(\sigma_0^{H_1/2}).$$

Moreover,

$$\mathsf{TV}(P_{f_0}^n, P_{\widetilde{f}_0}^n) = 1 - (1-R)^n = O(nR) = O(n\sigma_0^{H_1/2}).$$

With the choice $H_1 = 8\alpha + 4$, the proof is complete. □

Now we prove Theorem 2.4.3. We also use change of measure argument and show the concentration around $P_{\widetilde{f}_0}^{(n)}$ at first.

Proof of Theorem 2.4.3. We first present a latent variable version of Lemma 5.2.2,

$$\begin{aligned}
& P_{\widetilde{f}_0}^n \left[\widehat{Q}^{(\widehat{k})} n H \left(P_{\widehat{k},\theta^{(\widehat{k})}}, P_{\widetilde{f}_0} \right)^2 \right] \\
\leq & \inf_{a>0} \frac{1}{a} \left[\min_{k \in \mathcal{K}} \min_{\bar{Q}^{(k)} \in \bar{\mathcal{S}}_{\mathrm{MF}}^{(k)}} \left\{ D\left(\bar{Q}^{(k)} \| \bar{\Pi}^{(k)} \right) \right. \right. \\
& + \bar{Q}^{(k)} D\left(P_{\widetilde{f}_0}^n \| P(\cdot|k, z^{(k)}, \theta^{(k)}) \right) - \log \pi(k) \Big\} \\
& \left. + P_{\widetilde{f}_0}^n \log \Pi \left(\exp\left(anH \left(P_{k,\theta^{(k)}}, P_{\widetilde{f}_0} \right)^2 \right) \Big| X^{(n)} \right) \right], \qquad (5.21)
\end{aligned}$$

where $\widetilde{f}_0$ is defined in the Lemma 5.1.14. The proof of this inequality follows the same argument as in the proof of Lemma 5.2.2 and thus we omit it. Note that the parametrization of the density $p(x|k, \theta^{(k)})$ in (2.36) does not rely on the latent variables. Therefore, when the conditions of Theorem 2.4.2 are satisfied, for some small constant $a > 0$, we have

$$P_{\widetilde{f}_0}^n \log \Pi \left(\exp\left(anH \left(P_{k,\theta^{(k)}}, P_{\widetilde{f}_0} \right)^2 \right) \Big| X^{(n)} \right) \lesssim n^{\frac{1}{2\alpha+1}} (\log n)^{\frac{2\alpha r}{2\alpha+1}},$$

based on the Jensen's Inequality, Lemma 5.1.3 and Lemma 5.1.4.

Now we need to choose some $k \in \mathcal{K}$ and $\bar{Q}^{(k)} \in \mathcal{S}_{\mathrm{MF}}^{(k)}$ to bound the remaining terms of

(5.21). We consider $d\bar{Q}^{(k)}(z^{(k)},\theta^{(k)}) = dQ_z^{(k)}(z^{(k)})dQ_\theta^{(k)}(\theta^{(k)})$ and $Q_z^{(k)}(z_i^{(k)}=j)=\gamma_{ij}$, where $\sum_{j=1}^k \gamma_{ij}=1$. We sometimes shorthand $Q_\theta^{(k)}$ by $Q^{(k)}$ when the context is clear. Write $z_{ij}^{(k)} = \mathbb{I}\{z_i^{(k)}=j\}$, and we have

$$
\begin{aligned}
& D\left(\bar{Q}^{(k)}\|\bar{\Pi}^{(k)}\right) + \bar{Q}^{(k)} D\left(P_{\widetilde{f}_0}^n \| P(\cdot|k,z^{(k)},\theta^{(k)})\right) - \log\pi(k) \\
= \;& P_{\widetilde{f}_0}^n \sum_{z^{(k)}} \prod_{i=1}^n \prod_{j=1}^k \gamma_{ij}^{z_{ij}^{(k)}} \int \log \frac{\prod_{i=1}^n \prod_{j=1}^k \gamma_{ij}^{z_{ij}^{(k)}} dQ_\theta^{(k)}(\theta^{(k)})}{d\Pi^{(k)}(\theta^{(k)}) \prod_{i=1}^n \prod_{j=1}^k \left(w_j \psi_\sigma(X_i-\mu_j)\right)^{z_{ij}^{(k)}}} dQ_\theta^{(k)}(\theta^{(k)}) \\
& + P_{\widetilde{f}_0}^n \log p_{\widetilde{f}_0}^n(X^{(n)}) - \log\pi(k) \\
= \;& \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij}\log\gamma_{ij} + D\left(Q^{(k)}\|\Pi^{(k)}\right) + P_{\widetilde{f}_0}^n \log p_{\widetilde{f}_0}^n(X^{(n)}) - \log\pi(k) \\
& - \sum_{z^{(k)}} \sum_{i,j} z_{ij}^{(k)} \prod_{i=1}^n \prod_{j=1}^k \gamma_{ij}^{z_{ij}^{(k)}} \int \log w_j \psi_\sigma(X_i-\mu_j) dQ^{(k)}(\theta^{(k)}) \\
= \;& \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij} \log \frac{\gamma_{ij}}{\exp\left(\int \log w_j \psi_\sigma(X_i-\mu_j) dQ^{(k)}(\theta^{(k)}\right))} \\
& + D\left(Q^{(k)}\|\Pi^{(k)}\right) + P_{\widetilde{f}_0}^n \log p_{\widetilde{f}_0}^n(X^{(n)}) - \log\pi(k).
\end{aligned}
$$

Thus, the optimal choice of γ_{ij} is that

$$
\gamma_{ij} = \frac{\exp\left(\int \log w_j \psi_\sigma(X_i-\mu_j) dQ^{(k)}(\theta^{(k)})\right)}{\sum_{r=1}^k \exp\left(\int \log w_r \psi_\sigma(X_i-\mu_r) dQ^{(k)}(\theta^{(k)})\right)},
$$

and we fix this choice as our $Q_z^{(k)}$. We then have

$$
\begin{aligned}
& D\left(\bar{Q}^{(k)}\|\bar{\Pi}^{(k)}\right) + \bar{Q}^{(k)} D\left(P_{\widetilde{f}_0}^n \| P(\cdot|k,z^{(k)},\theta^{(k)})\right) - \log\pi(k) \qquad (5.22) \\
= \;& -\sum_{i=1}^n \log\left[\sum_{r=1}^k \exp\left(\int \log w_r \psi_\sigma(X_i-\mu_r) dQ^{(k)}(\theta^{(k)})\right)\right] \\
& + D\left(Q^{(k)}\|\Pi^{(k)}\right) + P_{\widetilde{f}_0}^n \log p_{\widetilde{f}_0}^n(X^{(n)}) - \log\pi(k).
\end{aligned}
$$

We now specify the choice of $k \in \mathcal{K}$ and $Q^{(k)} = Q_\theta^{(k)}$ in (5.22). According to Lemma 5.1.13, for $k = k_{\sigma_0} = O(\sigma_0^{-1}|\log\sigma_0|^{p/\xi_4})$, when $\theta^{(k)} = (\mu, w, \sigma)$ such that

$$\sigma \in [\sigma_0, \sigma_0 + \sigma_0^{H_1+2\alpha+2}], \qquad \|\mu - \mu_{\sigma_0}\|_1 \le \sigma_0^{H_1+2\alpha+2}, \qquad w \in \Delta_{k_{\sigma_0}}(w_{\sigma_0}, \sigma_0^{H_1+2\alpha+1}),$$

we have

$$D_2\left(P_{\widetilde{f}_0} \| P_{k,\theta^{(k)}}\right) \lesssim \sigma_0^{2\alpha}. \tag{5.23}$$

Suppose $i_0 = \text{argmax}_i\, w_{\sigma_0,i}$, then $w_{\sigma_0,i_0} \ge k_{\sigma_0}^{-1}$ and $w_{\sigma_0,i_0} - \frac{k_{\sigma_0}-1}{2k_{\sigma_0}}\sigma_0^{H_1+2\alpha+1} > \frac{1}{2k_{\sigma_0}}\sigma_0^{H_1+2\alpha+1}$ when $\sigma_0 \to 0$. Then consider

$$w^*_{\sigma_0,j} = w_{\sigma_0,j} + \mathbb{I}\{j \ne i_0\}\frac{1}{2k_{\sigma_0}}\sigma_0^{H_1+2\alpha+1} - \mathbb{I}\{j = i_0\}\frac{k_{\sigma_0}-1}{2k_{\sigma_0}}\sigma_0^{H_1+2\alpha+1}.$$

Obviously, $w^*_\sigma \in \Delta_{k_{\sigma_0}}(w_{\sigma_0}, \sigma_0^{H_1+2\alpha+1})$ and $w^*_{\sigma_0,i} \ge \frac{1}{2k_{\sigma_0}}\sigma_0^{H_1+2\alpha+1}$ for all $1 \le i \le k_{\sigma_0}$. Set

$$\widetilde{\Theta}_{\mu_j}^{(k_{\sigma_0})} = [\mu_{\sigma_0,j} - k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}, \mu_{\sigma_0,j} + k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}],$$

$$\widetilde{\Theta}_w^{(k_{\sigma_0})} = \Delta_{k_{\sigma_0}}\left(w^*_{\sigma_0}, \frac{\nu}{2k_{\sigma_0}}\sigma_0^{H_1+2\alpha+1}\right).$$

$$\widetilde{\Theta}_\sigma^{(k_{\sigma_0})} = \left[\widetilde{\sigma}_0, \widetilde{\sigma}_0 + \frac{1}{2}\widetilde{\sigma}_0^{H_1+2\alpha+2}\right],$$

where $\widetilde{\sigma}_0 = (1+\epsilon)\sigma_0$ with $\epsilon > 0$ and $0 < \nu < 1$ to be determined later. Choose $k = k_{\sigma_0}$ and $dQ^{(k_{\sigma_0})}(\theta^{(k_{\sigma_0})}) = dQ_w^{(k_{\sigma_0})}(w)dQ_\tau^{(k_{\sigma_0})}(\tau)\prod_{j=1}^{k_{\sigma_0}} dQ_{\mu_j}^{(k_{\sigma_0})}(\mu_j)$ in (5.22), where

$$dQ_w^{(k_{\sigma_0})}(w) = \frac{p_w^{(k_{\sigma_0})}(w)\mathbb{I}\{w \in \widetilde{\Theta}_w^{(k_{\sigma_0})}\}dw}{\int_{\widetilde{\Theta}_w^{(k_{\sigma_0})}} p_w^{(k_{\sigma_0})}(w)dw},$$

$$dQ_{\mu_j}(\mu_j) = \frac{p_\mu(\mu_j)\mathbb{I}\{\mu_j \in \widetilde{\Theta}_{\mu_j}^{(k_{\sigma_0})}\}d\mu_j}{\int_{\widetilde{\Theta}_{\mu_j}^{(k_{\sigma_0})}} p_\mu(\mu_j)d\mu_j}, \qquad \text{for } 1 \le j \le k_{\sigma_0},$$

$$dQ_\tau(\tau) = \frac{p_\tau(\tau)\mathbb{I}\{\tau^{-1/2} \in \widetilde{\Theta}_\sigma^{(k_{\sigma_0})}\}d\tau}{\int_{\tau^{-1/2}\in\widetilde{\Theta}_\sigma^{(k_{\sigma_0})}} p_\tau(\tau)d\tau}.$$

Now, we build an upper bound for

$$-\sum_{i=1}^{n}\log\left[\sum_{r=1}^{k_{\sigma_0}}\exp\left(\int \log w_r\psi_\sigma(X_i-\mu_r)dQ^{(k_{\sigma_0})}(\theta^{(k_{\sigma_0})})\right)\right], \tag{5.24}$$

or equivalently, construct lower bounds for $\int \log w_r\psi_\sigma(X_i-\mu_r)dQ^{(k_{\sigma_0})}(\theta^{(k_{\sigma_0})})$ for all $1 \le i \le n$ and $1 \le r \le k_{\sigma_0}$. Set $\tau_{\min}^{1/2} = (\widetilde{\sigma}_0 + \frac{1}{2}\widetilde{\sigma}_0^{H_1+2\alpha+2})^{-1}$ and $\tau_{\max}^{1/2} = \widetilde{\sigma}_0^{-1}$, and then for any $w \in \widetilde{\Theta}_w^{(k_{\sigma_0})}$, $\mu_r \in \widetilde{\Theta}_{\mu_r}^{(k_{\sigma_0})}$, $\tau^{-1/2} = \sigma \in \widetilde{\Theta}_{\sigma_0}^{(k_{\sigma_0})}$, we have

$$w_r \ge w_{\sigma_0,r}^* - \frac{\nu}{2k_{\sigma_0}}\sigma_0^{H_1+2\alpha+1} \ge (1-\nu)w_{\sigma_0,r}^*,$$

and

$$\begin{aligned}\psi_\sigma(X_i-\mu_r) &= \frac{\tau^{1/2}}{2\Gamma\left(1+1/p\right)}\exp\left(-\tau^{p/2}|X_i-\mu_r|^p\right)\\ &\ge \frac{\tau_{\min}^{1/2}}{2\Gamma(1+1/p)}\exp\left(-\tau_{\max}^{p/2}\left[|X_i-\mu_{\sigma_0,r}| + k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right]^p\right).\end{aligned}$$

Now we build the upper bounds of $\left[|X_i-\mu_{\sigma_0,r}| + k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right]^p$ in two cases:

- If $k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2} \le \epsilon|X_i-\mu_{\sigma_0,r}|$, then

$$\left[|X_i-\mu_{\sigma_0,r}| + k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right]^p \le (1+\epsilon)^p|X_i-\mu_{\sigma_0,r}|^p.$$

- If $k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2} > \epsilon|X_i-\mu_{\sigma_0,r}|$, then

$$\left[|X_i-\mu_{\sigma_0,r}| + k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right]^p \le \left((1+\epsilon^{-1})k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right)^p.$$

Therefore,

$$\left[|X_i - \mu_{\sigma_0,r}| + k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right]^p \leq (1+\epsilon)^p |X_i - \mu_{\sigma_0,r}|^p + \left((1+\epsilon^{-1})k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right)^p,$$

and then

$$\begin{aligned}
& \psi_\sigma(X_i - \mu_r) \\
\geq\ & \exp\left(-\left(\tau_{\max}^{1/2}(1+\epsilon^{-1})k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right)^p\right) \\
& \times \frac{\tau_{\min}^{1/2}}{2\Gamma(1+1/p)} \exp\left(-((1+\epsilon)^2\tau_{\max})^{p/2}|X_i - \mu_{\sigma_0,r}|^p\right) \\
=\ & \frac{\tau_{\min}^{1/2}}{(1+\epsilon)\tau_{\max}^{1/2}} \exp\left(-\left(\tau_{\max}^{1/2}(1+\epsilon^{-1})k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right)^p\right) \psi_{\sigma_0}(|X_i - \mu_{k_{\sigma_0},r}|).
\end{aligned}$$

where the last step we apply the fact that $(1+\epsilon)^{-1}\tau_{\max}^{-1/2} = (1+\epsilon)^{-1}\widetilde{\sigma}_0 = \sigma_0$. Thus, we have

$$\begin{aligned}
& \int \log w_r \psi_\sigma(X_i - \mu_r) dQ^{(k_{\sigma_0})}(\theta^{(k_{\sigma_0})}) \\
\geq\ & \log\left[(1-\nu)\frac{\tau_{\min}^{1/2}}{(1+\epsilon)\tau_{\max}^{1/2}} \exp\left(-\left(\tau_{\max}^{1/2}(1+\epsilon^{-1})k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right)^p\right)\right] + \log\psi_{\sigma_0}(|X_i - \mu_r|),
\end{aligned}$$

so

$$\begin{aligned}
& -\sum_{i=1}^n \log\left[\sum_{r=1}^{k_{\sigma_0}} \exp\left(\int \log w_r \psi_\sigma(X_i - \mu_r) dQ^{(k_{\sigma_0})}(\theta^{(k_{\sigma_0})})\right)\right] \\
\leq\ & -n\log\left[(1-\nu)\frac{\tau_{\min}^{1/2}}{(1+\epsilon)\tau_{\max}^{1/2}} \exp\left(-\left(\tau_{\max}^{1/2}(1+\epsilon^{-1})k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right)^p\right)\right] \\
& - \log p^n_{k_{\sigma_0},\theta^{*(k_{\sigma_0})}}(X^{(n)})
\end{aligned}$$

where $\theta^{*(k_{\sigma_0})} = (\mu_{\sigma_0}, w^*_{\sigma_0}, \sigma_0)$.

We plug the above upper bound into (5.22), and then

$$\begin{aligned}
& D\left(\bar{Q}^{(k_{\sigma_0})}\|\bar{\Pi}^{(k_{\sigma_0})}\right) + \bar{Q}^{(k_{\sigma_0})} D\left(P^n_{\widetilde{f}_0}\|P(\cdot|k_{\sigma_0}, z^{(k_{\sigma_0})}, \theta^{(k_{\sigma_0})})\right) - \log \pi(k_{\sigma_0}) \\
\leq\ & n\left[\log\frac{1+\epsilon}{1-\nu} + \frac{1}{2}\log\frac{(1+\epsilon)\tau_{\max}}{\tau_{\min}} + \left(\tau_{\max}^{1/2}(1+\epsilon^{-1})k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right)^p\right] \\
& + D\left(P^n_{\widetilde{f}_0}\|P^n_{k_{\sigma_0},\theta^{*(k_{\sigma_0})}}\right) + D\left(Q^{(k_{\sigma_0})}\|\Pi^{(k_{\sigma_0})}\right).
\end{aligned}$$

Now we build the upper bound for each term in the right hand side above. For the first term, when $\nu \leq 1/2$, $\frac{1}{1-\nu} \leq 1+2\nu$, we have

$$\begin{aligned}
& \log\frac{1+\epsilon}{1-\nu} + \frac{1}{2}\log\frac{(1+\epsilon)\tau_{\max}}{\tau_{\min}} + \left(\tau_{\max}^{1/2}(1+\epsilon^{-1})k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right)^p \\
\leq\ & \log(1+\epsilon)(1+2\nu) + \frac{1}{2}\log(1+\epsilon)(1+\widetilde{\sigma}_0^{H_1+2\alpha+1}) + \epsilon^{-p}|\log\sigma_0|^{-p^2/\xi_4}\sigma_0^{(H_1+2\alpha+3)p} \\
\leq\ & 2\epsilon + 2\nu + \epsilon + \frac{1}{2}(1+\epsilon)^{H_1+2\alpha+1}\sigma_0^{H_1+2\alpha+1} + \epsilon^{-p}\sigma_0^{(H_1+2\alpha+3)p}|\log\sigma_0|^{-p^2/\xi_4}.
\end{aligned}$$

Choose $\epsilon = \sigma_0^{2\alpha}$ and $\nu = \sigma_0^{2\alpha}$. When $H_1 > 2\alpha$ and $\sigma_0 \to 0$, we have

$$\log\frac{1+\epsilon}{1-\nu} + \frac{1}{2}\log\frac{(1+\epsilon)\tau_{\max}}{\tau_{\min}} + \left(\tau_{\max}^{1/2}(1+\epsilon^{-1})k_{\sigma_0}^{-1}\sigma_0^{H_1+2\alpha+2}\right)^p \lesssim \sigma_0^{2\alpha}.$$

Next, for the second term, as $w^*_{\sigma_0} \in \Delta_{k_{\sigma_0}}(w_{\sigma_0}, \sigma_0^{H_1+2\alpha+1})$, by (5.23), we have

$$D\left(P^n_{\widetilde{f}_0}\|P^n_{k_{\sigma_0},\theta^{*(k_{\sigma_0})}}\right) = nD\left(P_{\widetilde{f}_0}\|P_{k_{\sigma_0},\theta^{*(k_{\sigma_0})}}\right) \leq nD_2\left(P_{\widetilde{f}_0}\|P_{k_{\sigma_0},\theta^{*(k_{\sigma_0})}}\right) \lesssim n\sigma_0^{2\alpha}.$$

Finally, for the last term,

$$\begin{aligned}
& D\left(Q^{(k_{\sigma_0})}\|\Pi^{(k_{\sigma_0})}\right) \\
=\ & -\log\pi(k_{\sigma_0}) - \log\Pi_w^{(k_{\sigma_0})}(\widetilde{\Theta}_w^{(k_{\sigma_0})}) - \sum_{j=1}^{k_{\sigma_0}}\log\Pi_{\mu_j}(\widetilde{\Theta}_{\mu_j}^{(k_{\sigma_0})}) - \log\Pi_\sigma(\widetilde{\Theta}_\sigma^{(k_{\sigma_0})})
\end{aligned}$$

Then, by the same arguments as in the proof of Lemma 5.1.15, when $\sigma_0 = n^{-\frac{1}{2\alpha+1}}(\log n)^{\frac{2\alpha r}{2\alpha+1}}$,

we can further obtain that

$$-\log \Pi_{\mu_j}\left(\widetilde{\Theta}_{\mu_j}^{(k_{\sigma_0})}\right) \lesssim |\log \sigma_0|^{\max\{1, c_6/\xi_4\}},$$

$$-\log \Pi_w^{(k_{\sigma_0})}(\widetilde{\Theta}_w^{(k_{\sigma_0})}) \lesssim k_{\sigma_0}(\log k_{\sigma_0})^{d_3}|\log \frac{\sigma_0^{2\alpha}}{2k_{\sigma_0}}\sigma_0^{H_1+2\alpha+1}| \asymp k_{\sigma_0}(\log k_{\sigma_0})^{d_3}|\log \sigma_0|,$$

$$-\log \Pi_\sigma(\widetilde{\Theta}_{\sigma_0}^{(k_{\sigma_0})}) \lesssim |\log \widetilde{\sigma}_0| + \widetilde{\sigma}_0^{-b_6} \lesssim \sigma_0^{-1},$$

$$-\log \pi(k_{\sigma_0}) \lesssim k_{\sigma_0}\log k_{\sigma_0}.$$

Therefore, with the choice of $\xi_4 = \min\{p, \xi_3\}$, we have

$$D\left(Q^{(k_{\sigma_0})}\|\Pi^{(k_{\sigma_0})}\right) \lesssim \sigma_0^{-1}(\log \sigma_0)^r,$$

where r is the same defined in Theorem 2.4.2. Combining all the bounds above, we have

$$\begin{aligned} & D\left(\bar{Q}^{(k_{\sigma_0})}\|\bar{\Pi}^{(k_{\sigma_0})}\right) + \bar{Q}^{(k_{\sigma_0})} D\left(P_{\widetilde{f}_0}^n \| P(\cdot|k_{\sigma_0}, z^{(k_{\sigma_0})}, \theta^{(k_{\sigma_0})})\right) - \log \pi(k_{\sigma_0}) \\ \lesssim\ & n\sigma_0^{2\alpha} + \sigma_0^{-1}(\log \sigma_0)^r \asymp n^{\frac{1}{2\alpha+1}}(\log n)^{\frac{2\alpha r}{2\alpha+1}}. \end{aligned}$$

Finally, we have

$$P_{\widetilde{f}_0}^n\left[\widehat{Q}^{(\widehat{k})} n H\left(P_{\widehat{k},\theta^{(\widehat{k})}}, P_{\widetilde{f}_0}\right)^2\right] \lesssim n^{\frac{1}{2\alpha+1}}(\log n)^{\frac{2\alpha r}{2\alpha+1}}.$$

With the same change of measure argument in the proof of Theorem 2.4.2, the proof is complete. □

5.1.7 Proof of Theorem 2.5.1

Proof of Theorem 2.5.1. Recall that

$$d\widehat{Q}_{[k]} = \prod_{j\le k} dN\left(\frac{n}{n+j^{2\beta+1}}Y_j, \frac{1}{n+j^{2\beta+1}}\right) \prod_{j=k+1}^{n} dN(0, e^{-jn}) \prod_{j>n} \delta_0.$$

Then, we can decompose the risk into

$$\begin{aligned}
& P_{\theta^*}^{(n)}\widehat{Q}_{[k]}\|\theta-\theta^*\|^2 = P_{\theta^*}^{(n)}\widehat{Q}_{[k]}\sum_{j\le k}(\theta_j-\theta_j^*)^2 + P_{\theta^*}^{(n)}\widehat{Q}_{[k]}\sum_{j>k}(\theta_j-\theta_j^*)^2 \\
= \;& P_{\theta^*}^{(n)}\sum_{j\le k}\left(\frac{n}{n+j^{2\beta+1}}Y_j-\theta_j^*\right)^2 + \sum_{j>k}\theta_j^{*2} + \sum_{j\le k}\frac{1}{n+j^{2\beta+1}} + \sum_{j=k+1}^{n} e^{-jn} \\
= \;& \sum_{j\le k}\left(\frac{j^{2\beta+1}}{n+j^{2\beta+1}}\right)^2\theta_j^{*2} + \sum_{j>k}\theta_j^{*2} + \sum_{j\le k}\frac{n}{(n+j^{2\beta+1})^2} \\
& + \sum_{j\le k}\frac{1}{n+j^{2\beta+1}} + \sum_{j=k+1}^{n} e^{-jn}.
\end{aligned}$$

For the upper bound, we have

$$P_{\theta^*}^{(n)}\widehat{Q}_{[k]}\|\theta-\theta^*\|^2 \le \sum_{j\le k}\left(\frac{j^{2\beta+1}}{n+j^{2\beta+1}}\right)^2\theta_j^{*2} + \sum_{j>k}\theta_j^{*2} + 2\sum_{j\le k}\frac{1}{n+j^{2\beta+1}} + 2e^{-kn}.$$

Now we discuss in the two cases:

- When $k \le n^{\frac{1}{2\beta+1}}$, we have

$$\sum_{j>k}\theta_j^{*2} \le k^{-2\alpha}B^2, \qquad \sum_{j\le k}\frac{1}{n+j^{2\beta+1}} \le \frac{k}{n},$$

and

$$\sum_{j\le k}\left(\frac{j^{2\beta+1}}{n+j^{2\beta+1}}\right)^2\theta_j^{*2} \le \sum_{j\le k}\frac{j^{4\beta+2-2\alpha}}{n^2}j^{2\alpha}\theta_j^{*2} \le \frac{1+k^{4\beta+2-2\alpha}}{n^2}B^2.$$

Therefore,

$$P_{\theta^*}^{(n)}\widehat{Q}_{[k]}\|\theta-\theta^*\|^2 \lesssim k^{-2\alpha}+\frac{k}{n}.$$

- When $k > n^{\frac{1}{2\beta+1}}$, we have

$$\sum_{j\le k}\frac{1}{n+j^{2\beta+1}} \le \frac{n^{\frac{1}{2\beta+1}}}{n}+\sum_{j>n^{\frac{1}{2\beta+1}}} j^{-2\beta-1} \lesssim n^{-\frac{2\beta}{2\beta+1}},$$

and

$$\sum_{j\le k}\left(\frac{j^{2\beta+1}}{n+j^{2\beta+1}}\right)^2\theta_j^{*2} \le \sum_{j\le n^{\frac{1}{2\beta+1}}}\frac{j^{4\beta+2-2\alpha}}{n^2}j^{2\alpha}\theta_j^{*2}+\sum_{j>n^{\frac{1}{2\beta+1}}}\theta_j^{*2} \lesssim n^{-\frac{2\alpha}{2\beta+1}}.$$

Thus, we have

$$P_{\theta^*}^{(n)}\widehat{Q}_{[k]}\|\theta-\theta^*\|^2 \lesssim n^{-\frac{2(\alpha\wedge\beta)}{2\beta+1}}.$$

Now we prove the lower bound. According to the risk decomposition, we have

$$P_{\theta^*}^{(n)}\widehat{Q}_{[k]}\|\theta-\theta^*\|^2 \ge \sum_{j\le k}\left(\frac{j^{2\beta+1}}{n+j^{2\beta+1}}\right)^2\theta_j^{*2}+\sum_{j>k}\theta_j^{*2}+\sum_{j\le k}\frac{1}{n+j^{2\beta+1}}.$$

- When $k \le n^{\frac{1}{2\beta+1}}$, we consider a θ^* with every coordinate 0 except that $\theta^*_{k+1} = (k+1)^{-\alpha}B$. It is easy to check that $\theta^* \in \Theta_\alpha(B)$. Then, we have $\sum_{j\le k}\frac{1}{n+j^{2\beta+1}} \ge \frac{k}{2n}$ and $\sum_{j>k}\theta_j^{*2} \ge B^2(k+1)^{-2\alpha}$. Therefore,

$$\sup_{\theta^*\in\Theta_\alpha(B)} P_{\theta^*}^{(n)}\widehat{Q}_{[k]}\|\theta-\theta^*\|^2 \gtrsim k^{-2\alpha}+\frac{k}{n}.$$

- When $k > n^{\frac{1}{2\beta+1}}$, we consider a θ^* with every coordinate 0 except that $\theta^*_{\lceil n^{\frac{1}{2\beta+1}}\rceil} =$

$\left(\lceil n^{\frac{1}{2\beta+1}}\rceil\right)^{-\alpha} B$, and it is easy to check that $\theta^* \in \Theta_\alpha(B)$. Then, we have

$$\sum_{j\le k}\left(\frac{j^{2\beta+1}}{n+j^{2\beta+1}}\right)^2\theta_j^{*2} \ge \frac{1}{4}\theta^{*2}_{\lceil n^{\frac{1}{2\beta+1}}\rceil} \gtrsim n^{-\frac{2\alpha}{2\beta+1}},$$

and

$$\sum_{j\le k}\frac{1}{n+j^{2\beta+1}} \gtrsim \sum_{j\le n^{\frac{1}{2\beta+1}}}\frac{1}{n+j^{2\beta+1}} \gtrsim n^{-\frac{2\beta}{2\beta+1}}.$$

This leads to the lower bound

$$\sup_{\theta^*\in\Theta_\alpha(B)} P^{(n)}_{\theta^*}\widehat{Q}_{[k]}\|\theta-\theta^*\|^2 \gtrsim n^{-\frac{2(\alpha\wedge\beta)}{2\beta+1}}.$$

Now the proof is complete. □

5.1.8 Proofs of Theorem 2.5.2 and Theorem 2.5.3

Proof of Theorem 2.5.2. Define $Q = N(\beta_0, \tau^2 I_p)$, and then

$$Q\log(dQ) = -\frac{p}{2}\log(2\pi\tau^2 e),$$

which is a constant with respect to β_0. Thus,

$$\begin{aligned}
\widehat{Q}_{\tau^2} &= \operatorname*{argmin}_{Q\in\mathcal{S}_{\tau^2}} D\left(Q\|\Pi(\cdot|y)\right) \\
&= \operatorname*{argmin}_{Q\in\mathcal{S}_{\tau^2}}\{Q\log(dQ) - Q\log(d\Pi) - Q\log p_\beta(y)\} \\
&= \operatorname*{argmin}_{Q\in\mathcal{S}_{\tau^2}}\left\{\frac{1}{2}Q\|X\beta-y\|^2 + \lambda\sum_{i=1}^p Q|\beta_i|\right\}.
\end{aligned}$$

where $p_\beta(y) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\|y - X\beta\|^2\right)$. With $Q = N(\beta_0, \tau^2 I_p)$, we have β_j's independently drawn from $N(\beta_{0j}, \tau^2)$, and therefore,

$$Q\|X\beta - y\|^2 = \|X\beta_0 - y\|^2 + \tau^2 \mathrm{tr}(X^T X),$$

$$Q|\beta_i| = \tau Q|\tau^{-1}\beta_{0i}| = \tau h(\tau^{-1}\beta_{0i}),$$

where

$$\begin{aligned} h(x) &= \int_{-\infty}^{\infty} |t|\phi(t-x)dt = \int_{-\infty}^{\infty} |t+x|\phi(t)dt \\ &= \int_{-\infty}^{-x} -(t+x)\phi(t)dt + \int_{-x}^{\infty} (t+x)\phi(t)dt \\ &= \phi(t)\Big|_{-\infty}^{-x} - \phi(t)\Big|_{-x}^{\infty} - x\Phi(-x) + x\Phi(x) \\ &= 2\phi(x) + x[\Phi(x) - \Phi(-x)]. \end{aligned}$$

The proof is complete. □

Lemma 5.1.17.

$$0 \le h(x) - |x| \le \sqrt{\frac{2}{\pi}} \qquad \text{for all } x.$$

Proof of Lemma 5.1.17. It is not hard to see that $h(-x) = h(x)$. Thus, we only need to show the inequality for $x \ge 0$. For $x \ge 0$, set $d(x) = h(x) - x$, and then

$$d'(x) = \Phi(x) - \Phi(-x) - 1 \le 0.$$

Thus, $d(x)$ is monotonically decreasing when $x > 0$ and $d(x) \le d(0) = 2\phi(0) = \sqrt{\frac{2}{\pi}}$. For the left part of inequality, notice that $\frac{1-\Phi(x)}{\phi(x)} \le \frac{1}{x}$ for all $x > 0$ in [48], and then we can directly obtain that $d(x) \ge 0$. □

Proof of Theorem 2.5.3. We use the notation $H_\tau(\beta) = \sum_{j=1}^p \tau h(\beta_j/\tau)$. By Lemma 5.1.17, we have $|H_\tau(\beta) - \|\beta\|_1| \le \tau \sum_{j=1}^p |h(\beta_j/\tau) - \beta_j/\tau| \le p\tau\sqrt{2/\pi}$. By rearranging the basic

inequality $\|y - X\widehat{\beta}\|^2 + 2\lambda H_\tau(\widehat{\beta}) \leq \|y - X\beta^*\|^2 + 2\lambda H_\tau(\beta^*)$, we have

$$\|X(\widehat{\beta} - \beta^*)\|^2 \leq 2\left|\left\langle X(\widehat{\beta} - \beta^*), y - X\beta^* \right\rangle\right| + 2\lambda H_\tau(\beta^*) - 2\lambda H_\tau(\widehat{\beta}).$$

For the terms on the right hand side of the above inequality, we have $\left|\left\langle X(\widehat{\beta} - \beta^*), y - X\beta^* \right\rangle\right| \leq \|X^T(y - X\beta^*)\|_\infty \|\widehat{\beta} - \beta^*\|_1$ and $H_\tau(\beta^*) - H_\tau(\widehat{\beta}) \leq \|\beta^*\|_1 - \|\widehat{\beta}\|_1 + 2p\tau\sqrt{2/\pi}$. Therefore, with the notation $\Delta = \widehat{\beta} - \beta^*$, we have

$$\|X\Delta\|^2 \leq \lambda\|\Delta\|_1 + 2\lambda\|\beta^*\|_1 - 2\lambda\|\beta^* + \Delta\|_1 + 2\lambda p\tau\sqrt{2/\pi}, \tag{5.25}$$

as long as $\lambda \geq 2\|X^T(y - X\beta^*)\|_\infty$. Note that the choice $\lambda = C\sqrt{n \log p}$ implies that the condition $\lambda \geq 2\|X^T(y - X\beta^*)\|_\infty$ holds with probability at least $1 - p^{-C'}$ by a union bound argument in [9]. With the decompositions $\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1$, $\|\beta^*\|_1 = \|\beta^*_S\|_1$ and $\|\beta^* + \Delta\|_1 = \|\beta^*_S + \Delta_S\|_1 + \|\Delta_{S^c}\|_1$, the inequality (5.25) becomes

$$\|X\Delta\|^2 \leq \lambda\left(3\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1 + 2p\tau\sqrt{2/\pi}\right). \tag{5.26}$$

The inequality (5.26) immediately implies what is known as the generalized cone condition defined in [27],

$$\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1 + 2p\tau\sqrt{2/\pi}. \tag{5.27}$$

Another consequence of (5.26) is the error bound

$$\|X\Delta\|^2 \leq \lambda\left(3\sqrt{s}\|\Delta\| + 2p\tau\sqrt{2/\pi}\right). \tag{5.28}$$

For the Δ that satisfies (5.27), we define

$$\Delta^{(1)} \quad = \quad \frac{3\|\Delta_S\|_1 + 2p\tau\sqrt{2/\pi}}{3\|\Delta_S\|_1}\Delta_S + \frac{3\|\Delta_S\|_1}{3\|\Delta_S\|_1 + 2p\tau\sqrt{2/\pi}}\Delta_{S^c},$$

$$\Delta^{(2)} = -\frac{2p\tau\sqrt{2/\pi}}{3\|\Delta_S\|_1}\Delta_S + \frac{2p\tau\sqrt{2/\pi}}{3\|\Delta_S\|_1 + 2p\tau\sqrt{2/\pi}}\Delta_{S^c}.$$

It is easy to see that $\Delta = \Delta^{(1)} + \Delta^{(2)}$. Since

$$\begin{aligned}\|\Delta^{(1)}_{S^c}\|_1 &= \frac{3\|\Delta_S\|_1}{3\|\Delta_S\|_1 + 2p\tau\sqrt{2/\pi}}\|\Delta_{S^c}\|_1 \\ \leq \frac{3\|\Delta_S\|_1}{3\|\Delta_S\|_1 + 2p\tau\sqrt{2/\pi}}&(3\|\Delta_S\|_1 + 2p\tau\sqrt{2/\pi}) = 3\|\Delta_S\|_1 \leq 3\|\Delta^{(1)}_S\|_1,\end{aligned}$$

we have $\frac{1}{\sqrt{n}}\|X\Delta^{(1)}\| \geq \kappa\|\Delta^{(1)}\|$ by the definition of κ in (2.53). We also bound $\|\Delta^{(2)}\|$ by

$$\|\Delta^{(2)}\| \leq \|\Delta^{(2)}\|_1 \leq \frac{2p\tau\sqrt{2/\pi}}{3} + \frac{2p\tau\sqrt{2/\pi}}{3\|\Delta_S\|_1 + 2p\tau\sqrt{2/\pi}}\|\Delta_{S^c}\|_1 \leq \frac{8p\tau\sqrt{2/\pi}}{3},$$

where the last inequality is by (5.27). Therefore,

$$\|\Delta\| \leq \|\Delta^{(1)}\| + \|\Delta^{(2)}\| \leq \frac{\|X\Delta^{(1)}\|}{\kappa\sqrt{n}} + \frac{8p\tau\sqrt{2/\pi}}{3}.$$

Since $\|X\Delta^{(2)}\| \leq \sqrt{n\max_{i,j}|X_{ij}|^2\|\Delta^{(2)}\|_1^2} \leq \sqrt{L^2n^2\|\Delta^{(2)}\|_1^2} \leq nL\frac{8p\tau\sqrt{2/\pi}}{3}$, we have

$$\|\Delta\| \leq \frac{\|X\Delta\|}{\kappa\sqrt{n}} + \frac{8p\tau\sqrt{2/\pi}}{3} + \frac{\|X\Delta^{(2)}\|}{\kappa\sqrt{n}} \leq \frac{\|X\Delta\|}{\kappa\sqrt{n}} + \left(1 + \frac{\sqrt{n}L}{\kappa}\right)\frac{8p\tau\sqrt{2/\pi}}{3}.$$

Combining the above inequality and (5.28), we have

$$\|\Delta\|^2 \lesssim \frac{\|X\Delta\|^2}{n\kappa^2} + \left(1 + \frac{n}{\kappa^2}\right)p^2\tau^2 \lesssim \left(\frac{\lambda\sqrt{s}}{n\kappa^2}\|\Delta\| + \frac{p\tau\lambda}{n\kappa^2}\right) + \left(1 + \frac{n}{\kappa^2}\right)p^2\tau^2,$$

which further leads to

$$\|\Delta\|^2 \lesssim \frac{\lambda^2 s}{n^2\kappa^4} + \frac{p\tau\lambda}{n\kappa^2} + \left(1 + \frac{n}{\kappa^2}\right)p^2\tau^2.$$

With $\lambda = C\sqrt{n\log p}$ and $\tau = O\left(\frac{1}{np}\right)$, we have $\|\Delta\|^2 \lesssim \frac{s\log p}{n\kappa^4}$, which completes the proof. $\square$

5.1.9 Proofs of Theorem 2.5.4, Theorem 2.5.5 and Theorem 2.5.6

To show Theorem 2.5.4, we need the following three lemmas.

Lemma 5.1.18. *If the conditions (C1)-(C3) in Theorem 2.2.1 are satisfied, then there exists a constant $M_0 > 0$ large enough such that when $n\epsilon^2 \geq n\epsilon_n^2 \geq M_0$ and $0 < a \leq \frac{m}{2C_1}$, we have*

$$P_0^{(n)}\left(\log \Pi\left[\exp\left(aL(P_\theta^{(n)}, P_0^{(n)})\right)\Big| X^{(n)}\right] \geq aC_1 n\epsilon^2 + \log 2\right) \leq \exp\left(-\frac{m}{2}n\epsilon^2\right),$$

where $m = \frac{1}{2}\min\{1, \rho - 1\}$.

Lemma 5.1.19. *Under the conditions (C1)-(C3) in Theorem 2.2.1, there exist some constants $M_0 > 1$, $M > 0$ and $c > 0$ such that when $n\epsilon^2 \geq n\epsilon_n^2 \geq M_0$,*

$$P_0^{(n)}\left(\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) > M\left(D\left(\widehat{Q}\|\Pi(\cdot|X^{(n)})\right) + n\epsilon^2\right)\right) \leq \exp\left(-cn\epsilon^2\right),$$

where $\widehat{Q}$ is the variational posterior distribution defined in (2.3).

Lemma 5.1.20. *Suppose the conditions (C1)-(C3) in Theorem 2.2.1 hold for $P_0^{(n)} = P_{\theta_0}^{(n)}$, when $n\epsilon^2 \geq n\epsilon_n^2 \geq M_0$, with any $p > 1$ as a constant*

$$\begin{aligned} & P_*^{(n)}\left(\widehat{Q}L(P_\theta^{(n)}, P_{\theta_0}^{(n)}) > M\left(D\left(\widehat{Q}\|\Pi(\cdot|X^{(n)})\right) + c^{-1}D_p\left(P_*^{(n)}\|P_{\theta_0}^{(n)}\right) + n\epsilon^2\right)\right) \\ \leq\ & \exp\left(-\frac{(p-1)c}{p}n\epsilon^2\right), \end{aligned}$$

where M_0, M and c are the same constants in Lemma 5.1.19 and $\widehat{Q}$ is the variational posterior distribution defined in (2.3).

Proof of Lemma 5.1.18. According to Lemma 5.1.3

$$P_0^{(n)}\Pi\left(L(P_\theta^{(n)}, P_0^{(n)}) > C_1 n\epsilon^2\Big| X^{(n)}\right) \leq 4\exp\left(-2mn\epsilon^2\right),$$

with $m = \frac{1}{2}\min\{1, \rho - 1\}$ for any $\epsilon \geq \epsilon_n$. Then by Markov inequality,

$$P_0^{(n)}\left[\Pi\left(L(P_\theta^{(n)}, P_0^{(n)}) > C_1 n\epsilon^2 \middle| X^{(n)}\right) > \exp\left(-mn\epsilon^2\right)\right] \leq 4\exp\left(-mn\epsilon^2\right).$$

Denote $B_j = \left\{X^{(n)} \middle| \Pi\left(L(P_\theta^{(n)}, P_0^{(n)}) > C_1 jn\epsilon^2 \middle| X^{(n)}\right) \leq \exp\left(-mjn\epsilon^2\right)\right\}$ and $B = \cap_{j=1}^{\infty} B_j$. Then,

$$P_0^{(n)}(B^c) \leq \sum_{j=1}^{\infty} P_0^{(n)}(B_j^c) \leq 4\sum_{j=1}^{\infty}\exp(-mjn\epsilon^2) \leq \frac{4}{1-\exp(-mn\epsilon^2)}\exp(-mn\epsilon^2).$$

When $M_0 = \frac{2\log 8}{m}$ and $n\epsilon^2 \geq M_0$, it is easy to check that

$$P_0^{(n)}(B^c) \leq \exp\left(-\frac{m}{2}n\epsilon^2\right).$$

Under the event B,

$$\begin{aligned}
& \Pi\left[\exp\left(aL(P_\theta^{(n)}, P_0^{(n)})\right) \middle| X^{(n)}\right] \\
\leq\ & \exp\left(aC_1 n\epsilon^2\right) + \int_{\exp(aC_1 n\epsilon^2)}^{\infty} \Pi\left[\exp\left(aL(P_\theta^{(n)}, P_0^{(n)})\right) \geq t \middle| X^{(n)}\right] dt \\
\leq\ & \exp\left(aC_1 n\epsilon^2\right) + \\
& \sum_{j=1}^{\infty}\left[\exp\left((j+1)aC_1 n\epsilon^2\right) - \exp\left(jaC_1 n\epsilon^2\right)\right]\Pi\left[L(P_\theta^{(n)}, P_0^{(n)}) \geq jC_1 n\epsilon^2 \middle| X^{(n)}\right] \\
\leq\ & \exp\left(aC_1 n\epsilon^2\right) + \exp\left(aC_1 n\epsilon^2\right)\sum_{j=1}^{\infty}\exp\left((aC_1 - m)jn\epsilon^2\right) \\
\leq\ & \exp\left(aC_1 n\epsilon^2\right)\left[1 + \sum_{j=1}^{\infty}\exp\left(-\frac{m}{2}jn\epsilon^2\right)\right] \\
=\ & \exp\left(aC_1 n\epsilon^2\right)\frac{1}{1-\exp\left(-\frac{m}{2}n\epsilon^2\right)} \leq 2\exp\left(aC_1 n\epsilon^2\right),
\end{aligned}$$

where we have used the condition that $0 < a \leq \frac{m}{2C_1}$. The conclusion of the lemma directly follows the result above. □

Proof of Lemma 5.1.19. According to Lemma 5.1.1, for any $a > 0$, we have

$$a\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) \le D\left(\widehat{Q}\|\Pi(\cdot|X^{(n)})\right) + \log\Pi\left[\exp\left(aL(P_\theta^{(n)}, P_0^{(n)})\right)\Big|X^{(n)}\right].$$

Choose $a = \frac{\min\{1,\rho-1\}}{4C_1}$, and then according to Lemma 5.1.18, under the event B (defined in the proof of Lemma 5.1.18), for $n\epsilon^2 > M_0 > 1$, we have

$$\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) \le \frac{1}{a}\left(D\left(\widehat{Q}\|\Pi(\cdot|X^{(n)})\right) + \log 2\right) + C_1 n\epsilon^2 \le M\left(D\left(\widehat{Q}\|\Pi(\cdot|X^{(n)})\right) + n\epsilon^2\right),$$

with $M = \max\left\{\frac{\log 2}{a} + C_1, \frac{1}{a}\right\}$. Therefore, the conclusion that

$$P_0^{(n)}\left(\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) > M\left(D\left(\widehat{Q}\|\Pi(\cdot|X^{(n)})\right) + n\epsilon^2\right)\right) \le \exp\left(-cn\epsilon^2\right),$$

is implied by $P_0^{(n)}(B^c) \le \exp\left(-cn\epsilon^2\right)$, where $c = \frac{\min\{1,\rho-1\}}{4}$. □

Proof of Lemma 5.1.20. By Hölder's inequality,

$$\begin{aligned}
& P_*^{(n)}\left(\widehat{Q}L(P_\theta^{(n)}, P_{\theta_0}^{(n)}) > M\left[D\left(\widehat{Q}\|\Pi(\cdot|X^{(n)})\right) + n\epsilon^2\right]\right) \\
\le\ & \left(P_{\theta_0}^{(n)}\left(\frac{dP_*^{(n)}}{dP_{\theta_0}^{(n)}}\right)^p\right)^{1/p}\left(P_{\theta_0}^{(n)}\left(\widehat{Q}L(P_\theta^{(n)}, P_{\theta_0}^{(n)}) > M\left[D\left(\widehat{Q}\|\Pi(\cdot|X^{(n)})\right) + n\epsilon^2\right]\right)\right)^{1-1/p} \\
\le\ & \exp\left(-\frac{p-1}{p}\left(cn\epsilon^2 - D_p\left(P_*^{(n)}\|P_{\theta_0}^{(n)}\right)\right)\right)
\end{aligned}$$

where $q = \frac{p}{p-1}$. Replace $n\epsilon^2$ by $n\epsilon^2 + c^{-1}D_p\left(P_*^{(n)}\|P_{\theta_0}^{(n)}\right)$, and we have

$$\begin{aligned}
& P_*^{(n)}\left(\widehat{Q}L(P_\theta^{(n)}, P_0^{(n)}) > M\left(D\left(\widehat{Q}\|\Pi(\cdot|X^{(n)})\right) + c^{-1}D_p\left(P_*^{(n)}\|P_{\theta_0}^{(n)}\right) + n\epsilon^2\right)\right) \\
\le\ & \exp\left(-\frac{(p-1)c}{p}n\epsilon^2\right),
\end{aligned}$$

where M_0, M and c are the same constants in Lemma 5.1.19. □

Now we can show Theorem 2.5.4.

Proof of Theorem 2.5.4. Define

$$Y = M^{-1}\widehat{Q}L(P_\theta^{(n)}, P_{\theta_0}^{(n)}) - \left[D\left(\widehat{Q}\|\Pi(\cdot|X^{(n)})\right) + c^{-1}D_p\left(P_*^{(n)}\|P_{\theta_0}^{(n)}\right) + n\epsilon_n^2\right].$$

Then Lemma 5.1.20 implies that

$$P_*^{(n)}(Y \geq t) \leq \exp\left(-\frac{(p-1)c}{p}(n\epsilon_n^2 + t)\right),$$

for all $t \geq 0$. Note that

$$\begin{aligned} & P_*^{(n)}\left(\widehat{Q}L\left(P_\theta^{(n)}, P_{\theta_0}^{(n)}\right)\right) \\ = \; & MP_*^{(n)}D\left(\widehat{Q}\|\Pi(\cdot|X^{(n)})\right) + Mc^{-1}D_p\left(P_*^{(n)}\|P_{\theta_0}^{(n)}\right) + Mn\epsilon_n^2 + MP_*^{(n)}Y. \end{aligned}$$

For the first term on the right hand side of the above equality, we have

$$\begin{aligned} P_*^{(n)}D\left(\widehat{Q}\|\Pi(\cdot|X^{(n)})\right) &= P_*^{(n)}\inf_{Q\in\mathcal{S}} D\left(Q\|\Pi(\cdot|X^{(n)})\right) \\ &\leq \inf_{Q\in\mathcal{S}} P_*^{(n)}D\left(Q\|\Pi(\cdot|X^{(n)})\right) = n\gamma_n^2. \end{aligned}$$

The term $P_*^{(n)}Y$ can be bounded by

$$\begin{aligned} P_*^{(n)}Y &\leq P_*^{(n)}Y\mathbb{I}\{Y \geq 0\} \leq \int_0^\infty P_*^{(n)}(Y \geq t)dt \\ &\leq \int_0^\infty \exp\left(-\frac{(p-1)c}{p}(n\epsilon_n^2 + t)\right)dt \leq \frac{p}{(p-1)c}\exp\left(-\frac{(p-1)c}{p}n\epsilon_n^2\right) \\ &\lesssim n\epsilon_n^2 \end{aligned}$$

The proof is complete by choosing $p = 2$. □

Proof of Theorem 2.5.5. Theorem 2.5.5 uses the same arguments in the proof of Theorem

2.2.3 with ϵ_n^2 replaced by $\epsilon_n^2 + \frac{1}{n} D_2\left(P_*^{(n)} \| P_{\theta_0}^{(n)}\right)$. Therefore, we omit the details here. □

In the end of this part, we will show Theorem 2.5.6, which directly implies Theorem 2.3.5. We want to check conditions (C1)-(C3) for $\theta_0 \in \Theta_{k_0}(B)$. For this aim, we establish the following lemmas.

Lemma 5.1.21. *The marginal sampling process of θ in the prior for piecewise constant model can be regarded as following procedure:*

- *Sample $k \sim \pi(k)$ with*

$$\pi(k) = \frac{\Gamma(k-1+\alpha_0)\Gamma(n-k+\beta_0)\Gamma(\alpha_0+\beta_0)(n-1)!}{\Gamma(n-1+\alpha_0+\beta_0)\Gamma(\alpha_0)\Gamma(\beta_0)(k-1)!(n-k)!}; \tag{5.29}$$

- *Conditioning on k, sample $k-1$ change points uniformly from $\{2, 3, \cdots, n\}$. In the other words, we uniformly sample a subset $S \subseteq \{2, 3, \cdots, n\}$ of size $k-1$ with probability $\binom{n-1}{k-1}^{-1}$;*

- *Conditioning on S, sample θ_i according to $\theta_i \sim g_i$ for all $i \in S$ and $\theta_i = \theta_{i-1}$ for all $i \notin S$.*

Moreover, when (2.26) is satisfied,

$$n^{-(C_2+1)(k-1)-1} \leq \pi(k) \leq n^{-(C_1-1)(k-1)}.$$

Proof. First of all, the density of marginal prior on θ can be written as

$$\begin{aligned}
\frac{d\Pi(\theta)}{d\theta} &= \int \sum_z \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} w^{\alpha_0+\sum_{i=2}^n z_i - 1}(1-w)^{\beta_0+n-1-\sum_{i=2}^n z_i - 1} \\
&\quad \times g(\theta_1) \prod_{z_i=1, i>1} g(\theta_i) \prod_{z_i=0, i>1} \delta_{\theta_{i-1}}(\theta_i) dw \\
&= \sum_{k=1}^n \pi(k) \binom{n-1}{k-1}^{-1} \sum_{|S|=k-1} g(\theta_1) \prod_{i \in S} g(\theta_i) \prod_{i \notin S} \delta_{\theta_{i-1}}(\theta_i),
\end{aligned}$$

where S above is the set of label $2 \leq i \leq n$ such that $z_i = 1$ and $\pi(k)$ is defined in (5.29), which implies the marginal sampling process of θ can be written as the procedure above.

Then the condition (2.26) indicates that

$$\frac{\pi(k+1)}{\pi(k)} = \frac{k-1+\alpha_0}{n-k+\beta_0-1}\frac{n-k}{k} \leq \frac{n(\alpha_0+n-1)}{\beta_0} \leq n^{1-C_1},$$

$$\frac{\pi(k+1)}{\pi(k)} = \frac{k-1+\alpha_0}{n-k+\beta_0-1}\frac{n-k}{k} \geq \frac{\alpha_0}{(\beta_0+n)n} \geq n^{-C_2-1},$$

which implies that

$$n^{-(C_2+1)(k-1)}\pi(1) \leq \pi(k) \leq n^{-(C_1-1)(k-1)}\pi(1).$$

When $C_1 > 1$, $C_2 > 0$, it is easy to see that $1/n < \pi(1) < 1$ as $\pi(k)$ is decreasing with respect to k. Hence, we have

$$n^{-(C_2+1)(k-1)-1} \leq \pi(k) \leq n^{-(C_1-1)(k-1)}.$$

□

Lemma 5.1.22. *Suppose $\theta_0 \in \Theta_{k_0}$. For some integer $m \geq k_0$, define*

$$\widehat{\theta}_m = \underset{\theta \in \Theta_m}{\operatorname{argmin}} \|\theta - X\|^2. \tag{5.30}$$

Then for any $t \geq 24\sigma^2 r \log \frac{en}{r}$ with $r = \min\{n, m+k_0\}$, we have

$$P_{\theta^*}^{(n)}(\|\widehat{\theta}_m - \theta^*\|^2 > t) \leq \exp\left(-\frac{t}{16\sigma^2}\right).$$

Proof. According to the definition,

$$\|\widehat{\theta}_m - X\|^2 \leq \|\theta^* - X\|^2.$$

Using the identity $\|\widehat{\theta}_m - X\|^2 = \|\widehat{\theta}_m - \theta^*\|^2 + \|\theta^* - X\|^2 + 2\left\langle \widehat{\theta}_m - \theta^*, \theta^* - X \right\rangle$, we get

$$\|\widehat{\theta}_m - \theta^*\| \leq 2 \left| \left\langle \frac{\widehat{\theta}_m - \theta^*}{\|\widehat{\theta}_m - \theta^*\|}, X - \theta^* \right\rangle \right|.$$

Since $\frac{\widehat{\theta}_m - \theta^*}{\|\widehat{\theta}_m - \theta^*\|} \in \Theta_r$, we have

$$\|\widehat{\theta}_m - \theta^*\|^2 \leq 4\sigma^2 \sup_{\|u\|=1: u \in \Theta_r} \left| \sum_{i=1}^n u_i Z_i \right|^2,$$

where $Z_i \sim N(0,1)$. Then,

$$\begin{aligned} & P_{\theta^*}^{(n)}(\|\widehat{\theta}_m - \theta^*\|^2 > t) \leq \mathbb{P}\left(\sup_{\|u\|=1: u \in \Theta_r} \left| \sum_{i=1}^n u_i Z_i \right|^2 \geq \frac{t}{4\sigma^2} \right) \\ \leq & \sum_{x_1 + x_2 + \cdots + x_r = n} \mathbb{P}\left(\sup_{\sum_{i=1}^r x_i \widetilde{u}_i^2 = 1} \left| \sum_{i=1}^r \sqrt{x_i} \widetilde{u}_i \widetilde{Z}_i \right|^2 \geq \frac{t}{4\sigma^2} \right) \\ = & \sum_{x_1 + x_2 + \cdots + x_r = n} \mathbb{P}\left(\|\widetilde{Z}\|^2 \geq \frac{t}{4\sigma^2} \right), \end{aligned}$$

where $r = \min\{m + k_0, n\}$, $\widetilde{Z} = (\widetilde{Z}_1, \widetilde{Z}_2, \cdots, \widetilde{Z}_r)^T \sim N(0, I_r)$. Then a standard chi-squared bound gives

$$\begin{aligned} & P_{\theta^*}^{(n)}(\|\widehat{\theta}_m - \theta^*\|^2 > t) \leq \exp\left(r \log \frac{en}{r} + \frac{r}{2} \log 2 \right) \exp\left(-\frac{t}{8\sigma^2} \right) \\ \leq & \exp\left(\frac{t}{16\sigma^2} \right) \exp\left(-\frac{t}{8\sigma^2} \right) = \exp\left(-\frac{t}{16\sigma^2} \right). \end{aligned}$$

The proof is complete. □

We want check conditions (C1)-(C3) with respect to θ_0. This step can be split into the following two lemmas.

Lemma 5.1.23. *Assume $\theta_0 \in \Theta_{k_0}(B)$. For the prior Π that satisfies (2.26), the conditions*

(C1) and (C2) hold for all $\epsilon > \sqrt{\frac{k_0 \log n}{n}}$.

Lemma 5.1.24. *Assume* $\theta_0 \in \Theta_{k_0}(B)$. *For the prior* Π *that satisfies (2.26) and (2.27), the conditions (C3) and (C4**) hold for* $\epsilon_n = \sigma\sqrt{\frac{k_0 \log n}{n}}$ *with both* $\mathcal{S} = \mathcal{S}_{\mathrm{MC}}$ *and* $\mathcal{S} = \mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$.

Now we start to prove Lemma 5.1.23 and Lemma 5.1.24.

Proof of Lemma 5.1.23. For any $\epsilon > \sqrt{\frac{k_0 \log n}{n}}$, we set $m = \lceil \frac{C_0 n\epsilon^2}{2\log n} \rceil$. Choose a sufficiently large C_0 so that $m \geq 2k_0 \geq 2$. We consider $\Theta_n(\epsilon) = \Theta_r$ with $r = \min\{k_0 + m, n\}$. Then by the condition (2.26) and Lemma 5.1.21, we have

$$\Pi(\Theta_n(\epsilon)^c) = \sum_{j=r+1}^{n} \pi(j) \leq n^{-(C_1-1)r} \sum_{j=1}^{n-r} \pi(j) \leq n^{-(C_1-1)r} \leq \exp\left(-(C_1-1)C_0 n\epsilon^2\right),$$

which implies (C2).

To show (C1), we consider the testing function $\phi_n = \mathbb{I}\{\|\widehat{\theta}_m - \theta_0\| \geq 5\sigma\sqrt{(C_0+1)n\epsilon^2}\}$, where $\widehat{\theta}_m$ is defined in (5.30). Note that

$$\begin{aligned} &(5\sigma\epsilon\sqrt{(C_0+1)n})^2 \geq 24(C_0+1)\sigma^2 n\epsilon^2 \\ \geq\ & 24(C_0 n\epsilon^2 + k_0 \log n)\sigma^2 \geq 24(2m+k_0)\sigma^2 \log n, \end{aligned}$$

and we apply Lemma 5.1.22 to obtain

$$P_{\theta_0}^{(n)}\phi_n = P_{\theta_0}^{(n)}(\|\widehat{\theta}_m - \theta_0\|^2 \geq 25(C_0+1)\sigma^2 n\epsilon^2) \leq \exp\left(-\frac{25}{16}(C_0+1)n\epsilon^2\right).$$

Moreover, for any $\theta \in \Theta_n(\epsilon)$ and $\|\theta - \theta_0\|^2 \geq 10\sigma\epsilon\sqrt{(C_0+1)n}$, we have

$$(5\sigma\epsilon\sqrt{(C_0+1)n})^2 \geq 24(2m+k_0)\sigma^2 \log n \geq 24(m+r)\sigma^2 \log n,$$

and then,

$$\begin{aligned} P_\theta^{(n)}(1-\phi_n) &= P_\theta^{(n)}(\|\widehat{\theta}_m - \theta_0\| \le 5\sigma\epsilon\sqrt{(C_0+1)n}) \\ &\le P_\theta^{(n)}(\|\widehat{\theta}_m - \theta\| \ge 5\sigma\epsilon\sqrt{(C_0+1)n}) \\ &\le \exp\left(-\frac{25}{16}(C_0+1)n\epsilon^2\right). \end{aligned}$$

Therefore, (C1) is satisfied with a sufficiently large C_0. □

Proof of Lemma 5.1.24. We first verify condition (C3). Note that for any $\rho > 0$,

$$D_\rho\left(P_{\theta_0}^{(n)} \| P_\theta^{(n)}\right) = \frac{\rho}{2\sigma^2}\|\theta - \theta_0\|^2.$$

Consider the set $\Theta = \cup_{i=1}^n [\theta_{0i} - n^{-1/2}, \theta_{0i} + n^{-1/2}]$, then for $n > 1$,

$$\Theta \subseteq \left\{\theta : D_\rho\left(P_{\theta_0}^{(n)} \| P_\theta^{(n)}\right) \le \frac{\rho}{\sigma^2} \le \frac{\rho}{\sigma^2 \log 2} k_0 \log n\right\},$$

and

$$\begin{aligned} \Pi(\Theta) &\ge \pi(k_0) \prod_{i \in S(\theta_0)} \int_{\theta_{0i}-n^{-1/2}}^{\theta_{0i}+n^{-1/2}} g(\theta_j) d\theta_j \ge n^{-(C_2+1)(k_0-1)-1}\left(2cn^{-1/2}\right)^{k_0} \\ &\ge \exp\left(-\left(\frac{5}{2} + C_2 - \log(2c)\right) n\epsilon_n^2\right), \end{aligned}$$

where $S(\theta_0) = \{i : i > 1, \theta_{0i} \neq \theta_{0(i-1)}\} \cup \{1\}$ and C_2 is given in (2.26). Therefore, condition (C3) is satisfied.

Now we check condition (C4**) for both $\mathcal{S} = \mathcal{S}_{\mathrm{MC}}$ and $\mathcal{S} = \mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$. When $\mathcal{S} = \mathcal{S}_{\mathrm{MC}}$, assume $|S(\theta_0)| = \widetilde{k}_0$ and $S(\theta_0) = \{a_0 + 1, a_1 + 1, \cdots, a_{\widetilde{k}_0 - 1} + 1\}$ with $0 = a_0 < a_1 < \cdots < a_{\widetilde{k}_0} = n$. Since $\theta_0 \in \Theta_{k_0}(B)$, we must have $\widetilde{k}_0 \le k_0$. Define

$$\Theta_i = \left(\theta_{0i} - n^{-1/2}, \theta_{0i} + n^{-1/2}\right), \text{ for } i \in S(\theta_0).$$

Then we define

$$\Theta = \{\theta : \theta_i \in \Theta_i, \text{ for } i \in S(\theta_0), \theta_i = \theta_{i-1}, \text{ for } i \notin S(\theta_0)\}.$$

Then choose $dQ(\theta) = \frac{d\Pi(\theta)\mathbb{I}_\Theta(\theta)}{\Pi(\Theta)}$. As

$$dQ(\theta) = \prod_{i \in S(\theta_0)} \frac{g(\theta_i)\mathbb{I}_{\Theta_i}(\theta_i)}{\int_{\Theta_i} g(\theta_i)d\theta_i} \prod_{i \notin S(\theta_0)} \delta_{\theta_{i-1}}(\theta_i)d\theta,$$

we have $Q \in \mathcal{S}_{\mathrm{MC}}$. For any $\theta \in \mathrm{supp}(Q) = \Theta$,

$$\begin{aligned} D\left(P_{\theta^*}^{(n)} \| P_\theta^{(n)}\right) &= \frac{1}{2\sigma^2}\|\theta^* - \theta\|^2 \le \frac{1}{\sigma^2}\|\theta^* - \theta_0\|^2 + \frac{1}{\sigma^2}\|\theta_0 - \theta\|^2 \\ &\le \frac{1}{\sigma^2}\|\theta^* - \theta_0\|^2 + \frac{1}{\sigma^2} \le D_2\left(P_{\theta^*}^{(n)} \| P_\theta^{(n)}\right) + \sigma^{-2}k_0 \log n. \end{aligned}$$

Moreover,

$$\begin{aligned} D\left(Q\|\Pi\right) &= -\log \Pi(\Theta) = -\log \pi(\widetilde{k}_0) - \sum_{i \in S(\theta_0)} \log\left(\int_{\theta_{0i}-n^{-1/2}}^{\theta_{0i}+n^{-1/2}} g(\theta_i)d\theta_i\right) \\ &\lesssim \widetilde{k}_0 \log n \le k_0 \log n. \end{aligned}$$

Thus, condition (C4**) is satisfied for $\mathcal{S} = \mathcal{S}_{\mathrm{MC}}$.

When $\mathcal{S} = \mathcal{S}_{\mathrm{MC}}^{\mathrm{joint}}$. Choose $dQ^{\mathrm{joint}}(w, z, \theta) = dQ^{(w)}(w)\prod_{i=2}^n dQ_i^{(z)}(z_i)dQ^{(\theta)}(\theta)$, where

$$Q^{(w)} = \mathrm{Beta}(\widetilde{k}_0 - 1 + \alpha_0, n - \widetilde{k}_0 + \beta_0),$$

$$Q_i^{(z)}(z_i = 1) = \begin{cases} 0, & i \notin S(\theta_0), \\ 1, & i \in S(\theta_0), \end{cases} \qquad \text{for all } i > 1,$$

$$dQ^{(\theta)}(\theta) = \prod_{i \in S(\theta_0)} \frac{g(\theta_i)\mathbb{I}_{\Theta_i}(\theta_i)}{\int_{\Theta_i} g(\theta_i)d\theta_i} \prod_{i \notin S(\theta_0)} \delta_{\theta_{i-1}}(\theta_i)d\theta,$$

Obviously, we have $Q^{\text{joint}} \in \mathcal{S}_{\text{MC}}^{\text{joint}}$ and for any $\theta \in \text{supp}(Q^{(\theta)})$, we have shown that

$$D\left(P_{\theta^*}^{(n)} \| P_{\theta}^{(n)}\right) \lesssim D_2\left(P_{\theta^*}^{(n)} \| P_{\theta_0}^{(n)}\right) + k_0 \log n.$$

On the other hand, suppose $dQ^{(\theta)}(\theta) = q^{(\theta)}(\theta)d\theta$ and $dQ^{(w)}(w) = q^{(w)}(w)dw$, we have

$$\begin{aligned}
& D\left(Q^{\text{joint}}(w,z,\theta) \| \Pi(w,z,\theta)\right) \\
= & \iint q^{(w)}(w)q^{(\theta)}(\theta) \log \frac{q^{(w)}(w)q^{(\theta)}(\theta)}{\pi(w)w^{\widetilde{k}_0-1}(1-w)^{n-\widetilde{k}_0}\prod_{i\in S(\theta_0)} g(\theta_i)d\theta_i \prod_{i\notin S(\theta_0)} \delta_{\theta_{i-1}}(\theta_i)} d\theta dw \\
= & -\log \pi(\widetilde{k}_0) - \sum_{i\in S(\theta_0)} \log\left(\int_{\theta_{0i}-n^{-1/2}}^{\theta_{0i}+n^{-1/2}} g(\theta_i)d\theta_i\right) \\
\lesssim & \widetilde{k}_0 \log n \leq k_0 \log n.
\end{aligned}$$

Thus, condition (C4**) is satisfied for $\mathcal{S} = \mathcal{S}_{\text{MC}}^{\text{joint}}$. The proof is complete.

□

Proof of Theorem 2.5.6. By Lemma 5.1.23 and Lemma 5.1.24, together with Theorem 2.5.4 and Theorem 2.5.5, we have

$$P_{\theta^*}^{(n)}\widehat{Q}\|\theta - \theta_0\|^2 \lesssim k_0 \log n + \|\theta^* - \theta_0\|^2,$$

for both $\widehat{Q} = \widehat{Q}_{\text{MC}}$ and $\widehat{Q} = \widehat{Q}_{\text{MC}}^{\text{joint}}$. Then for every $1 \leq k_0 \leq n$ and $\theta_0 \in \Theta_{k_0}(B)$, we have

$$P_{\theta^*}^{(n)}\widehat{Q}\|\theta - \theta^*\|^2 \lesssim P_{\theta^*}^{(n)}\widehat{Q}\|\theta - \theta_0\|^2 + \|\theta_0 - \theta^*\|^2 \lesssim k_0 \log n + \|\theta^* - \theta_0\|^2.$$

Therefore, by taking minimum over $k_0 \in [n]$ and $\theta_0 \in \Theta_{k_0}(B)$, we can get

$$P_{\theta^*}^{(n)}\widehat{Q}\|\theta - \theta^*\|^2 \lesssim \min_{1\leq k\leq n}\left\{\inf_{\theta_0\in\Theta_k(B)} \|\theta^* - \theta_0\|^2 + k\log n\right\}.$$

The proof is complete. □

5.2 Proof in Chapter 3

Over all the proofs in this section, we treat $\mathcal{X}_Z$ as a matrix and use $\mathcal{X}_Z B$ to denote $\mathcal{X}_Z(B)$.

5.2.1 Proof of Proposition 3.2.1

Proposition 3.2.1 can be easily derived from the following Lemma.

Lemma 5.2.1. *For $a, b > 0$,*

$$\int_0^\infty \lambda^{-3/2} \exp\left(-\frac{1}{2}\left(a\lambda + \frac{b}{\lambda}\right)\right) d\lambda = \sqrt{\frac{2\pi}{b}} \exp(-\sqrt{ab}).$$

$$\int_0^\infty \lambda^{-1/2} \exp\left(-\frac{1}{2}\left(a\lambda + \frac{b}{\lambda}\right)\right) d\lambda = \sqrt{\frac{2\pi}{a}} \exp(-\sqrt{ab}).$$

Proof of Lemma 5.2.1. Set $t = \sqrt{a\lambda} - \sqrt{\frac{b}{\lambda}}$, then $t^2 = a\lambda + \frac{b}{\lambda} - 2\sqrt{ab}$

$$dt = \frac{1}{2}(\sqrt{a}\lambda^{-1/2} + \sqrt{b}\lambda^{-3/2})d\lambda = \frac{1}{2}(\sqrt{a\lambda} + \sqrt{\frac{b}{\lambda}})\lambda^{-1}d\lambda.$$

Then

$$\lambda^{-3/2}d\lambda = \frac{2}{\sqrt{\lambda}}\frac{1}{\sqrt{t^2 + 4\sqrt{ab}}}dt = \left(\frac{1}{\sqrt{b}} - \frac{t}{\sqrt{b(t^2 + 4\sqrt{ab})}}\right) dt.$$

$$\lambda^{-1/2}d\lambda = \frac{2\sqrt{\lambda}}{\sqrt{t^2 + 4\sqrt{ab}}}dt = \left(\frac{1}{\sqrt{a}} + \frac{t}{\sqrt{a(t^2 + 4\sqrt{ab})}}\right) dt.$$

Thus,

$$\int_0^\infty \lambda^{-3/2} \exp\left(-\frac{1}{2}\left(a\lambda + \frac{b}{\lambda}\right)\right) d\lambda = \int_{-\infty}^\infty \exp\left(-\frac{1}{2}t^2 - \sqrt{ab}\right)\left(\frac{1}{\sqrt{b}} - \frac{t}{\sqrt{b(t^2 + 4\sqrt{ab})}}\right) dt,$$

and

$$\int_0^\infty \lambda^{-1/2}\exp\left(-\frac{1}{2}\left(a\lambda+\frac{b}{\lambda}\right)\right)d\lambda = \int_{-\infty}^\infty \exp\left(-\frac{1}{2}t^2-\sqrt{ab}\right)\left(\frac{1}{\sqrt{a}}+\frac{t}{\sqrt{a(t^2+4\sqrt{ab})}}\right)dt.$$

According to symmetry, $\int_{-\infty}^\infty \exp(-\frac{1}{2}t^2-\sqrt{ab})\frac{t}{\sqrt{(t^2+4\sqrt{ab})}}dt = 0$, then

$$\int_0^\infty \lambda^{-3/2}\exp\left(-\frac{1}{2}(a\lambda+\frac{b}{\lambda})\right)d\lambda = \int_{-\infty}^\infty \exp(-\frac{1}{2}t^2-\sqrt{ab})\frac{1}{\sqrt{b}}dt = \sqrt{\frac{2\pi}{b}}\exp(-\sqrt{ab}).$$

$$\int_0^\infty \lambda^{-1/2}\exp\left(-\frac{1}{2}(a\lambda+\frac{b}{\lambda})\right)d\lambda = \int_{-\infty}^\infty \exp(-\frac{1}{2}t^2-\sqrt{ab})\frac{1}{\sqrt{a}}dt = \sqrt{\frac{2\pi}{a}}\exp(-\sqrt{ab}).$$

□

Proof of Proposition 3.2.1. Now we start to show Proposition 3.2.1. Suppose the marginal sampling distribution for B is $\widetilde{g}(B)$, then

$$\begin{aligned}
&\widetilde{g}(B)\\
=\ &\int \pi(\lambda;(\ell_\tau+1)/2,\beta) f_{\ell_\tau,\mathcal{X}_Z,\lambda}(B)d\lambda\\
=\ &\int \frac{\beta^{\frac{\ell_\tau+1}{2}}}{\Gamma\left(\frac{\ell_\tau+1}{2}\right)}\lambda^{-\frac{\ell_\tau+3}{2}}\exp\left(-\frac{\beta}{\lambda}\right)\lambda^{\ell_\tau/2}\det(\mathcal{X}_Z^T\mathcal{X}_Z)^{1/2}(2\pi)^{-\ell_\tau/2}\exp\left(-\frac{\lambda}{2}\|\mathcal{X}_ZB\|^2\right)d\lambda\\
=\ &\frac{\beta^{\frac{\ell_\tau+1}{2}}}{\Gamma\left(\frac{\ell_\tau+1}{2}\right)}\det(\mathcal{X}_Z^T\mathcal{X}_Z)^{1/2}(2\pi)^{-\ell_\tau/2}\int \lambda^{-\frac{3}{2}}\exp\left(-\frac{\lambda}{2}\|\mathcal{X}_ZB\|^2-\frac{\beta}{\lambda}\right)d\lambda\\
=\ &\frac{\beta^{\frac{\ell_\tau+1}{2}}}{\Gamma\left(\frac{\ell_\tau+1}{2}\right)}\det(\mathcal{X}_Z^T\mathcal{X}_Z)^{1/2}(2\pi)^{-\ell_\tau/2}\sqrt{\frac{\pi}{\beta}}\exp\left(-\sqrt{2\beta}\|\mathcal{X}_ZB\|\right).
\end{aligned}$$

Note that

$$g(B) = \frac{\sqrt{\det(\mathcal{X}_Z^T\mathcal{X}_Z)}}{2}\left(\frac{2\beta}{\pi}\right)^{\ell_\tau/2}\exp\left(-\sqrt{2\beta}\|\mathcal{X}_ZB\|\right)\frac{\Gamma(\ell_\tau/2)}{\Gamma(\ell_\tau)},$$

then

$$\begin{aligned}\frac{\widetilde{g}(B)}{g(B)} &= \frac{\Gamma(\ell_\tau)\sqrt{\pi}}{2^{\ell_\tau-1}\Gamma(\ell_\tau/2)\Gamma((\ell_\tau+1)/2)} \\ &= \frac{\Gamma(\ell_\tau)\sqrt{\pi}}{2^{\ell_\tau-1}\left((\ell_\tau-1)/2\right)\left((\ell_\tau-2)/2\right)\cdots\frac{1}{2}\Gamma(1)\Gamma(1/2)} \\ &= 1.\end{aligned}$$

□

5.2.2 Proof of Theorem 3.2.1

To show Theorem 3.2.1, we first prove the following lemma:

Lemma 5.2.2. *For any $\tau \in \mathcal{T}$ and $Q^{(\tau)} \in \mathcal{S}_{\mathrm{MF}}^{(\tau)}$, we have*

$$\begin{aligned}& Q^{(\tau)}\|\mathcal{X}_Z B - \theta^*\|^2 \\ \leq\ & \inf_{a>0}\frac{1}{a}\left[-F(Q^{(\tau)},\tau) + \log p_\Pi(Y) + \log\Pi\left(\exp\left(a\|\mathcal{X}_Z B - \theta^*\|^2\right)\Big|Y\right)\right],\end{aligned}$$

where Π is the prior sampling distribution,

$$p_\Pi = \int \phi_{\mathcal{X}_Z B} d\Pi = \sum_{\tau\in\mathcal{T}}\pi(\tau)\int \phi_{\mathcal{X}_Z B} d\Pi^{(\tau)}(Z,B).$$

This lemma is actually an enhanced version of Lemma 5.2.2, the proof is also likewise.

Proof of Lemma 5.2.2. A lower bound can be directly derived from the right hand side minus the left hand side. For any $a > 0$, any $\tau \in \mathcal{T}$, and any $Q^{(\tau)} \in \mathcal{S}_{\mathrm{MF}}^{(\tau)}$, we have

$$\begin{aligned}& -F(Q^{(\tau)},\tau) + \log p_\Pi(Y) - a\left[Q^{(\tau)}\|\mathcal{X}_Z B - \theta^*\|^2\right] \\ =\ & D\left(Q_Z^{(\tau)}Q_B^{(\tau)}Q_\lambda^{(\tau)}\|\Pi^{(\tau)}(Z,B,\lambda)\right) - \int \log\phi_{\mathcal{X}_Z B}(Y) dQ^{(\tau)}(Z,B) \\ & - \log\pi(\tau) + \log p_\Pi(Y) - a\left[Q^{(\tau)}\|\mathcal{X}_Z B - \theta^*\|^2\right]\end{aligned}$$

$$
\begin{aligned}
&= Q^{(\tau)} \log \frac{dQ^{(\tau)}(Z,B,\lambda)p_{\Pi}(Y)}{\pi(\tau)d\Pi^{(\tau)}(Z,B,\lambda)\phi_{\mathcal{X}_Z B}(Y)\exp\left(a\|\mathcal{X}_Z B-\theta^*\|^2\right)} \\
&= D\left(Q_Z^{(\tau)}Q_B^{(\tau)}Q_\lambda^{(\tau)}\|\widetilde{\Pi}^{(\tau)}(Z,B,\lambda)\right) \\
&\quad - \log \frac{\int \pi(\tau)\phi_{\mathcal{X}_Z B}(Y)\exp\left(a\|\mathcal{X}_Z B-\theta^*\|^2\right)d\Pi^{(\tau)}(Z,B)}{p_{\Pi}(Y)} \\
&\geq -P_{\theta^*}\log \frac{\sum_{\tau\in\mathcal{T}}\int \pi(\tau)\phi_{\mathcal{X}_Z B}(Y)\exp\left(a\|\mathcal{X}_Z B-\theta^*\|^2\right)d\Pi^{(\tau)}(Z,B)}{p_{\Pi}(Y)} \\
&= -P_{\theta^*}\log\Pi\left(\exp\left(a\|\mathcal{X}_Z B-\theta^*\|^2\right)\Big|Y\right),
\end{aligned}
$$

and

$$
d\widetilde{\Pi}^{(\tau)}(Z,B,\lambda) = \frac{\phi_{\mathcal{X}_Z B}(Y)\exp\left(a\|\mathcal{X}_Z B-\theta^*\|^2\right)d\Pi^{(\tau)}(Z,B,\lambda)}{\int \phi_{\mathcal{X}_Z B}(Y)\exp\left(a\|\mathcal{X}_Z B-\theta^*\|^2\right)d\Pi^{(\tau)}(Z,B,\lambda)}.
$$

The proof is complete. □

With Lemma 5.2.2 and the definition $\widehat{Q}^{(\widehat{\tau})}$, we have

$$
\begin{aligned}
&P_{\theta^*}\widehat{Q}^{(\widehat{\tau})}\|\mathcal{X}_Z B-\theta^*\|^2 \qquad (5.31)\\
\leq\ &\inf_{a>0}\frac{1}{a}P_{\theta^*}\Big[\inf_{\tau\in\mathcal{T}}\inf_{Q^{(\tau)}\in\mathcal{S}_{\mathrm{MF}}^{(\tau)}}\left(-F(Q^{(\tau)},\tau)\right)+\log p_{\Pi}(Y) \\
&+\log\Pi\left(\exp\left(a\|\mathcal{X}_Z B-\theta^*\|^2\right)\Big|Y\right)\Big].
\end{aligned}
$$

Then the proof of Theorem 3.2.1 can be established by showing the following lemmas.

Lemma 5.2.3. *For any $\delta_1 > 0$, $\tau^* \in \mathcal{T}$, $Z^* \in \mathcal{Z}_{\tau^*}$ and $B^* \in \mathbb{R}^{\ell_{\tau^*}}$, there exists a $\widetilde{\tau}$, $\widetilde{Q}^{(\widetilde{\tau})} \in \mathcal{S}_{\mathrm{MF}}^{(\widetilde{\tau})}$ that may depend on Y and a constant D_0 only related to δ_1 such that when $D > D_0$ in the prior, we have*

$$
P_{\theta^*}\left[-F(\widetilde{Q}^{(\widetilde{\tau})},\widetilde{\tau})+\log p_{\Pi}(Y)\right] \leq \delta_1\|\mathcal{X}_{Z^*}B^*-\theta^*\|^2 + C_1\epsilon_{\tau^*}, \qquad (5.32)
$$

for some constant $C_1 > 0$.

Lemma 5.2.4. *For any $\delta_2 > 0$, $\tau^* \in \mathcal{T}$, $Z^* \in \mathcal{Z}_{\tau^*}$ and $B^* \in \mathbb{R}^{\ell_{\tau^*}}$, there exists a constant*

$a > 0$, *such that*

$$P_{\theta^*} \log \Pi \left(\exp \left(a \|\mathcal{X}_Z B - \theta^*\|^2 \right) \Big| Y \right) \leq (1 + \delta_2) a \|\mathcal{X}_{Z^*} B^* - \theta^*\|^2 + C_2 \epsilon_{\tau^*}. \tag{5.33}$$

Proof of Lemma 5.2.3. The random variable part related to Y in the left hand side of (5.32) can be expanded as

$$\begin{aligned} & -F(\widetilde{Q}^{(\widetilde{\tau})}, \widetilde{\tau}) + \log p_\Pi(Y) \qquad (5.34) \\ = \;& D\left(\widetilde{Q}_Z^{(\widetilde{\tau})} \widetilde{Q}_B^{(\widetilde{\tau})} \widetilde{Q}_\lambda^{(\widetilde{\tau})} \| \Pi^{(\widetilde{\tau})}(Z, B, \lambda) \right) - \int \log \phi_{\mathcal{X}_Z B} d\widetilde{Q}^{(\widetilde{\tau})}(Z, B) - \log \pi(\widetilde{\tau}) + \log p_\Pi(Y). \end{aligned}$$

First of all, the prior joint density of (Z, λ, B) on $\widetilde{\tau}$ is given by

$$\pi^{(\widetilde{\tau})}(Z, \lambda, B) = \frac{1}{|\bar{\mathcal{Z}}_{\widetilde{\tau}}|} \sqrt{\frac{\beta^{\ell_{\widetilde{\tau}}+1} \det(\mathcal{X}_Z^T \mathcal{X}_Z)}{(2\pi)^{\ell_{\widetilde{\tau}}}}} \frac{1}{\Gamma\left(\frac{\ell_{\widetilde{\tau}}+1}{2}\right)} \lambda^{-\frac{3}{2}} \exp\left(-\frac{\lambda}{2} \|\mathcal{X}_Z B\|^2 - \frac{\beta}{\lambda} \right). \tag{5.35}$$

Now we assume

$$\widetilde{Q}_Z^{(\widetilde{\tau})}(Z = \widetilde{Z}) = 1,$$

$$\widetilde{Q}_B^{(\widetilde{\tau})} = N(\mu, \Sigma),$$

$$\frac{d\widetilde{Q}_\lambda^{(\widetilde{\tau})}}{d\lambda} = \sqrt{\frac{\beta}{\pi}} \lambda^{-3/2} \exp\left(\sqrt{2a\beta} - \frac{\beta}{\lambda} - \frac{a\lambda}{2} \right),$$

where $\widetilde{\tau} \in \mathcal{T}$, $\widetilde{Z} \in \bar{\mathcal{Z}}_{\widetilde{\tau}}$, $a \in \mathbb{R}$, $\mu \in \mathbb{R}^{\ell_{\widetilde{\tau}}}$ and $\Sigma \in \mathbb{R}^{\ell_{\widetilde{\tau}} \times \ell_{\widetilde{\tau}}}$ are to be determined later. Then the variational joint density of (Z, λ, B) on $\widetilde{\tau}$ is given by

$$\widetilde{q}^{(\widetilde{\tau})}(Z, \lambda, B) = \sqrt{\frac{2\beta}{(2\pi)^{\ell_{\widetilde{\tau}}+1} \det(\Sigma)}} \lambda^{-3/2} \exp\left(\sqrt{2a\beta} - \frac{\beta}{\lambda} - \frac{a\lambda}{2} - \frac{1}{2}(B - \mu)^T \Sigma^{-1} (B - \mu) \right). \tag{5.36}$$

Plug (5.35) and (5.36) into the first part of (5.34), we obtain

$$D\left(\widetilde{Q}_Z^{(\widetilde{\tau})} \widetilde{Q}_B^{(\widetilde{\tau})} \widetilde{Q}_\lambda^{(\widetilde{\tau})} \| \Pi^{(\widetilde{\tau})}(Z, B, \lambda) \right) - \int \log \phi_{\mathcal{X}_Z B} d\widetilde{Q}^{(\widetilde{\tau})}(Z, B)$$

$$
\begin{aligned}
= & -\frac{\ell_\tau}{2}\log\beta - \frac{1}{2}\log\pi - \frac{1}{2}\log\det(\Sigma) - \frac{1}{2}\log\det(\mathcal{X}_{\widetilde{Z}}^T\mathcal{X}_{\widetilde{Z}}) + \log|\bar{\mathcal{Z}}_{\widetilde{\tau}}| \\
& + \log\Gamma\left(\frac{\ell_{\widetilde{\tau}}+1}{2}\right) + \sqrt{2a\beta} \\
& + \frac{\widetilde{Q}_\lambda^{(\widetilde{\tau})}\lambda}{2}\left(\widetilde{Q}_B^{(\widetilde{\tau})}\|\mathcal{X}_{\widetilde{Z}}B\|^2 - a\right) - \frac{1}{2}\widetilde{Q}_B^{(\widetilde{\tau})}\left[(B-\mu)^T\Sigma^{-1}(B-\mu)\right] \\
& + \frac{1}{2}\widetilde{Q}_B^{(\widetilde{\tau})}\|Y - \mathcal{X}_{\widetilde{Z}}B\|^2 + \frac{N}{2}\log(2\pi).
\end{aligned}
$$

Since $\widetilde{Q}_B^{(\tau)} = N(\mu, \Sigma)$, we can get

$$
\widetilde{Q}_B^{(\widetilde{\tau})}\left[(B-\mu)^T\Sigma^{-1}(B-\mu)\right] = \ell_{\widetilde{\tau}},
$$

$$
\widetilde{Q}_B^{(\widetilde{\tau})}\|\mathcal{X}_{\widetilde{Z}}B\|^2 = \|\mathcal{X}_{\widetilde{Z}}\mu\|^2 + \mathsf{Tr}(\mathcal{X}_{\widetilde{Z}}\Sigma\mathcal{X}_{\widetilde{Z}}^T).
$$

$$
\widetilde{Q}_B^{(\widetilde{\tau})}\|\mathcal{X}_{\widetilde{Z}}B - Y\|^2 = \|\mathcal{X}_{\widetilde{Z}}\mu - Y\|^2 + \mathsf{Tr}(\mathcal{X}_{\widetilde{Z}}\Sigma\mathcal{X}_{\widetilde{Z}}^T).
$$

Moreover, according to the lemma 5.2.1, we have

$$
\widetilde{Q}_\lambda^{(\widetilde{\tau})}\lambda = \frac{\int \lambda^{-1/2}\exp\left(-\frac{\beta}{\lambda} - \frac{a\lambda}{2}\right)d\lambda}{\int \lambda^{-3/2}\exp\left(-\frac{\beta}{\lambda} - \frac{a\lambda}{2}\right)d\lambda} = \sqrt{\frac{2\beta}{a}}.
$$

Plug all above into the expression of $D\left(\widetilde{Q}_Z^{(\widetilde{\tau})}\widetilde{Q}_B^{(\widetilde{\tau})}\widetilde{Q}_\lambda^{(\widetilde{\tau})}\|\Pi^{(\widetilde{\tau})}(Z,B,\lambda)\right)$, we can get

$$
\begin{aligned}
& D\left(\widetilde{Q}_Z^{(\widetilde{\tau})}\widetilde{Q}_B^{(\widetilde{\tau})}\widetilde{Q}_\lambda^{(\widetilde{\tau})}\|\Pi^{(\widetilde{\tau})}(Z,B,\lambda)\right) - \int \log\phi_{\mathcal{X}_Z B}\, d\widetilde{Q}^{(\widetilde{\tau})}(Z,B) \\
= & -\frac{\ell_{\widetilde{\tau}}}{2}\log(\beta e) - \frac{1}{2}\log\det(\Sigma) - \frac{1}{2}\log\det(\mathcal{X}_{\widetilde{Z}}^T\mathcal{X}_{\widetilde{Z}}) + \log|\bar{\mathcal{Z}}_{\widetilde{\tau}}| \\
& + \sqrt{\frac{a\beta}{2}} + \sqrt{\frac{\beta}{2a}}\left(\|\mathcal{X}_{\widetilde{Z}}\mu\|^2 + \mathsf{Tr}(\mathcal{X}_{\widetilde{Z}}\Sigma\mathcal{X}_{\widetilde{Z}}^T)\right) \qquad (5.37) \\
& + \frac{1}{2}\|\mathcal{X}_{\widetilde{Z}}\mu - Y\|^2 + \frac{1}{2}\mathsf{Tr}(\mathcal{X}_{\widetilde{Z}}\Sigma\mathcal{X}_{\widetilde{Z}}^T) + \log\Gamma\left(\frac{\ell_{\widetilde{\tau}}+1}{2}\right) + \frac{N}{2}\log(2\pi) - \frac{1}{2}\log\pi.
\end{aligned}
$$

Now we set $a = \|\mathcal{X}_{\widetilde{Z}}\mu\|^2 + \mathsf{Tr}(\mathcal{X}_{\widetilde{Z}}\Sigma\mathcal{X}_{\widetilde{Z}}^T)$, then

$$
D\left(\widetilde{Q}_Z^{(\widetilde{\tau})}\widetilde{Q}_B^{(\widetilde{\tau})}\widetilde{Q}_\lambda^{(\widetilde{\tau})}\|\Pi^{(\widetilde{\tau})}(Z,B,\lambda)\right) - \int \log\phi_{\mathcal{X}_Z B}\, d\widetilde{Q}^{(\widetilde{\tau})}(Z,B)
$$

$$
\begin{aligned}
&= -\frac{\ell_{\widetilde{\tau}}}{2}\log(\beta e) - \frac{1}{2}\log\det(\Sigma) - \frac{1}{2}\log\det(\mathcal{X}_{\widetilde{Z}}^T\mathcal{X}_{\widetilde{Z}}) + \log|\bar{\mathcal{Z}}_{\widetilde{\tau}}| \\
&\quad + \sqrt{2\beta\left(\|\mathcal{X}_{\widetilde{Z}}\mu\|^2 + \mathsf{Tr}(\mathcal{X}_{\widetilde{Z}}\Sigma\mathcal{X}_{\widetilde{Z}}^T)\right)} + \frac{1}{2}\|\mathcal{X}_{\widetilde{Z}}\mu - Y\|^2 + \frac{1}{2}\mathsf{Tr}(\mathcal{X}_{\widetilde{Z}}\Sigma\mathcal{X}_{\widetilde{Z}}^T) \\
&\leq -\frac{\ell_{\widetilde{\tau}}}{2}\log(\beta e) - \frac{1}{2}\log\det(\Sigma) - \frac{1}{2}\log\det(\mathcal{X}_{\widetilde{Z}}^T\mathcal{X}_{\widetilde{Z}}) + \log|\bar{\mathcal{Z}}_{\widetilde{\tau}}| \\
&\quad + \sqrt{2\beta}\sqrt{\|\mathcal{X}_{\widetilde{Z}}\mu\|^2} + \sqrt{2\beta}\sqrt{\mathsf{Tr}(\mathcal{X}_{\widetilde{Z}}\Sigma\mathcal{X}_{\widetilde{Z}}^T)} \\
&\quad + \frac{1}{2}\|\mathcal{X}_{\widetilde{Z}}\mu - Y\|^2 + \frac{1}{2}\mathsf{Tr}(\mathcal{X}_{\widetilde{Z}}\Sigma\mathcal{X}_{\widetilde{Z}}^T) + \log\Gamma\left(\frac{\ell_{\widetilde{\tau}}+1}{2}\right) + \frac{N}{2}\log(2\pi) \\
&\overset{def}{=} L_1(\widetilde{\tau}, \widetilde{Z}, \Sigma, \mu) + \log\Gamma\left(\frac{\ell_{\widetilde{\tau}}+1}{2}\right) + \frac{N}{2}\log(2\pi).
\end{aligned}
$$

Now we assume $\Sigma = r(\mathcal{X}_{\widetilde{Z}}^T\mathcal{X}_{\widetilde{Z}})^{-1}$, then we denote

$$
\begin{aligned}
&h(r) \overset{def}{=} L_1(\widetilde{\tau}, \widetilde{Z}, r(\mathcal{X}_{\widetilde{Z}}^T\mathcal{X}_{\widetilde{Z}})^{-1}, \mu) \\
&= -\frac{\ell_{\widetilde{\tau}}}{2}\log(\beta e) - \frac{\ell_{\widetilde{\tau}}}{2}\log r + \log|\bar{\mathcal{Z}}_{\widetilde{\tau}}| + \sqrt{2\beta}\|\mathcal{X}_{\widetilde{Z}}\mu\| + \sqrt{2\beta\ell_{\widetilde{\tau}}r} + \frac{1}{2}\|\mathcal{X}_{\widetilde{Z}}\mu - Y\|^2 + \frac{\ell_{\widetilde{\tau}}r}{2}
\end{aligned}
$$

and it is easy to get that $h'(r) = -\frac{\ell_{\widetilde{\tau}}}{2r} + \sqrt{\frac{\beta\ell_{\widetilde{\tau}}}{2}}r^{-1/2} + \frac{1}{2}\ell_{\widetilde{\tau}}$, then the minimiser is obtained by setting $r^* = \frac{\sqrt{2\beta\ell_{\widetilde{\tau}}+4\ell_{\widetilde{\tau}}^2}-\sqrt{2\beta\ell_{\widetilde{\tau}}}}{2\ell_{\widetilde{\tau}}}$. Then, we have

$$
\sqrt{2\beta\ell_{\widetilde{\tau}}r^*} + \frac{\ell_{\widetilde{\tau}}r^*}{2} \leq \sqrt{2\beta\ell_{\widetilde{\tau}}r^*} + \ell_{\widetilde{\tau}}r^* = \ell_{\widetilde{\tau}},
$$

$$
\frac{1}{r^*} = \frac{\sqrt{2\beta\ell_{\widetilde{\tau}}+4\ell_{\widetilde{\tau}}^2}+\sqrt{2\beta\ell_{\widetilde{\tau}}}}{2\ell_{\widetilde{\tau}}} \leq \frac{2\sqrt{2\beta\ell_{\widetilde{\tau}}}+2\ell_{\widetilde{\tau}}}{2\ell_{\widetilde{\tau}}} = 1 + \sqrt{2\beta}\ell_{\widetilde{\tau}}^{-1/2},
$$

and

$$
-\frac{\ell_{\widetilde{\tau}}}{2}\log r^* \leq \frac{\ell_{\widetilde{\tau}}}{2}\log(1+\sqrt{2\beta}\ell_{\widetilde{\tau}}^{-1/2}) \leq \frac{\sqrt{2\beta}}{2}\ell_{\widetilde{\tau}}^{1/2} \leq \frac{\sqrt{2\beta}}{2}\ell_{\widetilde{\tau}}.
$$

Finally, we can obtain

$$
\begin{aligned}
&D\left(\widetilde{Q}_Z^{(\widetilde{\tau})}\widetilde{Q}_B^{(\widetilde{\tau})}\widetilde{Q}_\lambda^{(\widetilde{\tau})}\|\Pi^{(\widetilde{\tau})}(Z,B,\lambda)\right) - \int \log\phi_{\mathcal{X}_Z B}d\widetilde{Q}^{(\widetilde{\tau})}(Z,B) \\
&\leq L_1(\widetilde{\tau}, \widetilde{Z}, r^*(X_{\widetilde{Z}^T\widetilde{Z}})^{-1}, \mu) + \log\Gamma\left(\frac{\ell_{\widetilde{\tau}}+1}{2}\right) + \frac{N}{2}\log(2\pi)
\end{aligned}
$$

$$\leq \quad M_1\ell_{\widetilde{\tau}} + \log|\widetilde{Z}_{\widetilde{\tau}}| + \sqrt{2\beta}\|\mathcal{X}_{\widetilde{Z}}\mu\| + \frac{1}{2}\|X_{\widetilde{Z}}\mu - Y\|^2 + \log\Gamma\left(\frac{\ell_{\widetilde{\tau}}+1}{2}\right) + \frac{N}{2}\log(2\pi),$$

where $M_1 = \frac{1}{2}\left(\log(\frac{e}{\beta}) + \sqrt{2\beta}\right)$ is a constant.

Now we analyze the second part of (5.34),

$$\begin{aligned} & -\log\pi(\widetilde{\tau}) + \log p_\Pi(Y) \\ = \quad & \log\frac{\Gamma(\ell_{\widetilde{\tau}}/2)}{\Gamma(\ell_{\widetilde{\tau}})} + D\epsilon_{\widetilde{\tau}} - \frac{N}{2}\log(2\pi) + \log\sum_{\tau\in\mathcal{T}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\exp\left(-D\epsilon_\tau\right)\frac{1}{|\bar{\mathcal{Z}}_\tau|}M_Z, \end{aligned}$$

where

$$M_Z = \int \frac{\sqrt{\det(\mathcal{X}_Z^T\mathcal{X}_Z)}}{2}\left(\frac{2\beta}{\pi}\right)^{\frac{\ell_\tau}{2}}\exp\left(-\sqrt{2\beta}\|\mathcal{X}_Z B\| - \frac{1}{2}\|\mathcal{X}_Z B - Y\|^2\right)dB.$$

Combine the first part and second part of (14), we can get that there exists $\widetilde{\tau}$ and $\widetilde{Q}^{(\widetilde{\tau})} \in \mathcal{S}_{\mathrm{MF}}^{(\widetilde{\tau})}$.

$$\begin{aligned} & -F(\widetilde{Q}^{(\widetilde{\tau})}, \widetilde{\tau}) + \log p_\Pi(Y) \\ = \quad & D\left(\widetilde{Q}_Z^{(\widetilde{\tau})}\widetilde{Q}_B^{(\widetilde{\tau})}\widetilde{Q}_\lambda^{(\widetilde{\tau})}\|\Pi^{(\widetilde{\tau})}(Z,B,\lambda)\right) - \int\log\phi_{\mathcal{X}_Z B}d\widetilde{Q}^{(\widetilde{\tau})}(Z,B) - \log\pi(\widetilde{\tau}) + \log p_\Pi(Y) \\ \leq \quad & (M_1+1+D)\epsilon_{\widetilde{\tau}} + \log\frac{\Gamma\left(\frac{\ell_{\widetilde{\tau}}+1}{2}\right)\Gamma(\frac{\ell_{\widetilde{\tau}}}{2})}{\Gamma(\ell_{\widetilde{\tau}})} + \log\sum_{\tau\in\mathcal{T}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\exp\left(-D\epsilon_\tau\right)\frac{1}{|\bar{\mathcal{Z}}_\tau|}R(Z;\widetilde{\tau},\widetilde{Z},\mu) \\ \leq \quad & M_2\epsilon_{\widetilde{\tau}} + \log\sum_{\tau\in\mathcal{T}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\exp\left(-D\epsilon_\tau\right)\frac{1}{|\bar{\mathcal{Z}}_\tau|}R(Z;\widetilde{\tau},\widetilde{Z},\mu) \\ = \quad & \log\sum_{\tau\in\mathcal{T}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\exp\left(M_2\epsilon_{\widetilde{\tau}} - D\epsilon_\tau\right)\frac{1}{|\bar{\mathcal{Z}}_\tau|}R(Z;\widetilde{\tau},\widetilde{Z},\mu), \end{aligned}$$

where $M_2 = M_1 + 1 + D + \frac{1}{2}\log\pi$ and we have used the fact that $\frac{\Gamma\left(\frac{\ell_{\widetilde{\tau}}+1}{2}\right)\Gamma(\frac{\ell_{\widetilde{\tau}}}{2})}{\Gamma(\ell_{\widetilde{\tau}})} = \frac{\sqrt{\pi}}{2^{\ell_{\widetilde{\tau}}-1}}$. $R(Z;\widetilde{\tau},\widetilde{Z},\mu)$ above is defined as the following integral.

$$R(Z;\widetilde{\tau},\widetilde{Z},\mu) = \int\frac{\sqrt{\det(\mathcal{X}_Z^T\mathcal{X}_Z)}}{2}\left(\frac{2\beta}{\pi}\right)^{\frac{\ell_\tau}{2}} \tag{5.38}$$

$$\times \quad \exp\left(\sqrt{2\beta}(\|\mathcal{X}_{\widetilde{Z}}\mu\| - \|\mathcal{X}_Z B\|) - \frac{1}{2}\left(\|\mathcal{X}_Z B - Y\|^2 - \|\mathcal{X}_{\widetilde{Z}}\mu - Y\|^2\right)\right) dB.$$

Now for the given $\delta_1 > 0$, τ^*, $Z^* \in \bar{\mathcal{Z}}_{\tau^*}$ and $B^* \in \mathbb{R}^{\ell_{\tau^*}}$, we choose $(\widetilde{\tau}, \widetilde{Z}, \mu)$ by

$$(\widetilde{\tau}, \widetilde{Z}, \mu) = \underset{\substack{(\tau,Z,U)\\ \tau\in\mathcal{A}}}{\operatorname{argmin}} \left\{ \sqrt{2\beta}\|\mathcal{X}_Z U\| + \frac{1}{2}\|\mathcal{X}_Z U - Y\|^2 + M_2\epsilon_\tau \right\}, \tag{5.39}$$

where $\mathcal{A} = \left\{\tau \in \mathcal{T} : \epsilon_\tau \le \epsilon_{\tau^*} + \delta_1 M_3 \|\mathcal{X}_{Z^*}B^* - \theta^*\|^2\right\}$ and $M_3 > 0$ will be determined later.

- For $\tau \in \mathcal{A}$, with Pythagorean theorem, we can decompose $\|\mathcal{X}_Z B - Y\|^2$ by

$$\|\mathcal{X}_Z B - Y\|^2 = \|\mathcal{X}_Z B - \mathcal{X}_Z B_{Z,Y}\|^2 + \|\mathcal{X}_Z B_{Z,Y} - Y\|^2,$$

  where $B_{Z,Y} = (\mathcal{X}_Z^T\mathcal{X}_Z)^{-1}\mathcal{X}_Z^T Y \in \mathbb{R}^{\ell_\tau}$ and $\mathcal{X}_Z B_Y$ is the projection of Y on the subspace spanned by $\mathcal{X}_Z$. Then according to the definition of $(\widetilde{\tau}, \widetilde{Z}, \mu)$, we have

$$\begin{aligned}
& \sqrt{2\beta}(\|\mathcal{X}_{\widetilde{Z}}\mu\| - \|\mathcal{X}_Z B\|) - \frac{1}{2}(\|\mathcal{X}_Z B - Y\|^2 - \|\mathcal{X}_{\widetilde{Z}}\mu - Y\|^2) + M_2\epsilon_{\widetilde{\tau}} \\
\le\ & \sqrt{2\beta}(\|\mathcal{X}_Z B_{Z,Y}\| - \|\mathcal{X}_Z B\|) - \frac{1}{2}(\|\mathcal{X}_Z B - Y\|^2 - \|\mathcal{X}_Z B_{Z,Y} - Y\|^2) + M_2\epsilon_\tau \\
\le\ & \sqrt{2\beta}\|\mathcal{X}_Z(B - B_{Z,Y})\| - \frac{1}{2}\|\mathcal{X}_Z(B - B_{Z,Y})\|^2 + M_2\epsilon_\tau \\
\le\ & -\frac{1}{4}\|\mathcal{X}_Z(B - B_{Z,Y})\|^2 + 2\beta + M_2\epsilon_\tau.
\end{aligned}$$

  Then we assume that $\mathcal{X}_Z = U\Sigma V^T$ as its SVD and set $b = \Sigma V^T(B - B_{Z,Y})$, then

$$\begin{aligned}
& \exp(M_2\epsilon_{\widetilde{\tau}})\, R(Z; \widetilde{\tau}, \widetilde{Z}, \mu) \\
\le\ & \exp(M_2\epsilon_\tau) \int \frac{\sqrt{\det(\mathcal{X}_Z^T\mathcal{X}_Z)}}{2} \left(\frac{2\beta}{\pi}\right)^{\ell_\tau/2} \exp\left(-\frac{1}{4}\|\mathcal{X}_Z(B - B_{Z,Y})\|^2 + 2\beta\right) dB \\
=\ & \frac{1}{2}\exp(M_2\epsilon_\tau)\left(\frac{2\beta}{\pi}\right)^{\ell_\tau/2} e^{2\beta} \int \exp\left(-\frac{1}{4}\|b\|^2\right) db \\
=\ & \frac{1}{2}\exp(M_2\epsilon_\tau)\, e^{2\beta}(8\beta)^{\ell_\tau/2}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\sum_{\tau\in\mathcal{A}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\exp\left(M_2\epsilon_{\widetilde{\tau}}-D\epsilon_\tau\right)\frac{1}{|\bar{\mathcal{Z}}_\tau|}R(Z;\widetilde{\tau},\widetilde{Z},\mu) \qquad (5.40)\\
\leq\ & \frac{1}{2}e^{2\beta}\sum_{\epsilon_\tau\leq\epsilon_{\tau^*}+\delta_1 M_3\|\mathcal{X}_{Z^*}B^*-\theta^*\|^2}\exp\left((M_2-D+C_1)\epsilon_\tau\right)\\
\leq\ & 4\exp\Big((M_2-D+C_1+1)(\epsilon_{\tau^*}+\delta_1 M_3\|\mathcal{X}_{Z^*}B^*-\theta^*\|^2)\Big)\\
\leq\ & 4\exp\Big((M_1+C_1+3)(\epsilon_{\tau^*}+\delta_1 M_3\|\mathcal{X}_{Z^*}B^*-\theta^*\|^2)\Big). \qquad (5.41)
\end{aligned}$$

For the last inequality, we apply Lemma 7.2 in [28].

- For $\tau\notin\mathcal{A}$, note that $\tau^*\in\mathcal{A}$ and then

$$\begin{aligned}
&\sqrt{2\beta}(\|\mathcal{X}_{\widetilde{Z}}\mu\|-\|\mathcal{X}_Z B\|)-\frac{1}{2}\left(\|\mathcal{X}_Z B-Y\|^2-\|\mathcal{X}_{\widetilde{Z}}\mu-Y\|^2\right)+M_2\epsilon_{\widetilde{\tau}}\\
\leq\ & \sqrt{2\beta}(\|\mathcal{X}_{Z^*}B^*\|-\|\mathcal{X}_Z B\|)-\frac{1}{2}\left(\|\mathcal{X}_Z B-Y\|^2-\|\mathcal{X}_{Z^*}B^*-Y\|^2\right)+M_2\epsilon_{\tau^*}\\
\leq\ & \sqrt{2\beta}\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2+M_2\epsilon_{\tau^*}\\
&-\frac{1}{2}\left(\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2+2\left\langle\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*,\mathcal{X}_{Z^*}B^*-\theta^*-W\right\rangle\right)\\
\leq\ & \sqrt{2\beta}\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|-\frac{1}{2}\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\\
&+\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|\|\mathcal{X}_{Z^*}B^*-\theta^*\|+|\left\langle\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*,W\right\rangle|+M_2\epsilon_{\tau^*}\\
\leq\ & 4\beta+\frac{1}{8}\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2-\frac{1}{2}\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2+\frac{1}{8}\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\\
&+2\|\mathcal{X}_{Z^*}B^*-\theta^*\|^2+|\left\langle\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*,W\right\rangle|+M_2\epsilon_{\tau^*}\\
=\ & 4\beta-\frac{1}{4}\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2+2\|\mathcal{X}_{Z^*}B^*-\theta^*\|^2\\
&+|\left\langle\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*,W\right\rangle|+M_2\epsilon_{\tau^*}.
\end{aligned}$$

For any $t>0$, assume

$$\begin{aligned}
&E_Z(t)\\
=\ & \left\{W:|\left\langle\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*,W\right\rangle|\leq\sqrt{\epsilon_\tau^*+t}\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|,\ \text{for all } B\in\mathbb{R}^{\ell_\tau}\right\},
\end{aligned}$$

where $\epsilon^*_\tau = C_1\epsilon_\tau + C_2\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2$ and $C_1 > 1$, $C_2 > 0$. Then according to Lemma 7.1 in [28], for W satisfies condition (3.13),

$$P_{\theta^*}(E_Z(t)^c) \le 2\exp\left(-(\rho C_1/16 - 5)\epsilon_\tau - \rho C_2\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2/16 - \rho t/16\right). \quad (5.42)$$

Then under the event $\cap_{\tau\notin\mathcal{A}}\cap_{Z\in\bar{\mathcal{Z}}_\tau} E_Z(t)$,

$$\begin{aligned}
& \sqrt{2\beta}(\|\mathcal{X}_{\widetilde{Z}}\mu\| - \|\mathcal{X}_Z B\|) - \frac{1}{2}\left(\|\mathcal{X}_Z B - Y\|^2 - \|\mathcal{X}_{\widetilde{Z}}\mu - Y\|^2\right) + M_2\epsilon_{\widetilde{\tau}} \\
\le\ & 4\beta - \frac{1}{4}\|\mathcal{X}_Z B - \mathcal{X}_{Z^*}B^*\|^2 + 2\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 \\
& + \sqrt{\epsilon^*_\tau + t}\|\mathcal{X}_Z B - \mathcal{X}_{Z^*}B^*\| + M_2\epsilon_{\tau^*} \\
\le\ & 4\beta - \frac{1}{4}\|\mathcal{X}_Z B - \mathcal{X}_{Z^*}B^*\|^2 + 2\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + \frac{1}{8}\|\mathcal{X}_Z B - \mathcal{X}_{Z^*}B^*\|^2 \\
& + 2(\epsilon^*_\tau + t) + M_2\epsilon_{\tau^*} \\
=\ & 4\beta - \frac{1}{8}\|\mathcal{X}_Z B - \mathcal{X}_{Z^*}B^*\|^2 + 2(C_2+1)\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + 2C_1\epsilon_\tau + 2t + M_2\epsilon_{\tau^*}.
\end{aligned}$$

Therefore, under the event $\cap_{\tau\notin\mathcal{A}}\cap_{Z\in\bar{\mathcal{Z}}_\tau} E_Z(t)$, with the same argument as above,

$$\begin{aligned}
& \exp\left(M_2\epsilon_{\widetilde{\tau}}\right) R(Z;\widetilde{\tau},\widetilde{Z},\mu) \\
\le\ & \exp\left(4\beta + 2(C_2+1)\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + 2C_1\epsilon_\tau + M_2\epsilon_{\tau^*} + 2t\right) \\
& \times \int \frac{\sqrt{\det(\mathcal{X}_Z^T\mathcal{X}_Z)}}{2}\left(\frac{2\beta}{\pi}\right)^{\ell_\tau/2}\exp\left(-\frac{1}{8}\|\mathcal{X}_Z B - \mathcal{X}_{Z^*}B^*\|^2\right) dB \\
=\ & \frac{(16\beta)^{\ell_\tau/2}}{2}\exp\left(4\beta + 2(C_2+1)\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + 2C_1\epsilon_\tau + M_2\epsilon_{\tau^*} + 2t\right) \\
\le\ & \frac{1}{2}e^{4\beta}\exp\left(M_4\epsilon_\tau + 2(C_2+1)\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + M_2\epsilon_{\tau^*} + 2t\right),
\end{aligned}$$

where $M_4 = \frac{1}{2}\max\{\log(16\beta), 0\} + 2C_1 > 0$. Thus, if $D > M_4 + 2$, the second part of the summation is upper bound as following:

$$\sum_{\tau\notin\mathcal{A}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\exp\left(M_2\epsilon_{\widetilde{\tau}} - D\epsilon_\tau\right)\frac{1}{|\bar{\mathcal{Z}}_\tau|}R(Z;\widetilde{\tau},\widetilde{Z},\mu)$$

$$\leq \frac{1}{2}\exp\left(M_2\epsilon_{\tau^*} + 4\beta + 2(C_2+1)\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + 2t\right)\sum_{\tau\notin\mathcal{A}}\exp\left(-(D-M_4)\epsilon_\tau\right).$$

Set $\alpha = \epsilon_{\tau^*} + \delta_1\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2$ and $\beta = D - M_4$ in Lemma 7.2 of [28], then

$$\begin{aligned}
&\sum_{\tau\notin\mathcal{A}}\exp\left(-(D-M_4)\epsilon_\tau\right)\\
\leq\ & 4\exp\left(-(\beta-1)\alpha\right) = 4\exp\left(-(D-M_4-1)(\epsilon_{\tau^*} + \delta_1\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2)\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{\tau\notin\mathcal{A}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\exp\left(M_2\epsilon_{\widetilde{\tau}} - D\epsilon_\tau\right)\frac{1}{|\bar{\mathcal{Z}}_\tau|}R(Z;\widetilde{\tau},\widetilde{Z},\mu)\\
\leq\ & 2\exp\left(4\beta+2t\right)\exp\left((M_2 - D + M_4 + 1)\epsilon_{\tau^*} - [(D-M_4-1)\delta_1 - 2(C_2+1)]\right)\\
\leq\ & 2\exp\left(4\beta+2t\right)\exp\left((M_1 + M_4 + 3)\epsilon_{\tau^*} - [(D-M_4-1)\delta_1 - 2(C_2+1)]\right).
\end{aligned}$$

Then for $D > M_4 + 1 + 2\delta_1^{-1}(C_2+1)$,

$$\sum_{\tau\notin\mathcal{A}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\exp\left(-D\epsilon_\tau\right)\frac{1}{|\bar{\mathcal{Z}}_\tau|}R(Z;\widetilde{\tau},\widetilde{Z},\mu) \leq 2\exp\left(4\beta + 2t + (M_1+M_4+3)\epsilon_{\tau^*}\right). \tag{5.43}$$

Combine (5.57) and (5.56), when $D > D_0$ for $D_0 = M_4 + 1 + 2\delta_1^{-1}(C_2+1)$ and under the event $\cap_{\tau\notin\mathcal{A}}\cap_{Z\in\bar{\mathcal{Z}}_\tau} E_Z(t)$,

$$\begin{aligned}
&\log\sum_{\tau\in\mathcal{T}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\exp\left(M_2\epsilon_{\widetilde{\tau}} - D\epsilon_\tau\right)\frac{1}{|\bar{\mathcal{Z}}_\tau|}R(Z;\widetilde{\tau},\widetilde{Z},\mu)\\
\leq\ & \log\Big(4\exp\left((M_1+C_1+3)(\epsilon_{\tau^*} + \delta_1 M_3\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2)\right)\\
&+2\exp\left(4\beta+2t+(M_1+M_4+3)\epsilon_{\tau^*}\right)\Big)\\
\leq\ & \log 6 + 4\beta + 2t + M\epsilon_{\tau^*} + (M_1+C_1+3)M_3\delta_1\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2\\
\leq\ & C\epsilon_{\tau^*} + (M_1+C_1+3)M_3\delta_1\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + 2t,
\end{aligned}$$

where $M = \max\{M_1+C_1+3, M_1+M_4+3\}$ and $C = M+\log 6+4\beta$. Set $M_3 = (M_1+C_1+3)^{-1}$, then under the event $\cap_{\tau\notin\mathcal{A}} \cap_{Z\in\bar{\mathcal{Z}}_\tau} E_Z(t)$,

$$\log \sum_{\tau\in\mathcal{T}} \sum_{Z\in\bar{\mathcal{Z}}_\tau} \exp\left(M_2\epsilon_{\widetilde{\tau}} - D\epsilon_\tau\right) \frac{1}{|\bar{\mathcal{Z}}_\tau|} R(Z;\widetilde{\tau},\widetilde{Z},\mu) \le C\epsilon_{\tau^*} + \delta_1\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + 2t. \quad (5.44)$$

To sum up, for $D > D_0$, there exists a $\widetilde{\tau}$, $\widetilde{Q}^{(\widetilde{\tau})} \in \mathcal{S}_{\mathrm{MF}}^{(\widetilde{\tau})}$ such that

$$\begin{aligned} & P_{\theta^*}\left(-F(\widetilde{Q}^{(\widetilde{\tau})},\widetilde{\tau}) + \log p_\Pi(Y) > C\epsilon_{\tau^*} + \delta_1\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + 2t\right) \\ \le & \sum_{\tau\notin\mathcal{A}} \sum_{Z\in\bar{\mathcal{Z}}_\tau} P_{\theta^*}(E_Z(t)^c) \\ \le & \ 2\exp\left(-\rho C_2\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2/16 - \rho t/16\right) \sum_{\tau\notin\mathcal{A}} \sum_{Z\in\bar{\mathcal{Z}}_\tau} \exp\left(-(\rho C_1/16 - 5)\epsilon_\tau\right). \end{aligned}$$

For $C_1 > 192/\rho$, we have $\sum_{\tau\notin\mathcal{A}}\sum_{Z\in\bar{\mathcal{Z}}} \exp\left(-(\rho C_1/16-5)\epsilon_\tau\right) \le \sum_{\tau\notin\mathcal{A}} \exp\left(-6\epsilon_\tau\right) \le 6$. Then

$$P_{\theta^*}\left(-F(\widetilde{Q}^{(\widetilde{\tau})},\widetilde{\tau}) + \log p_\Pi(Y) > C\epsilon_{\tau^*} + \delta_1\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + 2t\right) \le 12\exp\left(-\rho t/16\right). \quad (5.45)$$

Finally, we have

$$\begin{aligned} & P_{\theta^*}\left[-F(\widetilde{Q}^{(\widetilde{\tau})},\widetilde{\tau}) + \log p_\Pi(Y)\right] \\ \le & \ C\epsilon_{\tau^*} + \delta_1\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + \int_0^\infty 12\exp\left(-\rho t/32\right) \\ \le & \ C'\epsilon_{\tau^*} + \delta_1\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2, \end{aligned} \quad (5.46)$$

where $C' = C + 384/\rho$. □

Proof of Lemma 5.2.4. For Lemma 5.2.4, the proof of Theorem 4.1 in [28] directly indicates that for any θ^*, τ^*, $Z^* \in \bar{\mathcal{Z}}_{\tau^*}$ and $B^* \in \mathbb{R}^{\ell_{\tau^*}}$, $\delta_2 > 0$, there exists D_0 only depending on β,

δ_2, ρ, such that

$$\begin{aligned} & P_{\theta^*}\Pi\left(\|\mathcal{X}_Z B - \theta^*\|^2 > (1+\delta_2)\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + M\epsilon_{\tau^*} + Mt \Big| Y\right) \\ \leq \;& \exp\left(-C(\epsilon_{\tau^*} + \|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + t)\right), \end{aligned}$$

for any $D > D_0$, $t > 0$ and some constants M, C only depending on β, δ_2, ρ, D. Then for $T = \|\mathcal{X}_Z B - \theta^*\|^2 - (1+\delta_2)\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 - M\epsilon_{\tau^*}$, we have

$$\begin{aligned} & P_{\theta^*}\Pi\left[\exp\left(a\|\mathcal{X}_Z B - \theta^*\|^2\right)\Big| Y\right] \\ \leq \;& \exp\left(a\left[(1+\delta_2)\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + M\epsilon_{\tau^*}\right]\right) P_{\theta^*}\Pi\left[\exp\left(aT\right)|Y\right], \end{aligned}$$

and

$$\begin{aligned} & P_{\theta^*}\Pi\left[\exp\left(aT\right)|Y\right] \\ \leq \;& P_{\theta^*}\Pi\left[\exp\left(aT\mathbb{I}\{T>0\}\right)|Y\right] \leq 1 + \int_1^\infty P_{\theta^*}\Pi\left[\exp\left(aT\right) > t|Y\right] dt \\ \leq \;& 1 + a\int_0^\infty P_{\theta^*}\Pi\left[T > r|Y\right]\exp\left(ar\right) dr \\ \leq \;& 1 + a\int_0^\infty \exp\left(-\left(\frac{C}{M} - a\right)r\right) dr. \end{aligned}$$

Choose $a = \frac{C}{2M}$, then

$$P_{\theta^*}\Pi\left[\exp\left(aT\mathbb{I}\{T>0\}\right)|Y\right] \leq 3,$$

and

$$P_{\theta^*}\Pi\left[\exp\left(a\|\mathcal{X}_Z B - \theta^*\|^2\right)\Big| Y\right] \leq 3\exp\left(a\left[(1+\delta_2)\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + M\epsilon_{\tau^*}\right]\right).$$

Finally,

$$P_{\theta^*}\log\Pi\left(\exp\left(a\|\mathcal{X}_Z B - \theta^*\|^2\right)\Big| Y\right) \leq \log P_{\theta^*}\Pi\left(\exp\left(a\|\mathcal{X}_Z B - \theta^*\|^2\right)\Big| Y\right)$$

$$\leq \quad a(1+\delta_2)\|\mathcal{X}_{Z^*}B^* - \theta^*\|^2 + M'\epsilon_{\tau^*} \tag{5.47}$$

where $M' = M + \log 3$. □

Proof of Theorem 3.2.1. The proof of Theorem 3.2.1 can be directly derived from Lemma 5.2.2, Lemma 5.2.3 and Lemma 5.2.4. □

5.2.3 Derivations of All Algorithms

In this section, we provides the derivations for the general algorithm and the specific algorithms in Theorem 5.2.1 and 5.2.2 respectively.

Theorem 5.2.1. *The algorithm derived by (3.21), (3.18) and (3.19), can be explicitly expressed in Algorithm 2 and Algorithm 3.*

Proof of Theorem 5.2.1. Assume $Q_B = N(\mu, \Sigma)$, Q_Z is the measurement such that $Q_Z(Z = \widetilde{Z}) = 1$ and

$$\frac{dQ_\lambda}{d\lambda} = \sqrt{\frac{\beta}{\pi}}\lambda^{-3/2}\exp\left(\sqrt{2a\beta} - \frac{\beta}{\lambda} - \frac{a\lambda}{2}\right).$$

Then according to (5.37),

$$\begin{aligned}
& L(\tau, \widetilde{Z}, a, \mu, \Sigma) \\
= \quad & \frac{1}{2}Q_BQ_Z\|\mathcal{X}_Z B - Y\|^2 + D\left(Q_BQ_ZQ_\lambda\|\Pi^{(\tau)}(B, Z, \lambda)\right) - \log\pi(\tau) \\
= \quad & -\frac{\ell_\tau}{2}\log(\beta e) - \frac{1}{2}\log\det(\Sigma) - \frac{1}{2}\log\det(\mathcal{X}_{\widetilde{Z}}^T\mathcal{X}_{\widetilde{Z}}) + \log|\tilde{\mathcal{Z}}_\tau| + \sqrt{\frac{a\beta}{2}} \\
& +\sqrt{\frac{\beta}{2a}}\left(\|\mathcal{X}_{\widetilde{Z}}\mu\|^2 + \mathsf{Tr}(\mathcal{X}_{\widetilde{Z}}\Sigma\mathcal{X}_{\widetilde{Z}}^T)\right) \\
& +\frac{1}{2}\|\mathcal{X}_{\widetilde{Z}}\mu - Y\|^2 + \frac{1}{2}\mathsf{Tr}(\mathcal{X}_{\widetilde{Z}}\Sigma\mathcal{X}_{\widetilde{Z}}^T) + \log\frac{\Gamma\left(\frac{\ell_\tau+1}{2}\right)\Gamma\left(\frac{\ell_\tau}{2}\right)}{\Gamma(\ell_\tau)} + D\epsilon_\tau + \text{const.}
\end{aligned} \tag{5.48}$$

The "const" above refers to a constant not depending on τ, μ, Σ, $\widetilde{Z}$, a. Then by some simple computations we can get the solution of (3.21), (3.19), (3.18) are, (3.22), (3.23) and (3.24).

Finally, assume $(Z^{[t]}, a^{[t]}, \mu^{[t]}, \Sigma^{[t]})$ to be the parameters after t-th iteration, then

$$\widehat{\mu}^{[t]} = \left(1+\sqrt{\frac{2\beta}{\widehat{a}^{[t]}}}\right)^{-1} (X_{\widetilde{Z}^{[t]}}^T X_{\widetilde{Z}^{[t]}})^{-1} X_{\widetilde{Z}^{[t]}}^T Y, \quad \widehat{\Sigma}^{[t]} = \left(1+\sqrt{\frac{2\beta}{\widehat{a}^{[t]}}}\right)^{-1} (X_{\widetilde{Z}^{[t]}}^T X_{\widetilde{Z}^{[t]}})^{-1}. \tag{5.49}$$

Plug (5.49) into (5.48) and with the fact that $\frac{\Gamma\left(\frac{\ell_\tau+1}{2}\right)\Gamma(\frac{\ell_\tau}{2})}{\Gamma(\ell_\tau)} = \frac{\sqrt{\pi}}{2^{\ell_\tau-1}}$, we have

$$\begin{aligned}
& L(\tau, \mu^{[t]}, \Sigma^{[t]}, \widehat{Z}^{[t]}, \widehat{a}^{[t]}) \\
= & -\frac{\ell_\tau}{2}\log(\beta e) + \frac{\ell_\tau}{2}\log\left(1+\sqrt{\frac{2\beta}{\widehat{a}^{[t]}}}\right) + \log|\bar{\mathcal{Z}}_\tau| + \sqrt{\frac{\widehat{a}^{[t]}\beta}{2}} \\
& + \sqrt{\frac{\beta}{2\widehat{a}^{[t]}}}\left(\|X_{\widehat{Z}^{[t]}}\widehat{\mu}^{[t]}\|^2 + \frac{\ell_\tau}{2}\left(1+\sqrt{\frac{2\beta}{\widehat{a}^{[t]}}}\right)^{-1}\right) \\
& + \frac{1}{2}\|X_{\widehat{Z}^{[t]}}\mu^{[t]} - Y\|^2 + \frac{\ell_\tau}{2}\left(1+\sqrt{\frac{2\beta}{\widehat{a}^{[t]}}}\right)^{-1} + \ell_\tau\log\frac{1}{2} + D\epsilon_\tau + \text{const} \\
= & \frac{\ell_\tau}{2}\log\frac{\widehat{\delta}^{[t]}}{4\beta} + \log|\widetilde{Z}_\tau| + \sqrt{\frac{\widehat{a}^{[t]}\beta}{2}} - \frac{1}{2}Y^T X_{\widehat{Z}^{[t]}}\widehat{\mu}^{[t]} + D\epsilon_\tau + \text{const},
\end{aligned}$$

where $\widehat{\delta}^{[t]} = 1+\sqrt{\frac{2\beta}{\widehat{a}^{[t]}}}$. The proof is complete. □

Theorem 5.2.2. *Algorithm 2 specialized for stochastic block model, biclustering model, sparse linear regression model, multiple regression with group sparsity, multi-task learning and dictionary learning is given in Algorithm 4, 5, 6, 7, 8 and 9.*

Proof of Theorem 5.2.2. We organize the proof from models to models:

- **Stochastic Block Model.**

  In this model $\mathcal{X}_Z = Z \otimes Z$ and $Y = \text{Vec}(A)$, then

$$\mathcal{X}_Z^T \mathcal{X}_Z = \text{diag}\left((n_c n_d)_{1\le c\le k, 1\le d\le k}\right),$$

$$\mathcal{X}_Z^T Y = \left(\sum_{i=1}^n\sum_{j=1}^n A_{ij}\mathbb{I}\{z_i = c, z_j = d\}\right)_{1\le c\le k, 1\le d\le k},$$

and $\mathsf{Tr}\left(\mathcal{X}_Z^T\mathcal{X}_Z\right) = \left(\sum_{c=1}^k n_c\right)^2 = n^2$, then the initialization and update for a, μ and Σ are correct in Algorithm 4.

Now we analyze the update procedure for Z. According to (3.22), for each $1 \le i \le n$ and each $1 \le c \le k$, we need to compute

$$\begin{aligned} v_{ic} &= -\frac{1}{2}\log\det\left(\mathcal{X}_{Z_{z_i}(c)}^T\mathcal{X}_{Z_{z_i}(c)}\right) - Y^T\mathcal{X}_{Z_{z_i}(c)}\mu \\ &\quad +\frac{1}{2}\left[\left\|\mathcal{X}_{Z_{z_i}(c)}\mu\right\|^2 + \mathsf{Tr}\left(\mathcal{X}_{Z_{z_i}(c)}\Sigma\mathcal{X}_{Z_{z_i}(c)}^T\right)\right], \end{aligned}$$

and update z_i by $z_i = \mathrm{argmin}_c\, v_{ic}$. As this procedure only rely on i and c, we can treat all irrelevant terms as constants. Then,

$$\log\det\left(\mathcal{X}_{Z_{z_i}(c)}^T\mathcal{X}_{Z_{z_i}(c)}\right) = 2\sum_{d\neq c}\log n_d(i) + 2\log(n_c(i)+1) = 2\log\left(1+n_c(i)^{-1}\right) + \text{const.}$$

$$Y^T\mathcal{X}_{Z_{z_i}(c)}\mu = \sum_{j_1\neq i}\sum_{j_2\neq i} A_{j_1j_2}\mu_{z_{j_1}z_{j_2}} + 2\sum_{j\neq i} A_{ij}\mu_{cz_j} + A_{ii}\mu_{cc} = 2\sum_{j\neq i} A_{ij}\mu_{cz_j} + \text{const.}$$

$$\left\|\mathcal{X}_{Z_{z_{ij}(c)}}\mu\right\|^2 = \sum_{j_1\neq i}\sum_{j_2\neq i}\mu_{z_{j_1}z_{j_2}}^2 + 2\sum_{j\neq i}\mu_{cz_j}^2 + \mu_{cc}^2 = 2\sum_{j\neq i}\mu_{cz_j}^2 + \mu_{cc}^2 + \text{const.}$$

$$\delta^{[t-1]}\,\mathsf{Tr}\left(\mathcal{X}_{Z_{z_i}(c)}\Sigma^{[t-1]}\mathcal{X}_{Z_{z_i}(c)}^T\right) = \left(\sum_{d\neq c}\frac{n_d(i)}{n_d^{[t-1]}} + \frac{n_c(i)+1}{n_c^{[t-1]}}\right)^2.$$

Thus, in SBM, Z should be updated by the procedure in Algorithm 4.

- **Biclustering Model**

  In this model $\mathcal{X}_Z = Z_1 \otimes Z_2$ and Y is replaced by $\mathrm{Vec}(Y)$, then

$$\mathcal{X}_Z^T\mathcal{X}_Z = \mathrm{diag}\left((n_cm_d)_{1\le c\le k, 1\le d\le l}\right),$$

$$\mathcal{X}_Z^TY = \left(\sum_{i=1}^n\sum_{j=1}^n Y_{ij}\mathbb{I}\{z_{1i}=c, z_{2j}=d\}\right)_{1\le c\le k, 1\le d\le l},$$

and $\mathsf{Tr}\left(\mathcal{X}_Z^T\mathcal{X}_Z\right) = \left(\sum_{c=1}^k n_c\right)\left(\sum_{d=1}^l m_c\right) = mn$, then the initialization and update for a, μ and Σ are correct in Algorithm 4.

For the procedure to update Z. When updating z_{1i}, z_{2j} for $1 \le j \le n$ are not updated, then we have

$$\log\det\left(\mathcal{X}_{Z_{z_i}(c)}^T\mathcal{X}_{Z_{z_i}(c)}\right) = \sum_{d\neq c}\log n_d(i) + \log(n_c(i)+1) = \log\left(1+n_c(i)^{-1}\right) + \text{const.}$$

$$Y^T\mathcal{X}_{Z_{z_{1i}}(c)}\mu = \sum_{i_1\neq i}\sum_{j=1}^m Y_{i_1 j}\mu_{z_{1i_1}z_{2j}} + \sum_{j=1}^m Y_{ij}\mu_{cz_{2j}} = \sum_{j=1}^m Y_{ij}\mu_{cz_{2j}} + \text{const.}$$

$$\left\|\mathcal{X}_{Z_{z_{1i}}(c)}\mu\right\|^2 = \sum_{i_1\neq i}\sum_{j=1}^n \mu^2_{z_{1i_1}z_{2j}} + \sum_{j=1}^n \mu^2_{cz_{2j}} = \sum_{j=1}^n \mu^2_{cz_{2j}} + \text{const.}$$

$$\delta^{[t-1]}\,\mathsf{Tr}\left(\mathcal{X}_{Z_{z_{1i}}(c)}\Sigma^{[t-1]}\mathcal{X}^T_{Z_{z_{1i}}(c)}\right) = l\left(\sum_{c_1\neq c}\frac{n_{c_1}(i)}{n_{c_1}^{[t-1]}} + \frac{n_c(i)+1}{n_c^{[t-1]}}\right) = \frac{1}{n_c^{[t-1]}} + \text{const.}$$

Then the update procedure for z_{1i} for $1 \le i \le n$ is correct in Algorithm 5. All results are symmetric for updating z_{2j} except the trace term as in this case, we have update the labels z_{1i}. Therefore,

$$\delta^{[t-1]}\,\mathsf{Tr}\left(X_{Z_{z_{2j}}(d)}\Sigma^{[t-1]}X^T_{Z_{z_{1i}}(c)}\right) = \left(\sum_{d_1\neq d}\frac{m_{d_1}(j)}{m_{d_1}^{[t-1]}} + \frac{m_d(i)+1}{m_d^{[t-1]}}\right)\left(\sum_{c=1}^k\frac{n_c^{[t]}}{n_c^{[t-1]}}\right)$$

$$= \frac{1}{m_d^{[t-1]}}\sum_{c=1}^k\frac{n_c^{[t]}}{n_c^{[t-1]}} + \text{const.}$$

- **Sparse Linear Regression** In the sparse linear regression, $\mathcal{X}_Z = X_Z = (X_{z_1}, \cdots, X_{z_s})$, the initialization and update for a, μ and Σ are automatically consistent in Algorithm

6. Then write $\mathcal{X}_Z = (X_{Z(-j)}, X_{z_j})$, we have

$$\mathcal{X}_Z^T \mathcal{X}_Z = \begin{pmatrix} X_{Z(-j)}^T X_{Z(-j)} & X_{Z(-j)}^T X_{z_j} \\ X_{z_j}^T X_{Z(-j)} & \|X_{z_j}\|^2 \end{pmatrix},$$

By Schur complement theorem, we have

$$\begin{aligned} \log\det\left(X_{Z_{z_i}(c)}^T X_{Z_{z_i}(c)}\right) &= \log\det\left(X_{Z(-j)}^T X_{Z(-j)}\right) + \log\left(\|X_c\|^2 - X_c^T H_{Z(-j)} X_c\right) \\ &= \log\left(\|X_c\|^2 - X_c^T H_{Z(-j)} X_c\right) + \text{const.} \end{aligned}$$

For the rest terms, we can get

$$Y^T X_{Z_{z_i}(c)} \mu = Y^T (X_{Z(-j)} \mu_{-j} + \mu_c X_c) = \mu_c Y^T X_c + \text{const.}$$

$$\begin{aligned} \|X_{Z_{z_i}(c)} \mu\|^2 &= \|X_{Z(-j)} \mu_{-j}\|^2 + 2\mu_j X_c^T X_{Z(-j)} \mu_{-j} + \|X_c\|^2 \\ &= 2\mu_j X_c^T X_{Z(-j)} \mu_{-j} + \|X_c\|^2 + \text{const} \end{aligned}$$

$$\begin{aligned} \mathsf{Tr}\left(\Sigma X_{Z_{z_i}(c)}^T X_{Z_{z_i}(c)}\right) &= \mathsf{Tr}\left(\Sigma_{-j,-j} X_{Z(-j)}^T X_{Z(-j)}\right) + 2\Sigma_{j,-j} X_{Z(-j)}^T X_c + \Sigma_{j,j} \|X_c\|^2 \\ &= 2\Sigma_{j,-j} X_{Z(-j)}^T X_c + \Sigma_{j,j} \|X_c\|^2 + \text{const.} \end{aligned}$$

Therefore, the procedure to update Z for sparse linear regression is Algorithm 6.

- **Multiple Regression with Group Sparsity** In this case $\mathcal{X}_Z = I_m \otimes X_Z$, but the derivation is similar to that for sparse linear regression, so we omit it.

- **Multi-task Learning** In multi-task learning model, $\mathcal{X}_Z = Z \otimes X$ and Y is replaced

by $\mathrm{Vec}(Y)$, then

$$\mathcal{X}_Z^T \mathcal{X}_Z = \mathrm{diag}\left(\left(n_c X^T X\right)_{1\le c\le k}\right), \qquad \mathcal{X}_Z^T \mathrm{Vec}(Y) = \left(\sum_{i=1}^{m} X^T Y_{\cdot i} \mathbb{I}\left\{z_i = c\right\}\right)_{1\le c\le k}.$$

After reshaping the vectors to matrices, we can get the initialization and update for a, δ, μ and Σ for multi-task learning model is given in Algorithm 8. Then for the procedure to update Z, we have

$$\begin{aligned}\log\det\left(X_{Z_{z_i}(c)}^T X_{Z_{z_i}(c)}\right) &= k\log(X^T X) + p\sum_{d\neq c}\log(n_d(i)) + p\log(n_c(i)+1)\\ &= p\log\left(1+n_c(i)^{-1}\right) + \text{const.}\end{aligned}$$

$$\mathrm{Vec}(Y)^T X_{Z_{z_i}(c)} \mathrm{Vec}(\mu) = \sum_{i_0\neq i} Y_{\cdot i_0}^T X \mu_{z_{i_0}} + Y_{\cdot i}^T X \mu_c = Y_{\cdot i}^T X \mu_c + \text{const.}$$

$$\|X_{Z_{z_i}(c)} \mathrm{Vec}(\mu)\|^2 = \sum_{i_0\neq i} \|X\mu_{z_{i_0}}\|^2 + \|X\mu_c\|^2 = \|X\mu_c\|^2 + \text{const.}$$

$$\delta^{[t-1]}\,\mathsf{Tr}\left(X_{Z_{z_i}(c)} \Sigma^{[t-1]} X_{Z_{z_i}(c)}^T\right) = p\sum_{c_1\neq c}\frac{n_{c_1}(i)}{n_{c_1}^{[t-1]}} + p\frac{n_c(i)+1}{n_c^{[t-1]}} = \frac{p}{n_c^{[t-1]}} + \text{const.}$$

Then the update of Z for multi-task learning model should be the procedure given in Algorithm 8.

- **Dictionary Learning** In dictionary learning, $\mathcal{X}_Z = I_n \otimes Z^T$. B is replaced by $\mathrm{Vec}(B^T)$. Y is replaced by $\mathrm{Vec}(Y^T)$, then

$$\mathcal{X}_Z^T \mathcal{X}_Z = I_n \otimes (ZZ^T), \mathrm{Vec}(\mu^T) = \delta^{-1}(\mathcal{X}_Z^T \mathcal{X}_Z)^{-1} \mathcal{X}_Z^T \mathrm{Vec}(Y^T).$$

Therefore, $\mu^T = (ZZ^T)^{-1} Z Y^T$ and $\mu = Y Z^T (ZZ^T)^{-1}$. The initialization and update for a, δ, μ and Σ is consistent in Algorithm 9. The procedure of updating Z can be obtained by straightforward derivation from Algorithm 2.

The proof is complete. □

5.3 Proof in Chapter 4

5.3.1 Proofs of Theorem 4.2.1, Corollary 4.2.1 and Theorem 4.2.2

We first show the following lemma to assist the proof of Theorem 4.2.1.

Lemma 5.3.1. *The variational posterior $\widehat{Q}$ with respect to the set $\mathcal{S}_{\mathrm{MF}}$ is a product measure, with the density for each coordinate in the form of*

$$q_j(\theta_j) = \begin{cases} g_{j1}(\theta_j), & j < k, \\ (1-p)g_{j1}(\theta_j) + pg_{j2}(\theta_j), & j = k, \\ g_{j2}(\theta_j), & j > k, \end{cases} \tag{5.50}$$

where $g_{j1} \in \mathcal{G}_{j1} = \left\{ g : \int_{\Theta_{j1}} g(\theta_j) d\theta_j = 1 \right\}$ and $g_{j2} \in \mathcal{G}_{j2} = \left\{ g : \int_{\Theta_{j2}} g(\theta_j) d\theta_j = 1 \right\}$ for all j, k is some integer, and $p \in [0, 1)$.

Proof. In order that $D(\widehat{Q} \| \Pi(\cdot | X^{(n)})) < \infty$, we must have

$$\mathrm{supp}(\widehat{Q}) \subseteq \mathrm{supp}(\Pi(\cdot | X^{(n)})) \subseteq \mathrm{supp}(\Pi).$$

In other words, for any set B such that $\Pi(B) = 1$, we must have $\widehat{Q}(B) = 1$. For each coordinate, we can assume that $q_j(\theta_j) = p_j g_{j1}(\theta_j) + (1 - p_j) g_{j2}(\theta_j)$, where $g_{j1} \in \mathcal{G}_{j1}$ and $g_{j2} \in \mathcal{G}_{j2}$. For each k, define

$$B_k = \left\{ \theta = (\theta_j)_{j=1}^{\infty} : \theta_j \in \Theta_{j1} \text{ for } j \leq k \text{ and } \theta_j \in \Theta_{j2} \text{ for } j > k \right\}.$$

Obverse that for $j \neq l$, $B_j \cap B_l = \varnothing$. Then, we can define the set $B = \cup_{k=0}^{\infty} B_k$. According to the sampling process of Π, $\Pi(B) = 1$, which implies that $\widehat{Q}(B) = 1$. Note that for each k,

$$\widehat{Q}(B_k) = \prod_{j \leq k} p_j \prod_{j > k} (1 - p_j),$$

and then

$$1 = \widehat{Q}(B) = \sum_{k=0}^{\infty} \prod_{j \le k} p_j \prod_{j>k} (1 - p_j). \tag{5.51}$$

For any $0 < k < s$,

$$\begin{aligned} 1 &= (1 - p_k + p_k)(1 - p_s + p_s) = (1 - p_k)p_s + [(1 - p_s)p_k + p_k p_s + \\ &\quad (1 - p_k)(1 - p_s)] \prod_{l \neq k,s} (1 - p_l + p_l) \\ &\geq (1 - p_k)p_s + \sum_{k=0}^{\infty} \prod_{j \le k} p_j \prod_{j>k} (1 - p_j) \\ &= (1 - p_k)p_s + 1. \end{aligned}$$

Therefore, $(1 - p_k)p_s = 0$ for all $0 < k < s$, and there are three possible cases:

- $p_j = 0$ for all j.
- $p_j = 1$ for all j.
- $p_j = 0$ for $j < k$, $p_j = 1$ for $j > k$, and $p_k \in [0, 1)$ for some $k \in \mathbb{N}$.

However, the first two cases do not satisfy the constraint (5.51). Thus, the variational posterior $\widehat{Q}$ is limited to the form (5.50), which completes the proof. □

Proof of Theorem 4.2.1. By Lemma 5.3.1, the variational posterior has the form

$$p \prod_{j<k} g_{j1}(\theta_j) \prod_{j \ge k} g_{j2}(\theta_j) + (1 - p) \prod_{j \le k} g_{j1}(\theta_j) \prod_{j>k} g_{j2}(\theta_j).$$

Now we need to determine k, p, g_{j1} for $j \le k$ and g_{j2} for $j \ge k$. We denote the above

distribution by Q_k. Then, it is easy to see that $Q_k(B_{k-1} \cup B_k) = 1$. This implies

$$
\begin{aligned}
& D(Q_k \| \Pi(\cdot|X^{(n)})) \\
= & \int_{B_{k-1}} p \prod_{j<k} g_{j1}(\theta_j) \prod_{j \geq k} g_{j2}(\theta_j) \log \frac{p \prod_{j<k} g_{j1}(\theta_j) \prod_{j \geq k} g_{j2}(\theta_j)}{\pi(k-1) p(X^{(n)}|\theta) \prod_{j<k} f_{j1}(\theta_j) \prod_{j \geq k} f_{j2}(\theta_j)} d\theta \\
& + \int_{B_k} (1-p) \prod_{j \leq k} g_{j1}(\theta_j) \prod_{j>k} g_{j2}(\theta_j) \log \frac{(1-p) \prod_{j \leq k} g_{j1}(\theta_j) \prod_{j>k} g_{j2}(\theta_j)}{\pi(k) p(X^{(n)}|\theta) \prod_{j \leq k} f_{j1}(\theta_j) \prod_{j>k} f_{j2}(\theta_j)} d\theta \\
& + \log p_\Pi(X^{(n)}) \\
= & \; p \log \frac{p}{\pi(k-1) \exp\left(m_{k-1}\left(X^{(n)}, (g_{j1})_{j=1}^{k-1} \cup (g_{j2})_{j=k}^{\infty} \right) \right)} \\
& + (1-p) \log \frac{1-p}{\pi(k) \exp\left(m_k\left(X^{(n)}, (g_{j1})_{j=1}^{k} \cup (g_{j2})_{j=k+1}^{\infty} \right) \right)} \\
& + \log p_\Pi(X^{(n)}),
\end{aligned} \tag{5.52}
$$

where $p_\Pi(X^{(n)}) = \int p(X^{(n)}|\theta) d\Pi(\theta)$. Minimizing $D(Q_k \| \Pi(\cdot|X^{(n)}))$ over p leads to

$$
\widetilde{p} = \frac{\pi(k-1) e^{m_{k-1}\left(X^{(n)}, (g_{j1})_{j=1}^{k-1} \cup (g_{j2})_{j=k}^{\infty} \right)}}{\pi(k-1) e^{m_{k-1}\left(X^{(n)}, (g_{j1})_{j=1}^{k-1} \cup (g_{j2})_{j=k}^{\infty} \right)} + \pi(k) e^{m_k\left(X^{(n)}, (g_{j1})_{j=1}^{k} \cup (g_{j2})_{j=k+1}^{\infty} \right)}}.
$$

Plugging $\widetilde{p}$ into (5.52), we have

$$
\begin{aligned}
& D(Q_k \| \Pi(\cdot|X^{(n)})) \\
= & \; -\log \left[\pi(k-1) e^{m_{k-1}\left(X^{(n)}, (g_{j1})_{j=1}^{k-1} \cup (g_{j2})_{j=k}^{\infty} \right)} + \pi(k) e^{m_k\left(X^{(n)}, (g_{j1})_{j=1}^{k} \cup (g_{j2})_{j=k+1}^{\infty} \right)} \right] \\
& + \log p_\Pi(X^{(n)}).
\end{aligned}
$$

Therefore, $\widetilde{k}$, $\widetilde{g}_{j1}^{(\widetilde{k})}$, $\widetilde{g}_{j2}^{(\widetilde{k})}$ are the solution to maximizxing the objective function

$$
\pi(k-1) e^{m_{k-1}\left(X^{(n)}, (g_{j1})_{j=1}^{k-1} \cup (g_{j2})_{j=k}^{\infty} \right)} + \pi(k) e^{m_k\left(X^{(n)}, (g_{j1})_{j=1}^{k} \cup (g_{j2})_{j=k+1}^{\infty} \right)},
$$

under the constraints that $g_{j1} \in \mathcal{G}_{j1}$ and $g_{j2} \in \mathcal{G}_{j2}$ for all j. The proof is complete. □

Proof of Corollary 4.2.1. If $p(X^{(n)}|\theta) = \prod_{j=1}^{\infty} p(X_j^{(n)}|\theta_j)$, then

$$\begin{aligned}
& m_{k-1}(X^{(n)}, (g_{j1})_{j=1}^{k-1} \cup (g_{j2})_{j=k}^{\infty}) \\
= & \sum_{j=1}^{k-1} \int g_{j1}(\theta_j) \log \frac{f_{j1}(\theta_j) p(X_j^{(n)}|\theta_j)}{g_{j1}(\theta_j)} d\theta_j + \sum_{j=k}^{\infty} \int g_{j2}(\theta_j) \log \frac{f_{j2}(\theta_j) p(X_j^{(n)}|\theta_j)}{g_{j2}(\theta_j)} d\theta_j \\
\leq & \sum_{j=1}^{k-1} \log \int_{\Theta_{j1}} f_{j1}(\theta_j) p(X_j^{(n)}|\theta_j) d\theta_j + \sum_{j=k}^{\infty} \log \int_{\Theta_{j2}} f_{j2}(\theta_j) p(X_j^{(n)}|\theta_j) d\theta_j,
\end{aligned}$$

and

$$\begin{aligned}
& m_k(X^{(n)}, (g_{j1})_{j=1}^{k} \cup (g_{j2})_{j=k+1}^{\infty}) \\
= & \sum_{j=1}^{k} \int g_{j1}(\theta_j) \log \frac{f_{j1}(\theta_j) p(X_j^{(n)}|\theta_j)}{g_{j1}(\theta_j)} d\theta_j + \sum_{j=k+1}^{\infty} \int g_{j2}(\theta_j) \log \frac{f_{j2}(\theta_j) p(X_j^{(n)}|\theta_j)}{g_{j2}(\theta_j)} d\theta_j \\
\leq & \sum_{j=1}^{k} \log \int_{\Theta_{j1}} f_{j1}(\theta_j) p(X_j^{(n)}|\theta_j) d\theta_j + \sum_{j=k+1}^{\infty} \log \int_{\Theta_{j2}} f_{j2}(\theta_j) p(X_j^{(n)}|\theta_j) d\theta_j.
\end{aligned}$$

The equalities above hold when $g_{j1}(\theta_j) \propto f_{j1}(\theta_j) p(X_j^{(n)}|\theta_j) \mathbb{I}_{\theta_j \in \Theta_{j1}}$ for $j \leq k$ and $g_{j2}(\theta_j) \propto f_{j2}(\theta_j) p(X_j^{(n)}|\theta_j) \mathbb{I}_{\theta_j \in \Theta_{j2}}$ for $j \geq k$. Plug these choices into the objective function, and then the objective function becomes

$$\begin{aligned}
& \pi(k-1) \prod_{j=1}^{k-1} \int_{\Theta_{j1}} f_{j1}(\theta_j) p(X_j^{(n)}|\theta_j) d\theta_j \prod_{j=k}^{\infty} \int_{\Theta_{j2}} f_{j2}(\theta_j) p(X_j^{(n)}|\theta_j) d\theta_j \\
& + \pi(k) \prod_{j=1}^{k} \int_{\Theta_{j1}} f_{j1}(\theta_j) p(X_j^{(n)}|\theta_j) d\theta_j \prod_{j=k+1}^{\infty} \int_{\Theta_{j2}} f_{j2}(\theta_j) p(X_j^{(n)}|\theta_j) d\theta_j.
\end{aligned}$$

This implies that $\widetilde{k}$ maximizes $\pi(k-1|X^{(n)}) + \pi(k|X^{(n)})$, where

$$\pi(k|X^{(n)}) \propto \pi(k) \prod_{j=1}^{k} \int_{\Theta_{j1}} f_{j1}(\theta_j) p(X_j^{(n)}|\theta_j) d\theta_j \prod_{j=k+1}^{\infty} \int_{\Theta_{j2}} f_{j2}(\theta_j) p(X_j^{(n)}|\theta_j) d\theta_j.$$

Therefore, $\widetilde{p}$ is given by

$$\widetilde{p} = \frac{\pi(k-1|X^{(n)})}{\pi(k-1|X^{(n)}) + \pi(k|X^{(n)})}.$$

The proof is complete. □

Proof of Theorem 4.2.2. Assume $Q \in \mathcal{S}_{\text{EB}}$, according to the definition, there exists a k such that

$$Q\left(B_k\right) = 1,$$

where $B_k = (\otimes_{j\leq k}\Theta_{j1}) \bigotimes (\otimes_{j>k}\Theta_{j2})$. Then

$$\begin{aligned} & D\left(Q\|\Pi(\cdot|X^{(n)})\right) = \int_{B_k} dQ(\theta) \log \frac{dQ(\theta) p_{\Pi}^{(n)}(X^{(n)})}{d\Pi(\theta) p(X^{(n)}|\theta)} \\ = & \int_{B_k} dQ(\theta) \log \frac{dQ(\theta)/d\theta}{\pi(k) \prod_{j\leq k} f_{j1}(\theta_j) \prod_{j>k} f_{j2}(\theta_j) p(X^{(n)}|\theta)} + \log p_{\Pi}^{(n)}(X^{(n)}), \end{aligned} \tag{5.53}$$

where $p_{\Pi}^{(n)}(X^{(n)}) = \int p^{(n)}(X^{(n)}|\theta) d\Pi(\theta)$.

Therefore, for a specific k, Q is chosen as

$$dQ^{(k)}(\theta) \propto \prod_{j\leq k} f_{j1}(\theta_j) \prod_{j>k} f_{j2}(\theta_j) p(X^{(n)}|\theta) d\theta,$$

to minimize $D\left(Q\|\Pi(\cdot|X^{(n)})\right)$ under the constraint that $Q\left(B_k\right) = 1$. Plug this form into the right hand side of (5.53), we can get $k = \widehat{k}$ selected as

$$\widehat{k} = \underset{k}{\text{argmax}}\, \pi(k) \int \prod_{j\leq k} f_{j1}(\theta_j) \prod_{j>k} f_{j2}(\theta_j) p(X^{(n)}|\theta) d\theta.$$

And therefore, $\widehat{Q}_{\text{EB}} = \widehat{Q}^{(\widehat{k})}$ is the variational posterior with the variational class $\mathcal{S}_{\text{EB}}$. □

5.3.2 Proof of Theorem 4.4.1

Proof of Theorem 4.4.1. First of all, we assume that ϕ is a testing function satisfying condition (4.15). For K defined in 4.17, we set an event as

$$H = \left\{ X : \int \frac{p_\theta(X)}{p_{\theta^*}(X)} d\Gamma_{Z^*}(\theta) \geq \exp\left(-(C_3+1)\epsilon(Z^*)^2 \right) \Gamma_{Z^*}(K) \right\}.$$

Suppose $B = \left\{ \theta \in \Theta : L(\theta, \theta^*) \geq M_1 \epsilon(Z^*)^2 \right\}$ for some constant $M_1 > 0$ to be determined later, then we have

$$P_{\theta^*} \widehat{\Pi}(B) \leq P_{\theta^*} \frac{\int_B p_\theta(X) d\Pi_{\widehat{\lambda}}(\theta)}{\int p_\theta(X) d\Pi_{\widehat{\lambda}}(\theta)} (1-\phi) \mathbb{I}_H + P_{\theta^*}(H^c) + P_{\theta^*}\phi. \tag{5.54}$$

According to the condition (4.15), the third term above is upper bounded by

$$P_{\theta^*}\phi \leq \exp\left(-C\epsilon(Z^*)^2 \right). \tag{5.55}$$

For the second term in (5.54), we have

$$\begin{aligned}
& P_{\theta^*}(H^c) \leq P_{\theta^*} \left(\int \frac{p_\theta(X)}{p_{\theta^*}(X)} d\Gamma_{Z^*}(\theta) \leq e^{-(C_3+1)\epsilon(Z^*)^2} \Gamma_{Z^*}(K) \right) \\
\leq \; & P_{\theta^*} \left(\int_K \frac{p_\theta(X)}{p_{\theta^*}(X)} d\Gamma_{Z^*}(\theta) \leq e^{-(C_3+1)\epsilon(Z^*)^2} \Gamma_{Z^*}(K) \right) \\
= \; & P_{\theta^*} \left(\int_K \frac{p_\theta(X)}{p_{\theta^*}(X)} d\widetilde{\Gamma}_{Z^*}(\theta) \leq e^{-(C_3+1)\epsilon(Z^*)^2} \right) \\
= \; & P_{\theta^*} \left(1 \leq e^{-\rho\epsilon(Z^*)^2} \left(\int_K \frac{p_\theta(X)}{p_{\theta^*}(X)} d\widetilde{\Gamma}_{Z^*}(\theta) \right)^{-\rho} \right) \\
\leq \; & e^{-(C_3+1)\rho\epsilon(Z^*)^2} \mathbb{E}_{\theta^*} \left(\int_K \frac{p_\theta(X)}{p_{\theta^*}(X)} d\widetilde{\Gamma}_{Z^*}(\theta) \right)^{-\rho} \\
\leq \; & e^{-(C_3+1)\rho\epsilon(Z^*)^2} \int_K \left(\int \frac{p_{\theta^*}(X)^{\rho+1}}{p_\theta(X)^\rho} dX \right) d\widetilde{\Gamma}_{Z^*}(\theta) \\
= \; & e^{-(C_3+1)\rho\epsilon(Z^*)^2} \int_K \exp\left[\rho D_{1+\rho}\left(P_{\theta^*} \| P_\theta \right) \right] d\widetilde{\Gamma}_{Z^*}(\theta),
\end{aligned}$$

where $\widetilde{\Gamma}_{Z^*}$ is a truncated probability measurement of Γ_{Z^*} on K, i.e. $\widetilde{\Gamma}_{Z^*}(A) = \frac{\Gamma_{Z^*}(A \cap K)}{\Gamma_{Z^*}(K)}$. According to the definition (4.17), we get

$$P_{\theta^*}(H^c) \leq \exp\left(-\rho\epsilon(Z^*)^2\right). \tag{5.56}$$

Finally, we consider the first term in (5.54). According to the definition of $\widehat{\lambda}$ in (4.13),

$$P_{\theta^*} \frac{\int_B p_\theta(X) d\Pi_{\widehat{\lambda}}(\theta)}{\int p_\theta(X) d\Pi_{\widehat{\lambda}}(\theta)} (1-\phi)\mathbb{I}_H \leq P_{\theta^*} \frac{\int_B \max_{\lambda\in\Lambda}\left\{w(\lambda)\frac{p_\theta(X)}{p_{\theta^*}(X)} d\Pi_\lambda(\theta)\right\}}{w(\lambda^*)\int \frac{p_\theta(X)}{p_{\theta^*}(X)} d\Pi_{\lambda^*}(\theta)} (1-\phi)\mathbb{I}_H.$$

For the numerator, we have

$$\begin{aligned}
&\int_B \max\left\{w(\lambda)\frac{p_\theta(X)}{p_{\theta^*}(X)} d\Pi_\lambda(\theta)\right\} = \int_B \max\left\{w(\lambda)\sum_{Z\in\mathcal{Z}} \nu_\lambda(Z)\frac{p_\theta(X)}{p_{\theta^*}(X)} d\Gamma_Z(\theta)\right\} \\
\leq\ & \int_B \sum_{Z\in\mathcal{Z}} \max_\lambda \{w(\lambda)\nu_\lambda(Z)\} \frac{p_\theta(X)}{p_{\theta^*}(X)} d\Gamma_Z(\theta) \\
=\ & \sum_{Z\in\mathcal{Z}} \gamma(Z) \int_B \frac{p_\theta(X)}{p_{\theta^*}(X)} d\Gamma_Z(\theta),
\end{aligned}$$

where $\gamma(Z) = \max_{\lambda\in\Lambda}\{w(\lambda)\nu_\lambda(Z)\}$.

Under the event H, we can lower bound the denominator by

$$\begin{aligned}
& w(\lambda^*)\int \frac{p_\theta(X)}{p_{\theta^*}(X)} d\Pi_{\lambda^*}(\theta) \geq w(\lambda^*)\nu_{\lambda^*}(Z^*)\int \frac{p_\theta(X)}{p_{\theta^*}(X)} d\Gamma_{Z^*}(\theta) \\
\geq\ & w(\lambda^*)\nu_{\lambda^*}(Z^*)\exp\left(-C_3\epsilon(Z^*)^2\right)\Gamma_{Z^*}(K).
\end{aligned}$$

Then we have

$$\begin{aligned}
& P_{\theta^*} \frac{\int_B p_\theta(X) d\Pi_{\widehat{\lambda}}(\theta)}{\int p_\theta(X) d\Pi_{\widehat{\lambda}}(\theta)} (1-\phi)\mathbb{I}_H \\
\leq\ & \exp\left(C_3\epsilon(Z^*)^2\right) \sum_{Z\in\mathcal{Z}} \frac{\gamma(Z)}{w(\lambda^*)\nu_{\lambda^*}(Z^*)} \frac{1}{\Gamma_{Z^*}(K)} \int_B P_\theta(1-\phi) d\Gamma_Z(\theta)
\end{aligned}$$

Note that B can be decomposed as $B = \cup_{Z\in\mathcal{Z}}\cup_{l=1}^{\infty}\Big(R_Z(\sqrt{(l+1)M_1}\epsilon(Z^*))\backslash R_Z(\sqrt{lM}\epsilon(Z^*))\Big)$, then according to the testing condition (4.15) and the prior ratio condition (4.18), we have

$$\begin{aligned}
&\frac{1}{\Gamma_{Z^*}(K)}\int_B P_\theta(1-\phi)d\Gamma_Z(\theta)\\
=\;& \frac{1}{\Gamma_{Z^*}(K)}\sum_{l=1}^{\infty}\exp\left(C_2(\epsilon(Z)^2+\epsilon(Z^*)^2)-C_1l\epsilon^2\right)\Gamma_Z\left(R_Z(\sqrt{(l+1)M_1}\epsilon(Z^*))\right)\\
&-\frac{1}{\Gamma_{Z^*}(K)}\sum_{l=1}^{\infty}\exp\left(C_2(\epsilon(Z)^2+\epsilon(Z^*)^2)-C_1l\epsilon^2\right)\Gamma_Z\left(R_Z(\sqrt{lM_1}\epsilon(Z^*))\right)\\
\le\;& \exp\left(C_2(\epsilon(Z)^2+\epsilon(Z^*)^2)\right)\sum_{l=1}^{\infty}\exp\left(-C_1lM\epsilon(Z^*)^2\right)\frac{\Gamma_Z\left(R_Z(\sqrt{(l+1)M}\epsilon(Z^*))\right)}{\Gamma_{Z^*}(K)}\\
\le\;& \frac{\delta(Z)}{\delta(Z^*)}\exp\left(C_2(\epsilon(Z)^2+\epsilon(Z^*)^2)\right)\\
&\times\sum_{l=1}^{\infty}\exp\left(-C_1lM\epsilon(Z^*)^2+c(l+1)M_1\epsilon(Z^*)^2+C_4\epsilon(Z)^2+C_5\epsilon(Z^*)^2\right)\\
\le\;& \frac{\delta(Z)}{\delta(Z^*)}\exp\left(C_2(\epsilon(Z)^2+\epsilon(Z^*)^2)\right)\\
&\times\sum_{l=1}^{\infty}\exp\left(-C_1lM\epsilon(Z^*)^2+2c\epsilon(Z^*)^2+C_4\epsilon(Z)^2+C_5\epsilon(Z^*)^2\right)\\
\le\;& \frac{\delta(Z)}{\delta(Z^*)}\exp\left(C_2(\epsilon(Z)^2+\epsilon(Z^*)^2+C_5\epsilon(Z^*)^2)\right)\sum_{l=1}^{\infty}\exp\left(-clM\epsilon(Z^*)^2+C_4\epsilon(Z)^2\right).
\end{aligned}$$

Without loss of generality, we can assume $\epsilon(Z^*)>1$ and set $M_1>1$, then $\frac{1}{1-\exp(-cM_1\epsilon^2)}<\exp\left(m_1\epsilon(Z^*)^2\right)$, where $m_1=-\log(1-e^{-c})<\infty$. Then we have

$$\frac{1}{\Gamma_{Z^*}(K)}\int_B P_\theta(1-\phi)d\Gamma_Z(\theta)\le\frac{\delta(Z)}{\delta(Z^*)}\exp\left((C_2+C_4)\epsilon(Z)^2+(m_1+C_2+C_5-cM_1)\epsilon(Z^*)^2\right).$$

Finally, with summability condition (4.19), we will have

$$\begin{aligned}
&P_{\theta^*}\frac{\int_B p_\theta(X)d\Pi_{\widehat{\lambda}}(\theta)}{\int p_\theta(X)d\Pi_{\widehat{\lambda}}(\theta)}(1-\phi)\mathbb{I}_H\\
\le\;& \exp\left((C_2+C_3+C_5+m_1-cM_1)\epsilon(Z^*)^2\right) \qquad (5.57)
\end{aligned}$$

$$\times \sum_{Z\in\mathcal{Z}} \frac{\gamma(Z)}{w(\lambda^*)\nu_{\lambda^*}(Z^*)}\frac{\delta(Z)}{\delta(Z^*)}\exp\left((C_2+C_4)\epsilon(Z)^2\right)$$
$$\leq \exp\left((C_2+C_3+C_5+C_6+m_1-cM_1)\epsilon(Z^*)^2\right). \tag{5.58}$$

Then we choose $M_1 = \max\{c^{-1}(C_2+C_3+C_5+C_6+m_1+1), 1, M_0\}$ and $M_2 = \min\{1, C, \rho\}$. Combine (5.57), (5.56), (5.55) together with (5.54), we will have

$$P_{\theta^*}\widehat{\Pi}\left(L(\theta,\theta^*) > M_1\epsilon(Z^*)^2\right) \leq 3\exp\left(-M_2\epsilon(Z^*)^2\right).$$

□

5.3.3 Proof of Theorem 4.5.1

Now we show Theorem 4.5.1 by applying Theorem 4.4.1. We choose $\rho = 1$, $C_3 = 1$, $L(\theta,\theta^*) = \|\theta-\theta^*\|^2$, $\epsilon(S)^2 = |S|\log\frac{ep}{|S|}$, $\delta(S) = 1$ to check the conditions. To show this Theorem, we start from the following lemmas.

Lemma 5.3.2. *There exists constants $M_0, C, C_2 > 0$ and a testing function ϕ for sparse sequence model, such that*

$$P_{\theta^*}\phi \leq \exp\left(-Cs^*\log\frac{ep}{s^*}\right),$$

and

$$\sup_{\theta\in\Theta_S:\|\theta-\theta^*\|^2\geq\epsilon^2} P_\theta(1-\phi) \leq \exp\left(C_2\left(s\log\frac{ep}{s}+s^*\log\frac{ep}{s^*}\right)-\frac{1}{8}\epsilon^2\right),$$

for any $\epsilon^2 > M_0 s^*\log\frac{ep}{s^*}$ and $S\in[p]$, where $s^* = |S^*|$ and $s = |S|$

Lemma 5.3.3. *There exists a constant $M_0 > 0$ such that*

$$\frac{\Gamma_S\left(\left\{\theta\in\Theta_S : \|\theta-\theta^*\|^2\leq\epsilon^2\right\}\right)}{\Gamma_{S^*}\left(\left\{\theta\in\Theta_{S^*}\|\theta-\theta^*\|^2\leq s^*\log\frac{ep}{s^*}\right\}\right)} \leq \exp\left(\frac{1}{32}\epsilon^2 + C_4\epsilon(S)^2 + C_5\epsilon(S^*)^2\right)$$

for all $\epsilon^2 > M_0 s^*\log\frac{ep}{s^*}$.

Lemma 5.3.4. *For any $C > 0$, there exists a constant $D > 0$. When (4.23) is satisfied for $\nu_1 > \nu_2 > D$, there exists $\lambda^* \in [0,1]$ and a constant $C' > 0$ such that*

$$\sum_{S \in [p]} \frac{\gamma(S)}{w(\lambda^*)\nu_{\lambda^*}(S^*)} \exp\left(C\epsilon(S)^2\right) \lesssim \exp\left(C'\epsilon(S^*)^2\right).$$

Proof of Theorem 4.5.1. This theorem can be directly proved by applying Theorem 4.4.1 with Lemma 5.3.2, 5.3.3 and 5.3.4. □

Proof of Lemma 5.3.2. Consider the testing function

$$\phi_S = \mathbb{I}\left\{\|X - \theta^*\|_S^2 > a\left(\epsilon(S) + \epsilon(S^*)\right)\right\},$$

where $\epsilon(S) = |S| \log \frac{ep}{|S|}$ for all $S \subseteq [p]$ and

$$\phi = \max_{S \subset [p]} \phi_S,$$

with $a > 0$ to be determined later. Then we show that ϕ is a test function satisfying the condition (4.15). Set $|S| = s$ and $|S^*| = s^*$, then we should note that

$$\begin{aligned} P_{\theta^*}\phi &\leq \sum_{s=1}^{p} \binom{p}{s} P\left(\xi_s^2 > a\left(s \log \frac{ep}{s} + s^* \log \frac{ep}{s^*}\right)\right) \\ &\leq \sum_{s=1}^{p} \exp\left(2s \log \frac{ep}{s} - \frac{s}{2}\log(1-2t) - at\left(s \log \frac{ep}{s} + s^* \log \frac{ep}{s^*}\right)\right) \end{aligned}$$

for any $t > 0$. We choose $t = 1/4$, then

$$P_{\theta^*}\phi \leq \sum_{s=1}^{p} \exp\left(-\left(\frac{a}{4} - 2 - \frac{1}{2}\log 2\right) s \log \frac{ep}{s} - \frac{a}{4} s^* \log \frac{ep}{s^*}\right).$$

Choose $a = 10$, we have $\frac{a}{4} - 2 - \frac{1}{2}\log 2 > 0$ and then

$$P_{\theta^*}\phi \leq \sum_{s=1}^{p} \exp\left(-\frac{5}{2}s^* \log\frac{ep}{s^*}\right) \leq \exp\left(-\frac{3}{2}s^* \log\frac{ep}{s^*}\right).$$

For any $\theta \in \Theta_S$ such that $\|\theta - \theta^*\|^2 \geq \epsilon^2$, we have

$$2\|X - \theta^*\|^2 + 2\|X - \theta\|^2 \geq \|\theta - \theta^*\|^2 > \epsilon^2.$$

Then $\|X - \theta\|^2 \geq \frac{1}{2}\epsilon^2 - \|X - \theta^*\|^2$ and in the same way, we have

$$\begin{aligned} & P_\theta(1-\phi) \leq P_\theta(1 - \phi_{S\cup S^*}) = P_\theta\left(\|X-\theta^*\|^2 \leq 10(\epsilon(S\cup S^*) + \epsilon(S^*))\right) \\ \leq\ & P_\theta\left(\|X-\theta\|^2 \geq \frac{1}{2}\epsilon^2 - 10\left(\epsilon(S\cup S^*) + \epsilon(S^*)\right)\right) \\ \leq\ & \exp\left(-\frac{|S\cup S^*|}{2}\log(1-2t) - \frac{1}{2}t\epsilon^2 + 10t(\epsilon(S\cup S^*) + \epsilon(S^*))\right). \end{aligned}$$

We also choose $t = 1/4$ in this case, then

$$P_\theta(1-\phi) \leq \exp\left(C_2(s\log\frac{ep}{s} + s^*\log\frac{ep}{s^*}) - \frac{1}{8}\epsilon^2\right),$$

where $C_2 = 5 + \frac{1}{2}\log 2$.

Finally, the proof is complete by setting $M_0 = 1$, $C = \frac{3}{2}$, $C_2 = 5 + \frac{1}{2}\log 2$. □

Proof of Lemma 5.3.3. According to the definition, we have

$$\frac{\Gamma_S\left(\left\{\theta \in \Theta_S : \|\theta - \theta^*\|^2 \leq \epsilon^2\right\}\right)}{\Gamma_{S^*}\left(\left\{\theta \in \Theta_{S^*} \|\theta - \theta^*\|^2 \leq \epsilon(S^*)^2\right\}\right)} \leq (\rho/2)^{s-s^*} \frac{\int_{\|\theta-\theta^*\|^2<\epsilon^2} \exp\left(-\rho\|\theta_S\|_1\right) d\theta_S}{\int_{\|\theta-\theta^*\|^2<\epsilon(S^*)^2} \exp\left(-\rho\|\theta_{S^*}\|_1\right) d\theta_{S^*}}.$$

Then for any $\theta \in \Theta_S$ such that $\|\theta - \theta^*\|^2 \leq \epsilon^2$ and any $\bar{\theta} \in \Theta_{S^*}$ such that $\|\bar{\theta} - \theta^*\|^2 \leq \epsilon(S^*)^2$, we have

$$\|\theta - \bar{\theta}\|^2 \leq 2\left(\|\theta - \theta^*\|^2 + \|\bar{\theta} - \theta^*\|^2\right) \leq 2(\epsilon^2 + \epsilon(S^*)^2).$$

Then we have

$$
\begin{aligned}
& -\rho\|\theta\|_1 + \rho\|\bar{\theta}\|_1 \leq \rho\|\theta - \bar{\theta}\|_1 \leq \rho\sqrt{s+s^*}\|\theta - \bar{\theta}\| \leq \xi(s+s^*) + \frac{1}{4}\xi^{-1}\|\theta - \bar{\theta}\|^2 \\
\leq\ & \frac{1}{2}\xi^{-1}\epsilon^2 + \xi\epsilon(S)^2 + \left(\frac{1}{2\xi} + \xi\right)\epsilon(S^*)^2,
\end{aligned}
$$

where $\xi > 0$ to be determined later. Therefore,

$$
\begin{aligned}
& \frac{\Gamma_S\left(\left\{\theta \in \Theta_S : \|\theta - \theta^*\|^2 \leq \epsilon^2\right\}\right)}{\Gamma_{S^*}\left(\left\{\theta \in \Theta_{S^*}\|\theta - \theta^*\|^2 \leq \epsilon(S^*)^2\right\}\right)} \\
\leq\ & \exp\left(\frac{1}{2\xi}\epsilon^2 + \left(\xi + |\log\frac{\rho}{2}|\right)\epsilon(S)^2 + \left(\frac{1}{2\xi} + \xi + |\log\frac{\rho}{2}|\right)\epsilon(S^*)^2\right)\frac{\mathrm{Vol}(B_s(\epsilon))}{\mathrm{Vol}(B_{s^*}(\epsilon(S^*)))},
\end{aligned}
$$

where $\mathrm{Vol}(B_s(r))$ denotes the volume of s dimensional ball with radius r. Then $\mathrm{Vol}(B_s(r)) = \frac{\pi^{\frac{s}{2}}}{\Gamma(s/2+1)}r^s$ and

$$
\begin{aligned}
& \frac{\Gamma_S\left(\left\{\theta \in \Theta_S : \|\theta - \theta^*\|^2 \leq \epsilon^2\right\}\right)}{\Gamma_{S^*}\left(\left\{\theta \in \Theta_{S^*}\|\theta - \theta^*\|^2 \leq \epsilon(S^*)^2\right\}\right)} \\
\leq\ & \exp\left(\frac{1}{2\xi}\epsilon^2 + \left(\xi + |\log\frac{\rho}{2}| + \frac{1}{2}\log(2\pi e)\right)\epsilon(S)^2 + \left(\frac{1}{2\xi} + \xi + |\log\frac{\rho}{2}|\right)\epsilon(S^*)^2\right) \\
& \times \left(\frac{\epsilon}{\sqrt{s}}\right)^s \left(\frac{\sqrt{s^*}}{\epsilon(S^*)}\right)^{s^*} \\
\leq\ & \exp\left(\frac{1}{2\xi}\epsilon^2 + \left(\xi + |\log\frac{\rho}{2}| + \frac{1}{2}\log(2\pi e)\right)\epsilon(S)^2\right. \\
& \left. + \left(\frac{1}{2\xi} + \xi + |\log\frac{\rho}{2}|\right)\epsilon(S^*)^2 + \frac{s}{2}\log\left(\frac{\epsilon^2}{s}\right)\right).
\end{aligned}
$$

With the fact that $\frac{s}{2}\log(\epsilon^2/s) \leq \frac{s}{2}\log\xi + \frac{s}{2}\log\left(1 + \frac{\epsilon^2}{\xi s}\right) \leq \frac{s}{2}\log\xi + \frac{\epsilon^2}{2\xi}$, we finally have

$$
\begin{aligned}
& \frac{\Gamma_S\left(\left\{\theta \in \Theta_S : \|\theta - \theta^*\|^2 \leq \epsilon^2\right\}\right)}{\Gamma_{S^*}\left(\left\{\theta \in \Theta_{S^*}\|\theta - \theta^*\|^2 \leq \epsilon(S^*)^2\right\}\right)} \\
\leq\ & \exp\left(\frac{1}{\xi}\epsilon^2 + \left(\xi + |\log\frac{\rho}{2}| + \frac{1}{2}\log(2\pi e\xi)\right)\epsilon(S)^2 + \left(\frac{1}{2\xi} + \xi + |\log\frac{\rho}{2}|\right)\epsilon(S^*)^2\right).
\end{aligned}
$$

Choose $\xi = 32$ and set $C_4 = 32 + |\log \frac{\rho}{2}| + \frac{1}{2}\log(64\pi e)$, $C_5 = 33 + |\log \frac{\rho}{2}|$. □

Proof of Lemma 5.3.4. For sparse sequence model, we have

$$w(\lambda) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\lambda^{\alpha-1}(1-\lambda)^{\beta-1}$$

and for each $S \subset [p]$,

$$\nu_\lambda(S) = \lambda^{|S|}(1-\lambda)^{p-|S|}.$$

Then,

$$\begin{aligned}\gamma(S) &= \max_{\lambda\in[0,1]}\{w(\lambda)\nu_\lambda(S)\} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\max_{\lambda\in[0,1]}\left\{\lambda^{\alpha+|S|-1}(1-\lambda)^{p-|S|+\beta-1}\right\} \\ &\quad \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{\alpha+|S|-1}{p+\alpha+\beta-2}\right)^{\alpha+|S|-1}\left(\frac{p-|S|+\beta-1}{p+\alpha+\beta-2}\right)^{p-|S|+\beta-1},\end{aligned}$$

Note that $\gamma(S)$ only depends on $|S|$, then we denote $\gamma_s = \gamma(S)$ for all S such that $|S| = s$ with

$$\gamma_s = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{\alpha+|S|-1}{p+\alpha+\beta-2}\right)^{\alpha+s-1}\left(\frac{p-s+\beta-1}{p+\alpha+\beta-2}\right)^{p-s+\beta-1}.$$

We choose $\lambda^* = \frac{\alpha+s^*-1}{p+\alpha+\beta-2}$, then $w(\lambda^*)\nu_{\lambda^*}(S^*) = \gamma_{s^*}$. Then,

$$\sum_{S\in[p]}\frac{\gamma(S)}{w(\lambda^*)\nu_{\lambda^*}(S^*)}\exp\left(C\epsilon(S)^2\right) \lesssim \sum_{s=1}^{p}\frac{\gamma_s}{\gamma_{s^*}}\exp\left((C+1)s\log\frac{ep}{s}\right).$$

When (4.23) is satisfied, we have $p^{-\nu_1} \le \gamma_{k+1}/\gamma_k \le p^{-\nu_2}$ and then $\frac{\gamma_s}{\gamma_{s^*}} = \frac{\gamma_s}{\gamma_1}\frac{\gamma_1}{\gamma_{s^*}} \le p^{\nu_1 s^* - \nu_2 s}$. When $s^* < p^a$ for some $a < 1$, $\epsilon(S^*) \asymp s^*\log p$, then for $\nu_1 > \nu_2 > C+2$ we will have

$$\begin{aligned}&\sum_{S\in[p]}\frac{\gamma(S)}{w(\lambda^*)\nu_{\lambda^*}(S^*)}\exp\left(C\epsilon(S)^2\right) \lesssim \sum_{s=1}^{p}\exp\left(\nu_1 s^*\log p - (\nu_2 - C - 1)s\log\frac{ep}{s}\right) \\ \le\ &\exp\left(\nu_1 s^*\log p - (\nu_2 - C - 2)\log p\right) \le \exp\left(\nu_1 s^*\log p\right).\end{aligned}$$

The proof is complete.

□

5.3.4 Proof of Theorem 4.5.2

To prove this theorem, we choose $\epsilon(S)$ as follows:

$$\epsilon(S)^2 = \begin{cases} |S| \log p & S \neq S^* \\ \frac{s^* \log p}{\kappa^2 \wedge 1} & S = S^* \end{cases} \tag{5.59}$$

For loss function, we choose the ℓ_2 loss between predictions: $L(\theta, \theta^*) = \|X(\theta - \theta^*)\|^2$. The calibrate parameter is $\delta(S) = 1$ as well. We also remain $\rho = 1$ and $C_3 = 1$ in the Theorem 4.4.1.

As the $\epsilon(S)$ in sparse linear regression is identical to that in the sparse sequence model except $S = S^*$, Lemma 5.3.4 can be generalised to this case and the summability condition holds. Likewise, we also propose some Lemmas to check the testing conditions and prior ratio conditions for sparse linear regression model.

Lemma 5.3.5. *There exists constants $M_0, C, C_2 > 0$ and a testing function ϕ for sparse linear regression model, such that*

$$P_\theta \phi \leq \exp\left(-C\epsilon(S^*)^2\right),$$

and

$$\sup_{\theta \in \Theta_S : \|X(\theta - \theta^*)\|^2} P_\theta(1 - \phi) \leq \exp\left(C_2(\epsilon(S)^2 + \epsilon(S^*)^2) - \frac{1}{8}\epsilon^2\right),$$

for any $\epsilon^2 > M_0 \epsilon(S^*)^2$ and $S \in [p]$.

Lemma 5.3.6. *There exists a constant $M_0 > 0$ such that*

$$\frac{\Gamma_S\left(\left\{\theta \in \Theta_S : \|X(\theta - \theta^*)\|^2 \leq \epsilon^2\right\}\right)}{\Gamma_{S^*}\left(\left\{\theta \in \Theta_{S^*} \|X(\theta - \theta^*)\|^2 \leq \epsilon(S^*)^2\right\}\right)} \leq \exp\left(\frac{1}{32}\epsilon^2 + C_4\epsilon(S)^2 + C_5\epsilon(S^*)^2\right)$$

for all $\epsilon^2 > M_0 s^* \log p$.

Proof of Theorem 4.5.2. The theorem can be directly proved by applying Theorem 4.4.1 with Lemma 5.3.5, 5.3.6 and a slightly modified Lemma 5.3.4. □

Now we start to show Lemma 5.3.5 and 5.3.6

Proof of Lemma 5.3.5. For each support set S, we consider

$$\phi_S = \mathbb{I}\left\{\|P_S(Y - X_{S^*}\theta^*_{S^*})\|^2 > 10\epsilon(S^*)^2 + 10\epsilon(S)^2\right\},$$

where P_S is the projection matrix on X_S defined by $P_S = X_S^T(X_S^T X_S)^- X_S$. Then overall, set

$$\phi = \max_{S\subset[p]} \phi_S.$$

Then, with almost the same procedure as the proof in Lemma 5.3.2 and the fact that $\epsilon(S^*) \geq s^* \log \frac{ep}{s^*} \geq \log p$, we have

$$\begin{aligned} P_{\theta^*}\phi &\leq \sum_{S\in[p]} P\left(\|P_S W\|^2 > 10(\epsilon(S)^2 + \epsilon(S^*)^2)\right) \\ &= \sum_{s=1}^{p} \binom{p}{s} P\left(\xi_s^2 > 10(\epsilon(S)^2 + \epsilon(S^*)^2)\right) \leq \exp\left(-\frac{3}{2}\epsilon(S^*)^2\right). \end{aligned}$$

In the same way, for any $\theta \in \Theta_S$ such that $\|X(\theta - \theta^*)\| \geq \epsilon$, we have

$$2\|P_{S\cup S^*}(Y - X_{S^*}\theta^*_{S^*})\|^2 + 2\|P_{S\cup S^*}(Y - X_S\theta_S)\|^2 \geq \|X(\theta - \theta^*)\|^2 \geq \epsilon^2,$$

then

$$\begin{aligned} & P_\theta(1-\phi) \leq P_\theta(1 - \phi_{S\cup S^*}) = P_\theta\left(\|P_{S\cup S^*}(Y - X\theta^*)\|^2 \leq 10(\epsilon(S\cup S^*)^2 + \epsilon(S^*)^2)\right) \\ \leq\ & P_\theta\left(\|P_{S\cup S^*}(Y - X\theta)\|^2 \geq \frac{1}{2}\epsilon^2 - 10(\epsilon(S\cup S^*)^2 + \epsilon(S^*)^2)\right) \end{aligned}$$

$$\leq \quad \exp\left(C_2(\epsilon(S)^2+\epsilon(S^*)^2)-\frac{1}{8}\epsilon^2\right).$$

The proof is complete. □

Proof of Lemma 5.3.6. First we upper bound the numerator, note that

$$\begin{aligned}
&\Gamma_S\left(\left\{\theta\in\Theta_S:\|X(\theta-\theta^*)\|^2\leq\epsilon^2\right\}\right)\\
=\;&\Gamma_S\left(\left\{\theta\in\Theta_S:\|X(\theta-\theta^*)\|^2\leq\epsilon^2,\|\theta_{S^{*c}}\|_1\leq 3\|(\theta-\theta^*)_{S^*}\|_1\right\}\right)\\
&+\Gamma_S\left(\left\{\theta\in\Theta_S:\|X(\theta-\theta^*)\|^2\leq\epsilon^2,\|\theta_{S^{*c}}\|_1> 3\|(\theta-\theta^*)_{S^*}\|_1\right\}\right)\\
\leq\;&\Gamma_S\left(\left\{\theta\in\Theta_S:\|\theta-\theta^*\|_1\leq\frac{\sqrt{s^*}\epsilon}{\kappa}\right\}\right)+\Gamma_S\left(\{\theta\in\Theta_S:\|\theta_{S^{*c}}\|_1>3\|(\theta-\theta^*)_{S^*}\|_1\}\right)
\end{aligned}$$

Suppose $|S|=s$, then for the first term, we have

$$\begin{aligned}
&e^{\rho\|\theta^*\|_1}\Gamma_S\left(\left\{\theta\in\Theta_S:\|\theta-\theta^*\|_1\leq\frac{\sqrt{s^*}\epsilon}{\kappa}\right\}\right)\\
\leq\;&\left(\frac{\rho}{2}\right)^s\int_{\|\theta-\theta^*\|_1\leq\frac{\sqrt{s^*}\epsilon}{\kappa}}\exp\left(\rho(\|\theta^*\|_1-\|\theta\|_1)\right)d\theta_S\\
\leq\;&\exp\left(s\log\frac{\rho}{2}+\frac{\rho\sqrt{s^*}\epsilon}{\kappa}\right)\mathrm{Vol}\left(\left\{\theta\in\Theta_S:\|\theta\|_1\leq\frac{\sqrt{s^*}\epsilon}{\kappa}\right\}\right)\\
\leq\;&\exp\left(s\log\frac{\rho}{2}+\frac{\rho\sqrt{s^*}\epsilon}{\kappa}+s\log\frac{e\sqrt{s^*}\epsilon}{2s\kappa}\right)\\
\leq\;&\exp\left(s\log\frac{\rho}{2}+\frac{\rho^2s^*}{2\xi_1\kappa^2}+\frac{\xi_1\epsilon^2}{2}+\frac{s}{2}\log\frac{\epsilon^2}{\xi_2 s}+\frac{s}{2}\log\frac{e^2s^*\xi_2}{4s\kappa^2}\right)\\
\leq\;&\exp\left(s\log\frac{\rho}{2}+\frac{\rho^2s^*}{2\xi_1\kappa^2}+\frac{\xi_1\epsilon^2}{2}+\frac{\epsilon^2}{2\xi_2}+\frac{e^2s^*\xi_2}{8\kappa^2}\right),
\end{aligned}$$

Set $\xi_1=1/32$ and $\xi_2=32$, then

$$\begin{aligned}
&e^{\rho\|\theta^*\|_1}\Gamma_S\left(\left\{\theta\in\Theta_S:\|\theta-\theta^*\|_1\leq\frac{\sqrt{s^*}\epsilon}{\kappa}\right\}\right)\\
\leq\;&\exp\left(|\log\frac{\rho}{2}|\epsilon(S)^2+(16\rho^2+4e^2)\epsilon(S^*)^2+\frac{1}{32}\epsilon^2\right)
\end{aligned}$$

For the second term,

$$
\begin{aligned}
& e^{\rho\|\theta^*\|_1}\Gamma_S\left(\{\theta\in\Theta_S:\|\theta_{S^{*c}}\|_1>3\|(\theta-\theta^*)_{S^*}\|_1\}\right)\\
\leq\ & \left(\frac{\rho}{2}\right)^s\int_{\|\theta_{S^{*c}}\|_1>3\|(\theta-\theta^*)_{S^*}\|_1}\exp\left(-\rho\|\theta\|_1+\rho\|\theta^*\|_1\right)d\theta_S\\
\leq\ & \left(\frac{\rho}{2}\right)^s\int_{\|\theta_{S^{*c}}\|_1>3\|(\theta-\theta^*)_{S^*}\|_1}\exp\left(-\rho\|\theta_{S^{*c}}\|_1+\rho\|(\theta-\theta^*)_{S^*}\|_1\right)d\theta_S\\
\leq\ & \left(\frac{\rho}{2}\right)^s\int_{\|\theta_{S^{*c}}\|_1>3\|(\theta-\theta^*)_{S^*}\|_1}\exp\left(-\frac{\rho}{2}\left(\|\theta_{S^{*c}}\|_1+\|(\theta-\theta^*)_{S^*}\|_1\right)\right)d\theta_S\\
=\ & \left(\frac{\rho}{2}\right)^s\int_{\|\theta_{S^{*c}}\|_1>3\|(\theta-\theta^*)_{S^*}\|_1}\exp\left(-\frac{\rho}{2}\|\theta-\theta^*\|_1\right)d\theta_S\\
\leq\ & \left(\frac{\rho}{2}\right)^s\int\exp\left(-\frac{\rho}{2}\|(\theta-\theta^*)_S\|_1\right)d\theta_S\\
=\ & 2^s\leq\exp\left(\epsilon(S)^2\log 2\right)
\end{aligned}
$$

Therefore, overall we have

$$
\begin{aligned}
& e^{\rho\|\theta^*\|_1}\Gamma_S\left(\left\{\theta\in\Theta_S:\|X(\theta-\theta^*)\|^2\leq\epsilon^2\right\}\right) \qquad (5.60)\\
\leq\ & \exp\left((|\log\rho|+3\log 2)\epsilon(S)^2+(16\rho^2+4e^2)\epsilon(S^*)^2+\frac{1}{32}\epsilon^2\right)
\end{aligned}
$$

Now we build the lower bound for the denominator, note that for $\theta\in\Theta_{S^*}$, we have

$$\|X(\theta-\theta^*)\|^2\leq\|X_{S^*}\|_F^2\|\theta-\theta^*\|_1^2\leq p^{a+C}\|\theta-\theta^*\|_1^2.$$

Then,

$$
\begin{aligned}
& e^{\rho\|\theta^*\|_1}\Gamma_{S^*}\left(\left\{\theta\in\Theta_{S^*}:\|X(\theta-\theta^*)\|^2\leq\epsilon(S^*)^2\right\}\right)\\
=\ & \left(\frac{\rho}{2}\right)^{s^*}\int_{\|X(\theta-\theta^*)\|^2\leq\epsilon(S^*)^2}\exp(-\rho\|\theta\|_1+\rho\|\theta^*\|_1)d\theta_{S^*}\\
\geq\ & \left(\frac{\rho}{2}\right)^{s^*}\int_{\|\theta-\theta^*\|_1\leq p^{-a-C}\epsilon(S^*)^2}\exp(-\rho\|\theta-\theta^*\|_1)d\theta_{S^*}\\
\geq\ & \left(\frac{\rho}{2}\right)^{s^*}\int_{\|\theta-\theta^*\|_1\leq p^{-a-C}}\exp(-\rho\|\theta-\theta^*\|_1)d\theta_{S^*}
\end{aligned}
$$

$$
\begin{aligned}
&\geq \exp\left(-s^*|\log\frac{\rho}{2}| - \rho p^{-a-C}\right)\mathrm{Vol}\left(\left\{\theta\in\Theta_{S^*}\|\theta-\theta^*\|_1\leq p^{-a-C}\right\}\right)\\
&\geq \exp\left(-\left(\frac{\rho^2+1}{2}+|\log\frac{\rho}{2}|\right)\epsilon(S^*)^2 - 2(a+C+1)\log p\right)\\
&\geq \exp\left(-\left(\frac{\rho^2+1}{2}+|\log\frac{\rho}{2}|+2(a+C+1)\right)\epsilon(S^*)^2\right) \qquad (5.61)
\end{aligned}
$$

for some constant $C'>0$. Combining (5.60) and (5.61), we have

$$
\frac{\Gamma_S\left(\left\{\theta\in\Theta_S:\|X(\theta-\theta^*)\|^2\leq\epsilon^2\right\}\right)}{\Gamma_{S^*}\left(\left\{\theta\in\Theta_{S^*}\|X(\theta-\theta^*)\|^2\leq\epsilon(S^*)\right\}\right)}\leq\exp\left(\frac{1}{32}\epsilon^2+C_4\epsilon(S)^2+C_5\epsilon(S^*)^2\right),
$$

where $C_4=|\log\rho|+3\log 2$, $C_5=17\rho^2+4e^2+1+|\log\frac{\rho}{2}|+2(a+C+1)$. □

5.3.5 Proof of Theorem 4.5.3

We also check the three conditions (4.15), (4.18) and (4.19) to prove Theorem 4.5.3.

Proof of Theorem 4.5.3. In this model, we simply choose $\epsilon(Z)^2=\epsilon_\tau$ for $Z\in\bar{\mathcal{Z}}_\tau$. The loss function is the ℓ_2 loss function $L(\mathcal{X}_ZB,\mathcal{X}_{Z^*}B^*)=\|\mathcal{X}_ZB-\mathcal{X}_{Z^*}B^*\|^2$. As all above, we choose $\rho=1$ and $C_3=1$ in Theorem 4.4.1.

First of all, for any given $Z\in\bar{\mathcal{Z}}_\tau$, we consider the following testing function:

$$
\phi_Z=\mathbb{I}\left\{\|P_{Z\cup Z^*}(Y-\mathcal{X}_{Z^*}B^*)\|^2>a(\epsilon_\tau+\epsilon_{\tau^*})\right\},
$$

where $P_{Z\cup Z^*}$ is the projection matrix on the space spanned by the column vectors of $\mathcal{X}_Z$ and X_{Z^*}, then $d(Z\cup Z^*)\overset{def}{=}\mathsf{Tr}(P_{Z\cup Z^*})=\mathsf{rank}(P_{Z\cup Z^*})\leq\ell_\tau+\ell_{\tau^*}$. Then set

$$
\phi=\max_{Z\in\cup_{\tau\in\mathcal{T}}\bar{\mathcal{Z}}_\tau}\phi_Z.
$$

When a is chosen to be a large constant, there exists some constant $C>3$ such that

$$
P_{\mathcal{X}_{Z^*}B^*}\phi\leq\sum_{\tau\in\mathcal{T}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}P_{\mathcal{X}_{Z^*}B^*}\left(\|P_{Z\cup Z^*}(Y-\mathcal{X}_{Z^*}B^*)\|^2>a(\epsilon_\tau+\epsilon_{\tau^*})\right)
$$

$$
\begin{aligned}
&\leq \sum_{\tau\in\mathcal{T}}\sum_{Z\in\bar{\mathcal{Z}}_\tau} P\left(\chi^2_{\ell_\tau+\ell_{\tau^*}} > 10(\epsilon_\tau+\epsilon_{\tau^*})\right)\\
&\leq \sum_{\tau\in\mathcal{T}}\sum_{Z\in\bar{\mathcal{Z}}_\tau} \exp\left(-C(\epsilon_{\tau^*}+\epsilon_\tau)\right)\\
&\leq \exp\left(-C\epsilon_{\tau^*}\right)\sum_{\tau\in\mathcal{T}}\exp\left(-(C-1)\epsilon_\tau\right)\\
&\leq \exp\left(-(C-\log 6)\epsilon_{\tau^*}\right),
\end{aligned}
$$

where we have used the third formula in Lemma 7.2 in [28]. For any $Z\in\bar{\mathcal{Z}}_\tau$ and $B\in\mathbb{R}^{\ell_\tau}$ such that $\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\geq\epsilon^2$, we have

$$
\begin{aligned}
P_{\mathcal{X}_Z B}(1-\phi) &\leq P_{\mathcal{X}_Z B}(1-\phi_Z) = P_{\mathcal{X}_Z B}\left(\|P_{Z\cup Z^*}(Y-\mathcal{X}_{Z^*}B^*)\|^2\leq 10(\epsilon_\tau+\epsilon_{\tau^*})\right)\\
&\leq P\left(\chi^2_{d(Z\cup Z^*)}\geq\frac{1}{2}\epsilon^2-10(\epsilon_\tau+\epsilon_{\tau^*})\right)\\
&\leq \exp\left(C_2(\epsilon_\tau+\epsilon_{\tau^*})-\frac{1}{8}\epsilon^2\right).
\end{aligned}
$$

Then the condition (4.15) in Theorem 4.4.1 holds. Now we check the condition (4.18). Consider the prior ratio

$$
\begin{aligned}
&\frac{\Gamma_Z\left(\left\{B\in\mathbb{R}^{\ell_\tau}:\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\leq\epsilon^2\right\}\right)}{\Gamma_{Z^*}\left(\left\{B\in\mathbb{R}^{\ell_{\tau^*}}:\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\leq\epsilon_{\tau^*}\right\}\right)}\\
=\ &\frac{\exp\left(\lambda\|\mathcal{X}_{Z^*}B^*\|\right)\Gamma_Z\left(\left\{B\in\mathbb{R}^{\ell_\tau}:\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\leq\epsilon^2\right\}\right)}{\exp\left(\lambda\|\mathcal{X}_{Z^*}B^*\|\right)\Gamma_{Z^*}\left(\left\{B\in\mathbb{R}^{\ell_{\tau^*}}:\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\leq\epsilon_{\tau^*}\right\}\right)}.
\end{aligned}
$$

At first, we build an upper bound for the numerator. Suppose $\mathcal{X}_Z = U_Z\Sigma_Z V_Z^T$ is the condensed singular value decomposition of $\mathcal{X}_Z$ and set $B_Z^* = (\mathcal{X}_Z^T\mathcal{X}_Z)^{-1}\mathcal{X}_Z^T\mathcal{X}_{Z^*}B^*$, $b=\Sigma_Z V_Z^T B$, $b^*=\Sigma_Z V_Z^* B_Z^*$. Then,

$$
\begin{aligned}
&\exp\left(\lambda\|\mathcal{X}_{Z^*}B^*\|\right)\Gamma_Z\left(\left\{B\in\mathbb{R}^{\ell_\tau}:\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\leq\epsilon^2\right\}\right)\\
=\ &\frac{\sqrt{\det(\mathcal{X}_Z^T\mathcal{X}_Z)}}{2}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell_\tau}\frac{\Gamma(\ell_\tau/2)}{\Gamma(\ell_\tau)}
\end{aligned}
$$

$$
\begin{aligned}
&\times \int_{B\in\mathbb{R}^{\ell_\tau}:\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\le\epsilon^2} \exp\left(\lambda\|\mathcal{X}_{Z^*}B^*\| - \lambda\|\mathcal{X}_Z B\|\right) dB \\
\le\ & \int_{B\in\mathbb{R}^{\ell_\tau}:\|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\le\epsilon^2} \frac{\sqrt{\det(\mathcal{X}_Z^T\mathcal{X}_Z)}}{2}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell_\tau}\frac{\Gamma(\ell_\tau/2)}{\Gamma(\ell_\tau)}\exp\left(\lambda\epsilon\right) dB \\
\le\ & \int_{B\in\mathbb{R}^{\ell_\tau}:\|\mathcal{X}_Z (B-B_Z^*)\|^2\le\epsilon^2} \frac{\sqrt{\det(\mathcal{X}_Z^T\mathcal{X}_Z)}}{2}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell_\tau}\frac{\Gamma(\ell_\tau/2)}{\Gamma(\ell_\tau)}\exp\left(\lambda\epsilon\right) dB \\
=\ & \int_{b\in\mathbb{R}^{\ell_\tau}:\|b-b_Z^*\|^2\le\epsilon^2} \frac{1}{2}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell_\tau}\frac{\Gamma(\ell_\tau/2)}{\Gamma(\ell_\tau)}\exp\left(\lambda\epsilon\right) db \\
\le\ & \frac{\Gamma(\ell_\tau/2)}{2\Gamma(\ell_\tau)}\exp\left(\lambda\epsilon + 2\ell_\tau\log\frac{\lambda\epsilon}{\sqrt{\ell_\tau}}\right).
\end{aligned}
$$

Note that

$$
\lambda\epsilon + 2\ell_\tau\log\frac{\lambda}{\sqrt{\ell_\tau}} \le \xi_1\lambda^2 + \frac{1}{4\xi}\epsilon^2 + 2\sqrt{\ell_\tau}\lambda\epsilon \le \xi_1\lambda^2 + \frac{1}{4\xi_1}\epsilon^2 + 4\xi_2\ell_\tau\lambda^2 + \frac{1}{4\xi_2}\epsilon^2.
$$

Choose $\xi_1 = \xi_2 = 16$, then $\lambda\epsilon + 2\ell_\tau\log\frac{\lambda}{\sqrt{\ell_\tau}} \le 80\lambda^2\epsilon_\tau + \frac{1}{32}\epsilon^2$. Plug this in the upper bound, we have

$$
\begin{aligned}
&\exp\left(\lambda\|\mathcal{X}_{Z^*}B^*\|\right)\Gamma_Z\left(\left\{B\in\mathbb{R}^{\ell_\tau} : \|\mathcal{X}_Z B - \mathcal{X}_{Z^*}B^*\|^2 \le \epsilon^2\right\}\right) \qquad (5.62)\\
\le\ & \frac{\Gamma(\ell_\tau/2)}{2\Gamma(\ell_\tau)}\exp\left(80\lambda^2\epsilon_\tau + \frac{1}{32}\epsilon^2\right).
\end{aligned}
$$

Now we build the lower bound for the denominator. In the same way, we assume $X_{Z^*} = U\Sigma V^T$, then set $b = \Sigma V^T B$ and $b^* = \Sigma V^T B^*$. Then we have

$$
\begin{aligned}
&\exp\left(\lambda\|\mathcal{X}_{Z^*}B^*\|\right)\Gamma_{Z^*}\left(\left\{B\in\mathbb{R}^{\ell_{\tau^*}} : \|\mathcal{X}_Z B - \mathcal{X}_{Z^*}B^*\|^2 \le \epsilon_{\tau^*}\right\}\right) \\
=\ & \frac{\sqrt{\det(X_{Z^*}^T X_{Z^*})}}{2}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell_{\tau^*}}\frac{\Gamma(\ell_{\tau^*}/2)}{\Gamma(\ell_{\tau^*})} \\
&\times \int_{B\in\mathbb{R}^{\ell_{\tau^*}}:\|X_{Z^*}(B-B^*)\|^2\le\epsilon_{\tau^*}} \exp\left(\lambda\|\mathcal{X}_{Z^*}B^*\| - \lambda\|X_{Z^*}B\|\right) dB \\
\ge\ & \frac{\sqrt{\det(X_{Z^*}^T X_{Z^*})}}{2}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell_{\tau^*}}\frac{\Gamma(\ell_{\tau^*}/2)}{\Gamma(\ell_{\tau^*})}
\end{aligned}
$$

$$
\begin{aligned}
&\times \int_{B\in\mathbb{R}^{\ell_{\tau^*}}:\|X_{Z^*}(B-B^*)\|^2\le\epsilon_{\tau^*}} \exp\left(-\lambda\|X_{Z^*}(B-B^*)\|\right) dB \\
\ge\ & \int_{b\in\mathbb{R}^{\ell_{\tau^*}}:\|b-b^*\|^2\le\epsilon_{\tau^*}} \frac{\Gamma(\ell_{\tau^*}/2)}{2\Gamma(\ell_{\tau^*})}\left(\frac{\lambda}{\sqrt{\pi}}\right)^{\ell_{\tau^*}} \exp\left(-\lambda\|b-b^*\|\right) db \\
\ge\ & \frac{\Gamma(\ell_{\tau^*}/2)}{2\Gamma(\ell_{\tau^*})}\exp\left(-\lambda\sqrt{\epsilon_{\tau^*}}-\ell_{\tau^*}\log\frac{\lambda}{\sqrt{\pi}}+\frac{\ell_{\tau^*}}{2}\log\frac{\pi\epsilon_{\tau^*}}{\ell_{\tau^*}}\right) \\
\ge\ & \frac{\Gamma(\ell_{\tau^*}/2)}{2\Gamma(\ell_{\tau^*})}\exp\left(-\lambda^2-\epsilon_{\tau^*}-\ell_{\tau^*}\log\frac{\lambda}{\sqrt{\pi}}+\frac{\ell_{\tau^*}}{2}\log\frac{\pi\epsilon_{\tau^*}}{\ell_{\tau^*}}\right)
\end{aligned}
$$

Note that $\epsilon_{\tau^*} \ge \ell_{\tau^*}$,

$$
\begin{aligned}
&\exp\left(\lambda\|\mathcal{X}_{Z^*}B^*\|\right)\Gamma_{Z^*}\left(\left\{B\in\mathbb{R}^{\ell_{\tau^*}} : \|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\le\epsilon_{\tau^*}\right\}\right) \\
\ge\ & \frac{\Gamma(\ell_{\tau^*}/2)}{2\Gamma(\ell_{\tau^*})}\exp\left(-\left(\lambda^2+1+\left|\log\frac{\lambda}{\sqrt{\pi}}\right|\right)\epsilon_{\tau^*}\right) \qquad (5.63)
\end{aligned}
$$

Combining (5.62) and (5.63), we have

$$
\frac{\Gamma_Z\left(\left\{B\in\mathbb{R}^{\ell_\tau} : \|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\le\epsilon^2\right\}\right)}{\Gamma_{Z^*}\left(\left\{B\in\mathbb{R}^{\ell_{\tau^*}} : \|\mathcal{X}_Z B-\mathcal{X}_{Z^*}B^*\|^2\le\epsilon_{\tau^*}\right\}\right)} \le \frac{\delta(Z)}{\delta(Z^*)}\exp\left(\frac{1}{32}\epsilon^2+C_4\epsilon_\tau+C_5\epsilon_{\tau^*}\right),
$$

where $\delta(Z)=\frac{\Gamma(\ell_\tau/2)}{\Gamma(\ell_\tau)}$ for $Z\in\bar{\mathcal{Z}}_\tau$ and $C_4=80\lambda^2$, $C_5=\lambda^2+1+\left|\log\frac{\lambda}{\sqrt{\pi}}\right|$.

Finally, we check condition (4.19) in Theorem 4.4.1. As $\Gamma_Z(\Theta_{Z'})=0$ for $Z\neq Z'$, we have $\gamma(Z)=w(\tau)\nu_\tau(Z)$ for $Z\in\bar{\mathcal{Z}}$. Then

$$
\begin{aligned}
&\sum_{\tau\in\mathcal{T}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\frac{\gamma(Z)\delta(Z)}{w(\tau^*)\nu_{\tau^*}(Z^*)\delta(Z^*)}\exp\left((C_2+C_4)\epsilon_\tau\right) \\
=\ & \sum_{\tau\in\mathcal{T}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\frac{\exp\left(-D\epsilon_\tau\right)\frac{1}{|\bar{\mathcal{Z}}_\tau|}}{\exp\left(-D\epsilon_{\tau^*}\right)\frac{1}{|\bar{\mathcal{Z}}_{\tau^*}|}}\exp\left((C_2+C_4)\epsilon_\tau\right) \\
\le\ & \exp\left((D+1)\epsilon_{\tau^*}\right)\sum_{\tau\in\mathcal{T}}\exp\left(-(D-C_2-C_4)\epsilon_\tau\right).
\end{aligned}
$$

By the third formula in Lemma 7.2 in [28], if we choose $D_0=C_2+C_4+2$, when $D>D_0$,

we will have

$$\sum_{\tau\in\mathcal{T}}\sum_{Z\in\bar{\mathcal{Z}}_\tau}\frac{\gamma(Z)\delta(Z)}{w(\tau^*)\nu_{\tau^*}(Z^*)\delta(Z^*)}\exp\left((C_2+C_4)\epsilon_\tau\right)\leq\exp\left((D+1+\log 6)\epsilon_{\tau^*}\right).$$

Finally, the proof is complete by applying Theorem 4.4.1.

□

REFERENCES

[1] Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *arXiv preprint arXiv:1706.09293*, 2017.

[2] Julyan Arbel, Ghislaine Gayraud, and Judith Rousseau. Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian journal of statistics*, 40(3):549–570, 2013.

[3] A Babenko and E Belitser. Oracle convergence rate of posterior under projection prior and bayesian model selection. *Mathematical methods of statistics*, 19(3):219–245, 2010.

[4] Andrew Barron, Mark J Schervish, and Larry Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

[5] Andrew R Barron. *The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions*. Department of Statistics, University of Illinois, 1988.

[6] Eduard Belitser and Farida Enikeeva. Empirical bayesian test of the smoothness. *Mathematical Methods of Statistics*, 17(1):1–18, 2008.

[7] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[8] Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943, 2013.

[9] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[10] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[11] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[12] Peter Carbonetto, Matthew Stephens, et al. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis*, 7(1):73–108, 2012.

[13] Ismaël Castillo. Lower bounds for posterior rates with gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299, 2008.

[14] Ismaël Castillo. On bayesian supremum norm contraction rates. *The Annals of Statistics*, 42(5):2058–2091, 2014.

[15] Ismaël Castillo, Romain Mismer, et al. Empirical bayes analysis of spike and slab posterior distributions. *Electronic Journal of Statistics*, 12(2):3953–4001, 2018.

[16] Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

[17] Ismaël Castillo, Botond Szabó, et al. Spike and slab empirical bayes sparse credible sets. *Bernoulli*, 26(1):127–158, 2020.

[18] Ismaël Castillo and Aad van der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101, 2012.

[19] Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Consistency of variational bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12(2):2995–3035, 2018.

[20] Bradley Efron. Empirical bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association*, 104(487):1015–1028, 2009.

[21] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[22] Bradley Efron, John D Storey, and Robert Tibshirani. *Microarrays empirical bayes methods, and false discovery rates*. Department of Statistics, Stanford University, 2001.

[23] Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[24] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

[25] Felix Friedrich, Angela Kempe, Volkmar Liebscher, and Gerhard Winkler. Complexity penalized m-estimation: fast computation. *Journal of Computational and Graphical Statistics*, 17(1):201–224, 2008.

[26] Chao Gao, Fang Han, and Cun-Hui Zhang. Minimax risk bounds for piecewise constant models. *arXiv preprint arXiv:1705.06386*, 2017.

[27] Chao Gao, Zongming Ma, and Harrison H Zhou. Sparse cca: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017.

[28] Chao Gao, Aad W van der Vaart, and Harrison H Zhou. A general framework for bayes structured linear models. *arXiv preprint arXiv:1506.02174*, 2015.

[29] Chao Gao and Harrison H Zhou. Rate exact bayesian adaptation with modified block priors. *The Annals of Statistics*, 44(1):318–345, 2016.

[30] Fengnan Gao, Aad van der Vaart, et al. Posterior contraction rates for deconvolution of dirichlet-laplace mixtures. *Electronic Journal of Statistics*, 10(1):608–627, 2016.

[31] Subhashis Ghosal, Jayanta K Ghosh, and Aad W van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000.

[32] Subhashis Ghosal and Aad Van Der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

[33] Qiyang Han. Bayes model selection. *arXiv preprint arXiv:1704.07513*, 2017.

[34] Marc Hoffmann, Judith Rousseau, and Johannes Schmidt-Hieber. On adaptive posterior concentration rates. *The Annals of Statistics*, 43(5):2259–2295, 2015.

[35] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[36] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[37] Wenhua Jiang, Cun-Hui Zhang, et al. General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

[38] Iain M Johnstone. Gaussian estimation: Sequence and wavelet models. *Manuscript, December*, 2011.

[39] Iain M Johnstone and Bernard W Silverman. Empirical bayes selection of wavelet thresholds. *Annals of Statistics*, pages 1700–1752, 2005.

[40] Iain M Johnstone, Bernard W Silverman, et al. Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

[41] Bartek T Knapik, Botond T Szabó, Aad W Van Der Vaart, and J Harry van Zanten. Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probability Theory and Related Fields*, 164(3-4):771–813, 2016.

[42] Willem Kruijer, Judith Rousseau, and Aad Van Der Vaart. Adaptive bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

[43] John D Lafferty and David M Blei. Correlated topic models. In *Advances in neural information processing systems*, pages 147–154, 2006.

[44] L LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

[45] Cathy Maugis-Rabusseau and Bertrand Michel. Adaptive density estimation for clustering with gaussian mixtures. *ESAIM: Probability and Statistics*, 17:698–724, 2013.

[46] Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational bayes. *arXiv preprint arXiv:1712.08983*, 2017.

[47] Sonia Petrone, Judith Rousseau, and Catia Scricciolo. Bayes and empirical bayes: do they merge? *Biometrika*, 101(2):285–302, 2014.

[48] Iosif Pinelis. Monotonicity properties of the relative error of a padé approximation for mills' ratio. *J. Inequal. Pure Appl. Math*, 3(2):1–8, 2002.

[49] Anil Raj, Matthew Stephens, and Jonathan K Pritchard. faststructure: variational inference of population structure in large snp data sets. *Genetics*, 197(2):573–589, 2014.

[50] Kolyan Ray et al. Bayesian inverse problems with non-conjugate priors. *Electronic Journal of Statistics*, 7:2516–2549, 2013.

[51] Kolyan Ray and Botond Szabo. Variational bayes for high-dimensional linear regression with sparse priors. *arXiv preprint arXiv:1904.07150*, 2019.

[52] Vincent Rivoirard and Judith Rousseau. Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis*, 7(2):311–334, 2012.

[53] Herbert Robbins. *An empirical Bayes approach to statistics*. Office of Scientific Research, US Air Force, 1955.

[54] Judith Rousseau and Botond Szabo. Asymptotic behaviour of the empirical bayes posteriors associated to maximum marginal likelihood estimator. *The Annals of Statistics*, 45(2):833–865, 2017.

[55] Lorraine Schwartz. On bayes procedures. *Probability Theory and Related Fields*, 4(1):10–26, 1965.

[56] Weining Shen and Subhashis Ghosal. Adaptive bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, 42(4):1194–1213, 2015.

[57] Xiaotong Shen and Larry Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

[58] Matthew Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

[59] Erik B Sudderth and Michael I Jordan. Shared segmentation of natural scenes using dependent pitman-yor processes. In *Advances in Neural Information Processing Systems*, pages 1585–1592, 2009.

[60] Botond Szabó, Aad W Van Der Vaart, JH van Zanten, et al. Frequentist coverage of adaptive nonparametric bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428, 2015.

[61] BT Szabó, AW van der Vaart, JH van Zanten, et al. Empirical bayes scaling of gaussian priors in the white noise model. *Electronic Journal of Statistics*, 7:991–1018, 2013.

[62] Ronald Thisted and Bradley Efron. Did shakespeare write a newly-discovered poem? *Biometrika*, 74(3):445–455, 1987.

[63] AW van der Vaart and JH van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

[64] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[65] Nicolas Verzelen et al. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

[66] Stephen Walker and Nils Lid Hjort. On bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821, 2001.

[67] Stephen G Walker, Antonio Lijoi, and Igor Prünster. On rates of convergence for posterior distributions in infinite-dimensional models. *The Annals of Statistics*, 35(2):738–746, 2007.

[68] Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, pages 1–15, 2018.

[69] Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *arXiv preprint arXiv:1710.03266*, 2017.

[70] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[71] Anderson Y Zhang and Harrison H Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *arXiv preprint arXiv:1710.11268*, 2017.

[72] Anderson Y Zhang, Harrison H Zhou, et al. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.

[73] Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *arXiv preprint arXiv:1712.02519*, 2017.

[74] Tong Zhang. From ϵ-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

[75] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.