# Ultra High-dimensional Multivariate Posterior Contraction Rate under Shrinkage Priors 

Ruoyang Zhang ${ }^{\text {a }}$, Malay Ghosh ${ }^{\text {a }}$<br>${ }^{a}$ Department of Statistics, University of Florida, Gainesville, FL 32611, United States


#### Abstract

In recent years, shrinkage priors have received much attention in high-dimensional data analysis from a Bayesian perspective. Compared with widely used spike-and-slab priors, shrinkage priors have better computational efficiency. But the theoretical properties, especially posterior contraction rate, which is important in uncertainty quantification, are not established in many cases. In this paper, we apply global-local shrinkage priors to high-dimensional multivariate linear regression with unknown covariance matrix. We show that when the prior is highly concentrated near zero and has heavy tail, the posterior contraction rates for both coefficients matrix and covariance matrix are nearly optimal. Our results hold when number of features p grows much faster than the sample size n , which is of great interest in modern data analysis. We show that a class of readily implementable scale mixture of normal priors satisfies the conditions of the main theorem.


Keywords: multivariate regression, unknown covariance matrix, Gaussian scale mixture

## 1. Introduction

Parameter estimation, variable selection and prediction in high dimensional regression models have received significant attention in these days, particularly when the number of regressors $p$ is much larger than the number of observations

[^0]n. Examples abound - brain imaging, microarray experiments, satellite data analysis, just to name a few. In many of these examples, one key issue is to address sparsity of effective regression parameters in the midst of a multitude of inactive ones. For example, there are only a few significant genes associated with Type I diabetes along with million others of no direct impact for such a disease.

In a frequentist framework, the most commonly used approach for inducing sparsity is by imposing regularization penalty on the parameters of interest. The most popular ones are $\ell_{1}$ (lasso) and $\ell_{2}$ (ridge) penalties or a combination of these (elastic net). The $\ell_{1}$ and $\ell_{2}$ regularization can naturally be extended to multivariate case where sparsity in the coefficient matrix is desired. Rothman et al. [26] used $\ell_{1}$ penalties on each entry of the coefficient matrix as well as on each off-diagonal element of the covariance matrix. Wilms and Croux 33 ] considered a model which put an $\ell_{2}$ penalty on the rows of coefficient matrix to shrink the entire row to zero, and an $\ell_{1}$ penalty on the off-diagonal elements of the inverse error covariance matrix. Li et al. 15 proposed a multivariate sparse group lasso imposing $\ell_{2}$ penalty on the rows of the regression matrix and in addition an $\ell_{1}$ penalty on individual coefficient of the regression matrix to perform sparse estimation and variable selection both at the between and within group levels.

In a Bayesian setting, spike-and-slab priors, originally introduced by Mitchell and Beauchamp [17] have become very popular for handling sparsity. Spike-and-slab priors are mixture densities with positive mass at zero to force some parameters to be zero, and a continuous density to model the nonzero coefficients. These priors have been used in a variety of contexts. For example, for Bayesian Group Lasso, Xu and Ghosh [34 used these priors for both variable selection and estimation. This work was extended by Liquet et al. [16] to the multivariate case. More recently, Ročková and George [24] introduced spike-and-slab lasso for variable selection and estimation. Deshpande et al. 11] extended it to multivariate case by putting spike-and-slab prior on each entry of the coefficient matrix as well as on each off-diagonal element of the precision
matrix.
Spike-and-slab priors face severe computational challenges, when $p$, the number of regressors, is very large. This is due to the fact that one needs to search over $2^{p}$ possible models. Bai and Ghosh 4] provided an alternative to these priors by introducing global-local shrinkage priors. These priors approximate the spike-and-slab priors well and are usually much easier to implement because they are continuous. Like spike-and-slab priors, global-local shrinkage priors also put significant probability around zero, but retain heavy enough tails so that the true signals are very unlikely to be missed.

Bai and Ghosh [4] considered the case when the number of regressors can grow at a sub-exponential rate when compared to the sample size. They established posterior consistency of their prior and showed that the insignificant regression coefficients converge to zero at an exponential rate. Song and Liang [28] provided some general posterior contraction rates in the context of variable selection and estimation in univariate regression models with unknown variance.

Our paper is a follow-up of the works by Bai and Ghosh [4] and Song and Liang [28]. In particular, unlike the former, we do not need to assume a known covariance matrix in the original regression model to establish exponential convergence rate of tail probabilities. We propose a set of general conditions on continuous prior for achieving nearly-optimal posterior contraction rate for both coefficient matrix and covariance matrix. This extends the work of Song and Liang [28] to the multivariate case. Also, we have demonstrated that these regulatory conditions are satisfied by a general class of global-local shrinkage priors. Our technical results borrowed tools developed by Song and Liang [28], but handling multivariate data presented some new challenges in proving the results.

Ning and Ghosal [19] also addressed the issue of variable selection with unknown covariance matrix and established posterior consistency result similar to ours. But their results are based on spike-and-slab priors instead of globallocal shrinkage priors and utilized different techniques from ours.

This paper is organized as follows. In Section 2, we establish general condi-
tions on priors for achieving nearly-optimal posterior contraction rate for both coefficient matrix and covariance matrix. In Section 3, a class of global-local shrinkage prior that satisfies these general conditions is proposed. In Section 4, finite sample performance of the proposed model is evaluated through numerical experiments. Some final remarks are made in Section 5. Most of the technical theorems and lemmas are relegated to the Appendix.

## 2. Posterior Contraction Rate

### 2.1. Problem Setting

We consider the following multivariate linear regression model

$$
\begin{equation*}
Y_{i}=X_{i} \boldsymbol{B}_{\boldsymbol{n}}+\varepsilon_{i} \boldsymbol{\Sigma}_{n}^{1 / 2}, \quad i=1, \cdots, n \tag{1}
\end{equation*}
$$

where $Y_{i}$ is a $1 \times q_{n}$ response vector, and the correlation of responses is assumed to be captured by the $q_{n} \times q_{n}$ covariance matrix $\boldsymbol{\Sigma}_{n} . \boldsymbol{B}_{\boldsymbol{n}}$ is a $p_{n} \times q_{n}$ coefficient matrix, $X_{i}$ is a $1 \times p_{n}$ regressor vector, $\varepsilon_{i}$ is a $1 \times q_{n}$ noise vector. Throughout this paper, $\varepsilon_{i}$ 's are assumed to have i.i.d multivariate normal $\mathcal{N}\left(0, I_{q_{n}}\right)$ distribution, $i=1, \cdots, n$. Subscripts $n$ denotes that the quantity can vary with $n$. In matrix form, Model (1) can be written as

$$
\begin{equation*}
\boldsymbol{Y}_{\boldsymbol{n}}=\boldsymbol{X}_{\boldsymbol{n}} \boldsymbol{B}_{\boldsymbol{n}}+\boldsymbol{\varepsilon}_{\boldsymbol{n}} \boldsymbol{\Sigma}_{n}^{1 / 2} \tag{2}
\end{equation*}
$$

where $\boldsymbol{Y}_{\boldsymbol{n}}=\left(Y_{1}^{T}, \cdots, Y_{n}^{T}\right)^{T}, \boldsymbol{X}_{\boldsymbol{n}}=\left(X_{1}^{T}, \cdots, X_{n}^{T}\right)^{T}$ and $\varepsilon_{\boldsymbol{n}}=\left(\varepsilon_{1}^{T}, \cdots, \varepsilon_{n}^{T}\right)^{T}$. Throughout the paper, for notational simplicity, subscript $n$ for $\boldsymbol{Y}_{n}, \boldsymbol{X}_{n}$ and $\boldsymbol{B}_{n}$ will be dropped when there is no ambiguity.

For estimation of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, we consider the following Bayesian multivariate linear regression model. This model puts independent prior on each row vector of $\boldsymbol{B}$ conditioning on $\boldsymbol{\Sigma}$ and an Inverse-Wishart prior for $\boldsymbol{\Sigma}$. General conditions for $\pi(\boldsymbol{B} \mid \boldsymbol{\Sigma})$ for establishing a satisfying posterior contraction rate of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$
is given in Theorem (1).

$$
\begin{align*}
& Y_{i} \mid X_{i}, \boldsymbol{B}, \boldsymbol{\Sigma} \stackrel{i n d}{\sim} \mathcal{N}_{q_{n}}\left(X_{i} \boldsymbol{B}, \boldsymbol{\Sigma}\right) \quad i=1, \cdots, n \\
& \boldsymbol{B}_{j} \mid \boldsymbol{\Sigma} \stackrel{i n d}{\sim} \pi\left(\boldsymbol{B}_{j} \mid \boldsymbol{\Sigma}\right) \quad j=1, \cdots, p_{n}  \tag{3}\\
& \boldsymbol{\Sigma} \sim \mathcal{I} \mathcal{W}_{q_{n}}(v, \Phi)
\end{align*}
$$

where $\boldsymbol{B}_{j}$ is the $j$ th row of $\boldsymbol{B} .{\mathcal{I} \mathcal{W}_{q_{n}}}(v, \Phi)$ means a $q_{n}$-dimensional InverseWishart distribution with degree of freedom $v>q_{n}-1$ and a $q_{n} \times q_{n}$ positive definite scale matrix $\Phi$.

### 2.2. Notations

First, a few notations used throughout the paper are defined. We write $a \vee b$ for $\max (a, b)$, where $a$ and $b$ are real numbers. Letters $C, c, k$ with subscripts denote generic positive constants that do not depend on $n$. For two sequences of positive real numbers $a_{n}$ and $b_{n}, a_{n} \lesssim b_{n}$ is equivalent to $a_{n}=\mathcal{O}\left(b_{n}\right)$, i.e. there exists constant $C>0$ such that $a_{n} \leq C b_{n}$ for all large $n$. $a_{n} \prec b_{n}$ means $a_{n}=o\left(b_{n}\right)$, that is, $a_{n} / b_{n} \rightarrow 0$ as $n \rightarrow \infty . a_{n} \simeq b_{n}$ denotes that there exists constants $0<C_{1} \leq C_{2}$ such that $C_{1} b_{n} \leq a_{n} \leq C_{2} b_{n}$.

For a vector $\boldsymbol{x} \in \mathbb{R}^{p},\|\boldsymbol{x}\|_{2}$ denotes the $\ell_{2}$ norm. For a $n \times m$ real matrix $A$ with entries $a_{i j},\|A\|_{F}:=\sqrt{\operatorname{tr}\left(A A^{T}\right)}$ denotes the Frobenius norm of $A$; $\|A\|_{2, \infty}:=\max _{1 \leq i \leq n}\left(\sum_{j=1}^{m} a_{i j}^{2}\right)^{1 / 2}$ denotes $A$ 's maximum row length; $\|A\|_{2,1}:=$ $\sum_{i=1}^{n}\left(\sum_{j=1}^{m} a_{i j}^{2}\right)^{1 / 2}$ denotes the sum of row lengths. For a symmetric real ma$\operatorname{trix} A, \lambda_{i}(A)$ denotes the $i t h$ smallest eigenvalue of $A .\|A\|=\lambda_{\max }(A)$ denotes the spectral norm of $A$, which is also the maximum eigenvalue of $A$.

### 2.3. Conditions for Posterior Contraction Rate

Suppose the data $\boldsymbol{Y}$ is generated by (1) with the true regression parameter $\boldsymbol{B}_{\mathbf{0}}$ and the true dispersion matrix $\boldsymbol{\Sigma}_{0}$. To achieve posterior contraction rate, we first state some assumptions for sparsity of $\boldsymbol{B}_{\mathbf{0}}$, the eigen-structure of design $\operatorname{matrix} \boldsymbol{X}$, and eigenvalues of $\boldsymbol{\Sigma}_{0}$.

Assumption 1. Sparsity of $\boldsymbol{B}_{\mathbf{0}}$ :
$A_{1}: s_{0} \log p_{n} \prec n$, where $s_{0}$ is the size of the true model, i.e., the number of nonzero rows in $\boldsymbol{B}_{0}$.

Assumption 2. Eigen-structure of the design matrix $\boldsymbol{X}$ :
$A_{2}(1)$ : Entries $(\boldsymbol{X})_{i j}$ in design matrix are uniformly bounded. For simplicity, assume they are bounded by 1.
$A_{2}(2): p_{n} \rightarrow \infty$ as $n \rightarrow \infty$.
$A_{2}(3)$ : There exist some integer $\bar{p}$ (depending on $n$ and $p_{n}$ ) and fixed constant $\lambda_{0}$ such that $\bar{p} \succ s_{0}$, and $\lambda_{\min }\left(X_{S}^{T} X_{S}\right) \geq n \lambda_{0}$ for any subset model $S$ with $|S| \leq \bar{p}$.

Assumption 3. Dimension and eigenvalues of $\boldsymbol{\Sigma}_{0}$ :
$A_{3}(1): q_{n} \lesssim \log p_{n}$.
$A_{3}(2): q_{n}^{2} \log n \prec n$.
$A_{3}(3): b_{1} \leq \lambda_{i}\left(\boldsymbol{\Sigma}_{0}\right) \leq b_{2}, \quad i=1, \cdots, q_{n}$.

Remark 2.1. Assumption (2) and Assumption (1) are the same as in [28]. Note that $A_{2}(2)$ does not restrict the rate of $p_{n}$ going to infinity. Along with $A_{1}, p_{n}$ can grow sub-exponentially fast with $n$ when $s_{0}$ is finite, e.g., $\log \left(p_{n}\right) \lesssim n^{c}$ for some $c \in(0,1)$, which is the ultrahigh dimensional setting in [4].

Remark 2.2. $A_{3}(2)$ and $A_{3}(3)$ are the same as in [19]. Different from many previous settings where the dimension of response $q_{n}$ is a fixed constant (4), [16]), here we allow $q_{n}$ to grow with $n$. However, the growth of $q_{n}$ is limited by constraints $A_{3}(1)$ and $A_{3}(2)$. When $q_{n}$ is a fixed constant, $A_{3}(1)$ and $A_{3}(2)$ are trivially satisfied.

Theorem 1. For the multivariate Bayesian model given in (3), suppose design matrix $\boldsymbol{X}$ satisfies Assumption (2) and true parameter $\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{\mathbf{0}}\right)$ satisfies Assumptions (1) and (3). Let the prior density of $\boldsymbol{B}$ be:

$$
\pi(\boldsymbol{B} \mid \boldsymbol{\Sigma})=\prod_{i=1}^{p_{n}}\left\{|\boldsymbol{\Sigma}|^{-1 / 2} g_{\tau}\left(\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right)\right\}
$$

If $g_{\tau}(\boldsymbol{x})$ satisfies

$$
\begin{align*}
& \int_{\|\boldsymbol{x}\|_{2} \geq a_{n}} g_{\tau}(\boldsymbol{x}) d \boldsymbol{x} \leq p_{n}^{-(1+u)} \text { for some } u>0  \tag{4}\\
& \log \left(\inf _{\|\boldsymbol{x}\|_{2} \leq M_{0}} g_{\tau}(\boldsymbol{x})\right) \gtrsim-\log p_{n} \tag{5}
\end{align*}
$$

where $M_{0}=\gamma\left\|\boldsymbol{B}_{0} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2, \infty}, \gamma>1, a_{n} \simeq \epsilon_{n} / p_{n}$.

Then the following posterior contraction result holds

$$
\begin{gathered}
\Pi_{n}\left(\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \geq M \epsilon_{n} \mid \boldsymbol{Y}\right) \rightarrow 0 \text { in } P_{\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)} \text {-probability } \\
\quad \Pi_{n}\left(\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\| \geq M\left\|\boldsymbol{\Sigma}_{0}\right\| \epsilon_{n} \mid \boldsymbol{Y}\right) \rightarrow 0 \text { in } P_{\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)} \text {-probability }
\end{gathered}
$$

where $\epsilon_{n}=\sqrt{s_{0} \log p_{n} / n} \vee \sqrt{q_{n}^{2} \log n / n} \vee \sqrt{q_{n} s_{0} \log n / n}$ and $M$ is a sufficiently large constant.

Remark 2.3. Conditions (4)) and (5) for $g_{\tau}(\cdot)$ have intuitive interpretation. (4) means that the prior has to be highly concentrated around a small neighborhood of $\mathbf{0}$, which corresponds to the sparsity structure of the model. Taking $\left\|\boldsymbol{B}_{\mathbf{0}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2, \infty}$ as the strength of true signal, (5) means that the prior needs to put enough mass around the true signal, which is often referred as heavy-tail condition in [21, 3, 2].

Remark 2.4. When $q_{n}$ is a fixed constant, the contraction rate $\epsilon_{n}$ becomes $\sqrt{s_{0} \log p_{n} / n}$, which is the same as the univariate optimal posterior contraction rates for regression coefficient with respect to $\ell_{1}$ and $\ell_{2}$ norm in [8, 25], where spike-and-slab priors are used. In addition, this rate is also comparable to the $\operatorname{minimax}$ rate $\sqrt{s_{0} \log \left(p_{n} / s_{0}\right) / n}$ of lasso and Dantzing selector for $\ell_{2} \operatorname{loss}$ in $\ell_{0}$ ball [23, 35. Two additional terms $\sqrt{q_{n}^{2} \log n / n}$ and $\sqrt{q_{n} s_{0} \log n / n}$ that may slower the convergence can be viewed as a compensation of allowing $q_{n} \rightarrow \infty$.

Remark 2.5. By the fact that $\left\|\boldsymbol{\Sigma}_{0}\right\|_{2} \leq\left\|\boldsymbol{\Sigma}_{0}\right\|_{F}$ and $\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\|_{F} \leq \sqrt{q_{n}} \| \boldsymbol{\Sigma}-$ $\boldsymbol{\Sigma}_{0} \|_{2}$, we get $\Pi_{n}\left(\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\|_{F} \geq M\left\|\boldsymbol{\Sigma}_{0}\right\|_{F \sqrt{q_{n}}} \epsilon_{n} \mid Y\right) \rightarrow 0$ in $P_{\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)}$-probability. Further, if $q_{n}$ is a constant, the posterior contraction rate of $\boldsymbol{\Sigma}$ under Frobenius norm is also $\sqrt{s_{0} \log p_{n} / n}$.

The complete proof of Theorem (1) is provided in Appendix. Here we briefly summarize the ideas and key steps. We applied the tools developed in 28. To extend univariate contraction results to multivariate case, spectral norm is used
for measuring matrix distance. With its relation to Frobenius norm, we are able to make straightforward interpretations.

For showing the posterior contraction results, auxiliary sets $A_{n}, B_{n}$ and $C_{n}$ are constructed as follow.

$$
\begin{aligned}
A_{n} & =\left\{\text { at least } \tilde{p} \text { entries }\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\| \text { is larger than } a_{n}\right\} \\
& \cup\left\{\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\| \geq M\left(\|\boldsymbol{\Sigma}\| \vee\left\|\boldsymbol{\Sigma}_{0}\right\|\right) \epsilon_{n}\right\} \\
& \cup\left\{\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \geq M \epsilon_{n}\right\}
\end{aligned}
$$

Define $B_{n}=\left\{\right.$ at least $\tilde{p}$ entries $\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2}$ is larger than $\left.a_{n}\right\}$, and $C_{n}=$ $A_{n} \backslash B_{n}$. Let $\theta=(\boldsymbol{B}, \boldsymbol{\Sigma})$ and $\theta_{0}=\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)$. It suffices to show $P_{\theta_{0}}\left(\Pi_{n}\left(A_{n} \mid \boldsymbol{Y}\right) \geq\right.$ $\left.e^{-\tilde{c_{1}} n \epsilon_{n}^{2}}\right) \leq e^{-\tilde{c_{2}} n \epsilon_{n}^{2}}$. By Lemma A. 4 in [28], the proof is composed of three parts:
(1) Construction of test $\phi_{n}$ satisfies $\mathbb{E}_{\theta_{0}} \phi_{n} \leq e^{-k_{2} n \epsilon_{n}^{2}}$ and $\sup _{\theta \in C_{n}} \mathbb{E}_{\theta}\left(1-\phi_{n}\right) \leq$ $e^{-k_{3} n \epsilon_{n}^{2}}$.
(2) Showing event $B_{n}$ has very small probability under the specified prior.
(3) Demonstrating the marginal probability of data is highly likely to be bounded away from 0 if data is generated with true parameters. Probability bounds of Inverse Wishart distribution 19 are applied in this part.

### 2.4. Variable Selection Consistency

Different from spike-and-slab priors, continuous global-local shrinkage priors put zero probability at the point $\mathbf{0}$, so the solution is not sparse naturally. In this subsection, variable selection criteria and corresponding selection consistency property are discussed. But we want to point out that variable selection is not always required. Sometimes in practice, lacking exact zeros is deemed to be more realistic and preferred [29].

By Condition (4) in Theorem (11), where $a_{n}$ acts like a partition for "spike" and "slab" parts, the posterior model selection rule is set to be $S_{n}=\{j$ : $\left.\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|>a_{n}\right\}$. Consistency of the selection rule is established in the following theorem.

Theorem 2. Suppose assumptions and conditions for Theorem (1) hold with $a_{n} \prec \sqrt{\log p_{n} / n} / p_{n}$ and $u>1$ in (4). Let $\mathcal{B}_{j, \epsilon_{n}}:=\operatorname{Ball}\left(\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}, c_{0} \epsilon_{n}\right)$, i.e.,
a $q_{n}$-dimensional ball centering at $\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}$ with radius $c_{0} \epsilon_{n}$, where $c_{0}>0$ is a constant. Let $S_{n}$ be the posterior subset model: $S_{n}=\left\{j:\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|>a_{n}\right\}$. Suppose $\boldsymbol{B}_{\mathbf{0}}$ and $g_{\tau}(\boldsymbol{x})$ satisfies

$$
\begin{align*}
& \min _{j \in S_{0}}\left\|\boldsymbol{B}_{\mathbf{0}_{j}}\right\| \geq M_{1} \epsilon_{n} \text { for some large constant } M_{1},  \tag{6}\\
& s_{0} \log l_{n} \prec \log p_{n} \text { where } l_{n}=\max _{j \in S_{0}} \sup _{\boldsymbol{x}_{1}, \boldsymbol{x}_{2} \in \mathcal{B}_{j, \epsilon_{n}}} \frac{g_{\tau}\left(\boldsymbol{x}_{1}\right)}{g_{\tau}\left(\boldsymbol{x}_{2}\right)}, \tag{7}
\end{align*}
$$

then $\Pi_{n}\left(S_{n}=S_{0} \mid \boldsymbol{Y}\right) \rightarrow 1$ in $P_{\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)}$.

Remark 2.6. Compared to Theorem (1), this theorem holds with a more concentrated prior peak implied by smaller $a_{n}$ and larger $u$. In addition, condition (6) requires minimal strength of the true coefficients. Intuitively, if a true parameter is too small, it would be hard to distinguish it from zero. In condition (7), $l_{n}$ can be viewed as a measurement of "flatness" of $g_{\tau}(\boldsymbol{x})$ around true coefficients. With enough mass around the truth and this flatness constraint, we can get sufficiently large prior density for points in a small neighborhood of the true parameters. With stronger conditions than Theorem (1), this result also gives a stronger posterior contraction that the false coefficients are bounded by $a_{n}$. As $a_{n} \rightarrow 0$ when $n \rightarrow \infty$, the false coefficients will diminish to zero in the limit.

## 3. Extended MBSP Model with Unknown Covariance Matrix

In previous section, we establish general conditions on the priors to obtain good posterior contraction and variable selection. In this section, we will propose a class of global-local shrinkage prior that satisfies conditions (4) and (5).

This class of priors we propose is scale mixture of Gaussians, which is closely related to the Multivariate Bayesian model with Shrinkage Priors (MBSP) introduced by [4]. In MBSP, $\boldsymbol{\Sigma}$ is assumed to be fixed and known. Here we put an Inverse-Wishart prior for $\boldsymbol{\Sigma}$, extending it to the unknown $\boldsymbol{\Sigma}$ case and obtain
the following Extended MBSP model:

$$
\begin{align*}
& \boldsymbol{B}_{j} \mid \xi_{j}, \boldsymbol{\Sigma} \stackrel{i \sim d}{\sim} \mathcal{N}_{q_{n}}\left(0, \tau_{n} \xi_{j} \boldsymbol{\Sigma}\right) \quad j=1, \cdots, p_{n} \\
& \xi_{j} \stackrel{i n d}{\sim} \pi\left(\xi_{j}\right)  \tag{8}\\
& \boldsymbol{\Sigma} \sim \mathcal{I} \mathcal{W}_{q_{n}}(v, \Phi)
\end{align*}
$$

In univariate case $\left(q_{n}=1\right)$, many priors can be expressed as scale mixtures of Gaussians [30. Table (1) lists such priors and corresponding mixing density $\pi(\xi)$.

Table 1: List of scale mixtures of Gaussian priors

| prior | $\pi(\xi)$ |
| :--- | :--- |
| Student's t | $\xi^{-a-1} \exp (-a / \xi)$ |
| TPBN [1] | $\xi^{u-1}(1+\xi)^{-a-u}$ |
| Horseshoe [7] | $\xi^{-1 / 2}(1+\xi)^{-1}$ |
| NEG [13] | $(1+\xi)^{-a-1}$ |
| GDP [2] | $\int_{0}^{\infty} \frac{\lambda^{2}}{2} \exp \left(-\frac{\lambda^{2} \xi}{2}\right) \lambda^{2 a-1} \exp (-\eta \lambda) d \lambda$ |
| HIB [22] | $\xi^{u-1}(1+\xi)^{-a-u} \exp \left(-\frac{s}{1+\xi}\right)\left(\phi^{2}+\frac{1-\phi^{2}}{1+\xi}\right)^{-1}$ |
| Horseshoe+ [5] | $\xi^{-1 / 2}(\xi-1)^{-1} \log \xi$ |

We now show that when the mixing component $\pi(\xi)$ follows certain polynomialtailed distribution, posterior contraction is obtained with proper global shrinkage parameter $\tau_{n}$.

Theorem 3. Suppose $\boldsymbol{B}$ follow the following prior:

$$
\begin{gather*}
\boldsymbol{B}_{j} \mid \xi_{j}, \boldsymbol{\Sigma} \stackrel{i n d}{\sim} \mathcal{N}_{q_{n}}\left(0, \tau_{n} \xi_{j} \boldsymbol{\Sigma}\right) \quad j=1, \cdots, p_{n} \\
\xi_{j} \stackrel{i n d}{\sim} \pi\left(\xi_{j}\right) \tag{9}
\end{gather*}
$$

where $\pi\left(\xi_{j}\right)$ is a polynomial-tailed distribution taking the form $\pi\left(\xi_{j}\right)=K \xi_{j}^{-r} L\left(\xi_{j}\right)$, $r>1, K>0$. If $L(\xi)$ satisfies either of the two following conditions for all $\xi>0$ :
$(C 1) 1-C_{11} \xi^{-t} \leq L(\xi) \leq C_{12}, \quad C_{12} \geq 1, C_{11}>0, t>0 ;$
$(C 2) C_{21} \xi^{-t} \leq L(\xi) \leq 1, \quad C_{21}>0, t>0$,
then (4) and (5) hold with $\tau_{n} \lesssim a_{n}^{2} p_{n}^{-\left(1+u^{\prime}\right) /(r-1)}$ for some $u^{\prime}>0$ and $\log \tau_{n} \gtrsim$ $-\log p_{n}$.

Remark 3.1. It is easy to see that $-\log p_{n} \lesssim \log \left(a_{n}^{2} p_{n}^{-\left(1+u^{\prime}\right) /(r-1)}\right)$, therefore such $\tau_{n}$ must exist.

Remark 3.2. Many commonly used shrinkage priors satisfy either ( $C 1$ ) or (C2). As shown in Table (22), mixing component $\pi(\xi)$ of student's t, TPBN (horseshoe, NEG are special cases of TPBN) and HIB satisfies ( $C 1$ ); horseshoe+ satisfies (C2). Proofs of these bounds are provided in Appendix.

Remark 3.3. For application, we recommend using TPBN prior. It has been shown in [4] that TPBN prior is easy for implementation using Gibbs sampling and relevant computation $R$ package MBSP is readily available. They compared their simulation results with other high-dimensional multivariate models.

Table 2: Bounds for $L(\xi)$

| prior | $\mathrm{L}(\xi)$ | lower bound | upper bound |
| :--- | :--- | :--- | :--- |
| Student's t | $\exp (-a / \xi)$ | $1-a \xi^{-1}$ | 1 |
| TPBN | $(\xi /(1+\xi))^{a+u}$ | $1-(a+u) \xi^{-1}$ | 1 |
| Horseshoe | $\xi /(1+\xi)$ | $1-\xi^{-1}$ | 1 |
| NEG | $(\xi /(1+\xi))^{a+1}$ | $1-(a+1) \xi^{-1}$ | 1 |
|  | $\int_{0}^{\infty} t^{a} \exp (-t-\eta \sqrt{2 t / \xi}) d t$ | $1-\sqrt{2} \eta \frac{\Gamma(a+3 / 2)}{\Gamma(a+1)} x^{-1 / 2}$ | 1 |
| GDP | $\times 1 / \Gamma(a+1)$ |  |  |
|  | $\exp \left(-\frac{s}{1+\xi}\right)\left(\phi^{2}+\frac{1-\phi^{2}}{1+\xi}\right)^{-1}$ |  | $1-(a+u) \xi^{-1}$ |
| HIB | $\times\left(\frac{\xi}{1+\xi}\right)^{a+u}\left(1 \vee \phi^{2}\right) e^{s}$ |  |  |
| Horseshoe+ | $\xi^{3 / 4}(\xi-1)^{-1} \log \xi / 4$ | 1 |  |

## 4. Numerical Experiments and Data Analysis

Through numerical experiments, we examine the uncertainty assessment for covariance matrix estimate using scale mixture of Gaussians proposed in Section 3. We explore how the difference between estimation and truth varies as $n$ and $p_{n}$ grows. More simulations that evaluate performances on coefficient matrix reconstruction, prediction as well as variable selection under various situations are presented in Bai and Ghosh [4]. A real data analysis is also given.

### 4.1. Numerical Experiments

In our simulation, horseshoe mixing density $\pi(\xi)=\xi^{-1 / 2}(1+\xi)^{-1}$ is used. We focus on performance on ultra high-dimensional and ultra-sparse setting, where $p$ is approximately $n^{1.5}$, proportion of nonzero coefficients ranges from $0.38 \%$ to $1.6 \%$. Six different experiments settings are listed below.

Experiment 1: $n=25, p=125, d=3, s_{0}=2, s_{0} / p=1.6 \%$.
Experiment 2: $n=50, p=354, d=3, s_{0}=3, s_{0} / p=0.85 \%$.
Experiment 3: $n=75, p=650, d=3, s_{0}=4, s_{0} / p=0.62 \%$.
Experiment 4: $n=100, p=1000, d=3, s_{0}=5, s_{0} / p=0.5 \%$.
Experiment 5: $n=125, p=1398, d=3, s_{0}=6, s_{0} / p=0.43 \%$.
Experiment 6: $n=150, p=1837, d=3, s_{0}=7, s_{0} / p=0.38 \%$.

In all six experiments, data are generated according to the multivariate linear regression model (1). Each row of $\boldsymbol{X}$ is generated independently from $\mathcal{N}_{p}(\mathbf{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma}_{i j}=0.5^{|i-j|}$. The true coefficient matrix $\boldsymbol{B}_{\mathbf{0}}$ is generated by uniformly selecting $s_{0}$ nonzero rows, and other rows are set to be zero. For nonzero rows, each entry is independently sampled from $\operatorname{Unif}([-5,-0.5] \cup[0.5,5])$. The true covariance matrix $\left(\boldsymbol{\Sigma}_{0}\right)_{i j}=\sigma^{2}(0.5)^{|i-j|}, \sigma^{2}=2$.

By Theorem (3), when Assumption (1)-(3) holds, the global shrinkage parameter $\tau_{n}$ for nearly minimax posterior contraction rate should satisfy $\tau_{n} \lesssim$ $a_{n}^{2} p_{n}^{-\left(1+u^{\prime}\right) /(r-1)}$ for some $u^{\prime}>0$ and $\log \tau_{n} \gtrsim-\log p_{n}$. But in application,
this value is very small, e.g., such $\tau_{n}$ in Experiment 3 would be around $10^{-13}$. Too small $\tau_{n}$ will cause problems in Gibbs sampling 31. Currently, inference for the global hyperparmeter is still an open problem[20]. Here, we set $\tau_{n}=1 /\left(p_{n} \sqrt{n \log n}\right)$, which achieves posterior consistency, although theoretical posterior contraction rate is not available [4. We use the Gibbs sampler in $R$ package MBSP, where the major computational complexity is linear in $p_{n}[6]$. Each experiment is repeated 100 times. In all experiments, Gibbs sampler is run for 15000 iterations, the first 5000 iterations are burn-in.

Posterior mean $\hat{\boldsymbol{\Sigma}}$ is taken to be the point estimators of $\boldsymbol{\Sigma}$. $\left\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right\|_{2}$ and $\left\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right\|_{F}$ are used to measure the difference between posterior estimates and the truth in two different norms. Figure (1) illustrates how $\left\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right\|_{2}$ and $\left\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right\|_{F}$ decrease as $n$ and $p$ increase. Although this trend is a finite sample behavior, it matches our posterior consistency result established in previous section.


Figure 1: Box-plots of difference between estimated and true covariance matrix. The x-axis indicates experiment number and the y-axis indicates $\left\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right\|_{2}$ (left) and $\left\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_{0}\right\|_{F}$ (right) respectively.

### 4.2. Data Analysis

We estimate the correlation between multiple responses on a yeast cell cycle data set. This data set was first analyzed by Chun and Keleş 10 and is available
in the spls package in $R$.
In molecular biology, transcription factors (TFs), also known as sequencespecific DNA-binding factors, are proteins that controls the rate of transcription of genetic information from DNA to mRNA, by binding to a specific DNA sequence. To understand the regulatory mechanism of TFs, it is important to reveal the relationship between TFs and their target genes.

In the original yeast cell cycle data set, the response Y consists of 542 cell-cycle-regulated genes from an $\alpha$ factor arrest method and mRNA levels measured every 7 minutes at 18 time points, i.e. $n=542, q=18$. The $542 \times 106$ design matrix X consists of 106 TFs' binding information, representing the strength of interaction between TFs and the target genes. This data set has been analyzed in Chen and Huang [9, Goh et al. [12], Bai and Ghosh [4] with various variable selection and estimation methods for regression matrix $\boldsymbol{B}$.

Here, we focus on the estimation of covariance $\boldsymbol{\Sigma}$ between responses. Because no sparsity is assumed in $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}^{-1}, q=18$ is too large to get accurate estimation. We only use the first four measurements, i.e. $q=4$. The point estimator obtained for covariance matrix between responses is

$$
\hat{\boldsymbol{\Sigma}}=\left[\begin{array}{cccc}
0.39 & 0.17 & 0.07 & -0.03 \\
0.17 & 0.30 & 0.13 & 0.04 \\
0.07 & 0.13 & 0.28 & 0.17 \\
-0.03 & 0.04 & 0.17 & 0.25
\end{array}\right]
$$

Note that $Y_{i}$ 's are mRNA levels measured every 7 minutes, it would be natural to observe autocorrelation as demonstrated in the above estimator.

## 5. Conclusion and Future Work

This paper has several contributions. First, we propose a set of general conditions for continuous prior $\pi(\boldsymbol{B} \mid \boldsymbol{\Sigma})$ in sparse multivariate Bayesian estimation that can achieve nearly-optimal posterior contraction rate. While previous Bayesian multivariate models usually assume $\boldsymbol{\Sigma}$ to be fixed and known, our work highlights the proof of posterior contraction of both coefficient matrix $\boldsymbol{B}$ and
covariance matrix $\boldsymbol{\Sigma}$. Moreover, we allow $p_{n}$ to grow nearly at an exponential rate with $n$ and response dimension $q_{n}$ to go to infinity. To the best of our knowledge, our work is the first paper showing the nearly-optimal contraction rate of continuous shrinkage priors under this setting. The tools we developed in proof can also be utilized in other multivariate Bayesian models. For application, we show that a large family of heavy-tailed priors, including Student's t prior, horseshoe and horseshoe+ prior, the generalized double Pareto prior, etc, satisfy the condition with good posterior contraction results.

Although we have established an informative $\ell_{2}$ reconstruction rate, there are still many important issues unexplored. One of them is the sparsity of $\boldsymbol{\Sigma}$ or its inverse. In our paper, where no structure of $\boldsymbol{\Sigma}$ is assumed, although dimension of response $q_{n}$ is allowed to grow, it has to be much smaller than sample size $n$ in order to keep $\boldsymbol{\Sigma}$ consistently estimable 19. Recently, to encourage sparsity of precision matrix, Li et al. [14] proposed a model putting horseshoe prior on regression coefficient and graphical horseshoe prior on precision matrix.

Another interesting problem is whether to adopt the joint scale-invariant prior framework. We use a scale-invariant prior in the paper, but this may result in underestimating the model error [18. Moran et al. [18] recommend independent priors for regression coefficient and error variance apriori for preventing distortion of the global-local shrinkage mechanism and obtaining better estimates of the error variance.

## Acknowledgments

The authors are grateful to Ray Bai and Qian Qin for helpful comments and suggestions.

## References

[1] Armagan, A., Clyde, M., Dunson, D. B., 2011. Generalized beta mixtures of gaussians. Advances in neural information processing systems, 523-531.
[2] Armagan, A., Dunson, D. B., Lee, J., 2013. Generalized double pareto shrinkage. Statistica Sinica 23 (1), 119.
[3] Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., Strawn, N., 2013. Posterior consistency in linear models under shrinkage priors. Biometrika 100 (4), 1011-1018.
[4] Bai, R., Ghosh, M., 2018. High-dimensional multivariate posterior consistency under global-local shrinkage priors. Journal of Multivariate Analysis 167, 157-170.
[5] Bhadra, A., Datta, J., Polson, N. G., Willard, B., et al., 2017. The horseshoe+ estimator of ultra-sparse signals. Bayesian Analysis 12 (4), 11051131.
[6] Bhattacharya, A., Chakraborty, A., Mallick, B. K., 2016. Fast sampling with gaussian scale mixture priors in high-dimensional regression. Biometrika, asw042.
[7] Carvalho, C. M., Polson, N. G., Scott, J. G., 2010. The horseshoe estimator for sparse signals. Biometrika 97 (2), 465-480.
[8] Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al., 2015. Bayesian linear regression with sparse priors. The Annals of Statistics 43 (5), 19862018.
[9] Chen, L., Huang, J. Z., 2012. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. Journal of the American Statistical Association 107 (500), 1533-1545.
[10] Chun, H., Keles, S., 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (1), 3-25.
[11] Deshpande, S. K., Rokov, V., George, E. I., 2019. Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. Journal of Computational and Graphical Statistics(to appear).
[12] Goh, G., Dey, D. K., Chen, K., 2017. Bayesian sparse reduced rank multivariate regression. Journal of multivariate analysis 157, 14-28.
[13] Griffin, J. E., Brown, P. J., et al., 2010. Inference with normal-gamma prior distributions in regression problems. Bayesian Analysis 5 (1), 171-188.
[14] Li, Y., Datta, J., Craig, B. A., Bhadra, A., 2019. Joint mean-covariance estimation via the horseshoe with an application in genomic data analysis. arXiv preprint arXiv:1903.06768.
[15] Li, Y., Nan, B., Zhu, J., 2015. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. Biometrics 71 (2), 354-363.
[16] Liquet, B., Mengersen, K., Pettitt, A., Sutton, M., et al., 2017. Bayesian variable selection regression of multivariate responses for group data. Bayesian Analysis 12 (4), 1039-1067.
[17] Mitchell, T. J., Beauchamp, J. J., 1988. Bayesian variable selection in linear regression. Journal of the American Statistical Association 83 (404), 10231032.
[18] Moran, G. E., Rocková, V., George, E. I., 2019. On variance estimation for bayesian variable selection. Bayesian Analysis(to appear).
[19] Ning, B., Ghosal, S., 2018. Bayesian linear regression for multivariate responses under group sparsity. arXiv preprint arXiv:1807.03439.
[20] Piironen, J., Vehtari, A., 20-22 Apr 2017. On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics 54, 905-913.
[21] Polson, N. G., Scott, J. G., 2010. Shrink globally, act locally: Sparse bayesian regularization and prediction. Bayesian statistics 9, 501-538.
[22] Polson, N. G., Scott, J. G., et al., 2012. On the half-cauchy prior for a global scale parameter. Bayesian Analysis 7 (4), 887-902.
[23] Raskutti, G., Wainwright, M. J., Yu, B., 2011. Minimax rates of estimation for high-dimensional linear regression over $\ell_{q}$-balls. IEEE transactions on information theory 57 (10), 6976-6994.
[24] Ročková, V., George, E. I., 2018. The spike-and-slab lasso. Journal of the American Statistical Association 113 (521), 431-444.
[25] Ročková, V., et al., 2018. Bayesian estimation of sparse signals with a continuous spike-and-slab prior. The Annals of Statistics 46 (1), 401-437.
[26] Rothman, A. J., Levina, E., Zhu, J., 2010. Sparse multivariate regression with covariance estimation. Journal of Computational and Graphical Statistics 19 (4), 947-962.
[27] Scott, D. W., 2015. Multivariate density estimation: theory, practice, and visualization. John Wiley \& Sons.
[28] Song, Q., Liang, F., 2017. Nearly optimal bayesian shrinkage for high dimensional regression. arXiv preprint arXiv:1712.08964.
[29] Stephens, M., Balding, D. J., 2009. Bayesian statistical methods for genetic association studies. Nature Reviews Genetics 10 (10), 681.
[30] Tang, X., Xu, X., Ghosh, M., Ghosh, P., 2018. Bayesian variable selection and estimation based on global-local shrinkage priors. Sankhya A 80 (2), 215-246.
[31] Van Der Pas, S., Kleijn, B., Van Der Vaart, A., et al., 2014. The horseshoe estimator: Posterior concentration around nearly black vectors. Electronic Journal of Statistics 8 (2), 2585-2618.
[32] Vershynin, R., 2012. Introduction to the non-asymptotic analysis of random matrices, 2011. arXiv preprint arXiv:1011.3027.
[33] Wilms, I., Croux, C., 2018. An algorithm for the multivariate group lasso with covariance estimation. Journal of Applied Statistics 45 (4), 668-681.
[34] Xu, X., Ghosh, M., 2015. Bayesian variable selection and estimation for group lasso. Bayesian Analysis 10 (4), 909-936.
[35] Ye, F., Zhang, C.-H., 2010. Rate minimaxity of the lasso and dantzig selector for the $\ell_{q}$ loss in $\ell_{r}$ balls. Journal of Machine Learning Research 11 (Dec), 3519-3540.

## Appendix

Proof. (Theorem 11) Define auxiliary sets $A_{n}, B_{n}$ and $C_{n}$ are constructed as follow.

$$
\begin{aligned}
A_{n} & =\left\{\text { at least } \tilde{p} \text { entries }\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\| \text { is larger than } a_{n}\right\} \\
& \cup\left\{\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\| \geq M\left(\|\boldsymbol{\Sigma}\| \vee\left\|\boldsymbol{\Sigma}_{0}\right\|\right) \epsilon_{n}\right\} \\
& \cup\left\{\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \geq M \epsilon_{n}\right\}
\end{aligned}
$$

$B_{n}=\left\{\right.$ at least $\tilde{p}$ entries $\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|$ is larger than $\left.a_{n}\right\}$, and $C_{n}=A_{n} \backslash B_{n}$. Let $\theta=(\boldsymbol{B}, \boldsymbol{\Sigma})$ and $\theta_{0}=\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)$. By Lemma A. 4 in Song and Liang [28], it suffices to show the following three parts:

$$
\begin{equation*}
\pi\left(B_{n}\right) \leq e^{-k_{1} n \epsilon_{n}^{2}} \tag{10}
\end{equation*}
$$

There exists a test function $\phi_{n}$ s.t.

$$
\begin{align*}
\mathbb{E}_{\theta_{0}} \phi_{n} & \leq e^{-k_{2} n \epsilon_{n}^{2}}  \tag{11}\\
\sup _{\theta \in C_{n}} \mathbb{E}_{\theta}\left(1-\phi_{n}\right) & \leq e^{-k_{3} n \epsilon_{n}^{2}} \tag{12}
\end{align*}
$$

And for sufficiently large $n$,

$$
\begin{equation*}
P_{\theta_{0}}\left(\frac{m\left(\mathcal{D}_{n}\right)}{f_{\theta_{0}}\left(\mathcal{D}_{n}\right)} \geq e^{-k_{4} n \epsilon_{n}^{2}}\right) \geq 1-e^{-k_{5} n \epsilon_{n}^{2}} \tag{13}
\end{equation*}
$$

for some constant $0<k_{4}<\min \left(k_{1}, k_{3}\right)$, where $m\left(\mathcal{D}_{n}\right)=\int_{\Theta} \pi(\theta) f_{\theta}\left(\mathcal{D}_{n}\right) d \theta$ is the marginal of $\mathcal{D}_{n}$.

So the proof is composed of three parts: (I)construction of test $\phi_{n}$ satisfying (11) and 12 , (II) showing that event $B_{n}$ has very probability under the specified prior, (III) showing that the marginal probability of data is highly likely to be bounded away from 0 if data is generated with true parameters.

Part I: Firstly, we show (11) and 12 by constructing testing function $\phi_{n}$ in the following way.
For given $S \subseteq\{1, \cdots, p\}$, consider the following testing functions $\phi_{n, S}^{(1)}$ and $\phi_{n, S}^{(2)}$ :

$$
\begin{aligned}
& \phi_{n, S}^{(1)}=1\left\{\left\|\frac{1}{n-|S|} \boldsymbol{Y}^{T}\left(I_{n}-H_{S}\right) \boldsymbol{Y}-\boldsymbol{\Sigma}_{0}\right\| \geq M\|\boldsymbol{\Sigma}\|_{0} \epsilon_{n} / 2\right\} \\
& \phi_{n, S}^{(2)}=1\left\{\left\|\left(\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \boldsymbol{X}_{S}^{T} \boldsymbol{Y}-\boldsymbol{B}_{\mathbf{0}_{S}}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \geq M \epsilon_{n} / 2\right\}
\end{aligned}
$$

where $\boldsymbol{X}_{S}$ is the submatrix of $\boldsymbol{X}$ composed of columns indexed by $S, \boldsymbol{B}_{S}$ and $\boldsymbol{B}_{\mathbf{0} S}$ are the submatrices of $\boldsymbol{B}$ and $\boldsymbol{B}_{\mathbf{0}}$ composed of rows indexed by $S$ respectively, $H_{S}=\boldsymbol{X}_{S}\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \boldsymbol{X}_{S}^{T}$.

We have the following two inequalities for $\mathbb{E}_{\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)} \phi_{n, S}^{(1)}$ and $\mathbb{E}_{\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)} \phi_{n, S}^{(2)}$.

$$
\begin{gathered}
\mathbb{E}_{\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)} \phi_{n, S}^{(1)}=P_{\left(\boldsymbol{B}_{0}, \boldsymbol{\Sigma}_{0}\right)}\left(\left\|\frac{1}{n-|S|} \boldsymbol{Y}^{T}\left(I_{n}-H_{S}\right) \boldsymbol{Y}-\boldsymbol{\Sigma}_{0}\right\| \geq M\|\boldsymbol{\Sigma}\|_{0} \epsilon_{n} / 2\right) \\
\leq e^{-c M^{2} n \epsilon_{n}^{2} /\left(4 K^{2}\right)} \quad \text { by Lemma } \\
\mathbb{E}_{\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)} \phi_{n, S}^{(2)}
\end{gathered} \quad \begin{aligned}
& \leq P\left(\left\|\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \boldsymbol{X}_{S}^{T} \boldsymbol{\varepsilon}\right\|_{F} \geq M \epsilon_{n} / 2\right) \\
& \leq P\left(\lambda_{\max }\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \operatorname{tr}\left(\varepsilon^{T} H_{S} \boldsymbol{\varepsilon}\right) \geq M^{2} \epsilon_{n}^{2} / 4\right) \\
& \leq P\left(\chi_{q_{n}|S|}^{2} \geq M^{2} \lambda_{0} n \epsilon_{n}^{2} / 4\right) \\
& \leq e^{-\lambda_{0} M^{2} n \epsilon_{n}^{2} / 16}
\end{aligned}
$$

The last inequality holds because $q_{n} \lesssim \log p_{n},|S| \simeq s_{0}$ and $P\left(\chi_{p}^{2} \geq x\right) \leq e^{-x / 4}$ if $x \geq 8 p[3]$.

Let $\phi_{n}=\max \left\{\phi_{n}^{(1)}, \phi_{n}^{(2)}\right\}$, where

$$
\phi_{n}^{(i)}=\max _{\left\{S \supset S_{0},|S| \leq \tilde{p}+s_{0}\right\}} \phi_{n, S}^{(i)}, \quad i=1,2 .
$$

$$
\begin{aligned}
\mathbb{E}_{\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)} \phi_{n} & \leq \mathbb{E}_{\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)} \sum_{\left\{S \supset S_{0},|S| \leq \tilde{p}+s_{0}\right\}}\left(\phi_{n, S}^{(1)}+\phi_{n, S}^{(2)}\right) \\
& \leq \sum_{i=0}^{\tilde{p}}\binom{p_{n}-s_{0}}{i} 2 e^{-k_{23} M n \epsilon_{n}^{2}} \quad \text { where } k_{23}=\min \left(\frac{c M}{4 K^{2}}, \frac{\lambda_{0} M}{16}\right) \\
& \leq 2(\tilde{p}+1)\left(p_{n}-s_{0}\right)^{\tilde{p}} e^{-k_{23} M n \epsilon_{n}^{2}}
\end{aligned}
$$

Taking logarithm on both sides,

$$
\begin{aligned}
\log \left(\mathbb{E}_{\left(\boldsymbol{B}_{0}, \boldsymbol{\Sigma}_{0}\right)} \phi_{n}\right) & \leq \tilde{p} \log \left(p_{n}-s_{0}\right)+\log (2(\tilde{p}+1))-k_{23} M n \epsilon_{n}^{2} \\
& \leq(\tilde{p}+1) \log p_{n}-k_{23} M n \epsilon_{n}^{2} \\
& \leq-k_{23} M n \epsilon_{n}^{2} / 2 \quad \text { where } \tilde{p}=\left\lfloor\frac{k_{23} M n \epsilon_{n}^{2}}{2 \log p_{n}}\right\rfloor-1
\end{aligned}
$$

Note that sufficient large $M\left(>6 / k_{23}\right)$ will ensure there is such $\tilde{p}(\geq 2)$ such that $\mathbb{E}_{\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right)} \phi_{n} \leq e^{-k_{2} n \epsilon_{n}^{2}}$.

Now we want to show $\sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n}} \mathbb{E}_{(\boldsymbol{B}, \boldsymbol{\Sigma})}\left(1-\phi_{n}\right) \leq e^{-k_{3} n \epsilon_{n}^{2}}$. Consider the following two sets $C_{n_{1}}$ and $C_{n, 2}$ :
$C_{n, 1}=\left\{\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\| \geq M\left(\|\boldsymbol{\Sigma}\| \vee\|\boldsymbol{\Sigma}\|_{0}\right) \epsilon_{n}\right\} \cap\left\{\right.$ at most $\tilde{p}$ entries $\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|$ is larger than $\left.a_{n}\right\}$, $C_{n, 2}=\left\{\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \geq M \epsilon_{n},\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\| \leq M\left(\|\boldsymbol{\Sigma}\| \vee\|\boldsymbol{\Sigma}\|_{0}\right) \epsilon_{n}\right\} \cap$
$\left\{\right.$ at most $\tilde{p}$ entries $\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|$ is larger than $\left.a_{n}\right\}$.
It's easy to verify that $C_{n} \subset C_{n, 1} \cup C_{n, 2}$, so we have

$$
\begin{aligned}
\sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n}} \mathbb{E}_{(\boldsymbol{B}, \boldsymbol{\Sigma})}\left(1-\phi_{n}\right) & =\sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n}} \mathbb{E}_{(\boldsymbol{B}, \boldsymbol{\Sigma})} \min \left(1-\phi_{n}^{(1)}, 1-\phi_{n}^{(2)}\right) \\
& \leq \max \left\{\sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 1}} \mathbb{E}_{(\boldsymbol{B}, \boldsymbol{\Sigma})}\left(1-\phi_{n}^{(1)}\right), \sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 2}} \mathbb{E}_{(\boldsymbol{B}, \boldsymbol{\Sigma})}\left(1-\phi_{n}^{(2)}\right)\right\}
\end{aligned}
$$

By definition of $\phi_{n}^{(1)}$, for $\forall S \supset S_{0},|S| \leq \tilde{p}+s_{0}$, we have

$$
\sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 1}} \mathbb{E}_{(\boldsymbol{B}, \boldsymbol{\Sigma})}\left(1-\phi_{n}^{(1)}\right) \leq \sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 1}} \mathbb{E}_{(\boldsymbol{B}, \boldsymbol{\Sigma})}\left(1-\phi_{n, S}^{(1)}\right)
$$

Taking $S=\left\{j:\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} \geq a_{n}\right\} \cup S_{0}$, note that when $(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 1}$,

$$
\begin{aligned}
&\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\|-M\|\boldsymbol{\Sigma}\|_{0} \epsilon_{n} / 2 \geq M\|\boldsymbol{\Sigma}\| \epsilon_{n} / 2, \text { then } \\
& \sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 1}} \mathbb{E}_{(\boldsymbol{B}, \boldsymbol{\Sigma})}\left(1-\phi_{n, S}^{(1)}\right) \\
&= \sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 1}} P\left(\left\|\frac{1}{n-|S|} \boldsymbol{Y}^{T}\left(I_{n}-H_{S}\right) \boldsymbol{Y}-\boldsymbol{\Sigma}_{0}\right\| \leq M\|\boldsymbol{\Sigma}\|_{0} \epsilon_{n} / 2\right) \\
& \leq \sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 1}} P\left(\left\|\frac{1}{n-|S|} \boldsymbol{Y}^{T}\left(I_{n}-H_{S}\right) \boldsymbol{Y}-\boldsymbol{\Sigma}\right\| \geq\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\|-M\|\boldsymbol{\Sigma}\|_{0} \epsilon_{n} / 2\right) \\
& \leq \sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 1}} P\left(\left\|\frac{1}{n-|S|} \boldsymbol{Y}^{T}\left(I_{n}-H_{S}\right) \boldsymbol{Y}-\boldsymbol{\Sigma}\right\| \geq M\|\boldsymbol{\Sigma}\| \epsilon_{n} / 2\right) \\
& \leq e^{-c M^{2} n \epsilon_{n}^{2} /\left(16 K^{2}\right)}
\end{aligned}
$$

The last inequality follows by Lemma (1) because $\boldsymbol{Y}=\boldsymbol{X} \boldsymbol{B}+\boldsymbol{\varepsilon} \boldsymbol{\Sigma}^{1 / 2}$ and for $(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 1}$,

$$
\begin{aligned}
\left\|\left(I_{n}-H_{S}\right) X \boldsymbol{B} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2} & =\left\|X_{S^{c}} \boldsymbol{B}_{S^{c}} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2} \\
& \leq\left\|X_{S^{c}}\right\|_{F}^{2}\left\|\boldsymbol{B}_{S^{c}} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2} \\
& \leq\left(n p_{n}\right)\left(p_{n} a_{n}^{2}\right) \\
& \lesssim n \epsilon_{n}^{2}
\end{aligned}
$$

Now we consider $\sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 2}} \mathbb{E}_{(\boldsymbol{B}, \boldsymbol{\Sigma})}\left(1-\phi_{n, S}^{(2)}\right)$. For $(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 2},\|\boldsymbol{\Sigma}\|-$ $\left\|\boldsymbol{\Sigma}_{0}\right\| \leq M\|\boldsymbol{\Sigma}\| \epsilon_{n}$ or $\|\boldsymbol{\Sigma}\|-\left\|\boldsymbol{\Sigma}_{0}\right\| \leq M\left\|\boldsymbol{\Sigma}_{0}\right\| \epsilon_{n}$, so $\|\boldsymbol{\Sigma}\| \leq 2\left\|\boldsymbol{\Sigma}_{0}\right\|$, which gives $\left\|\boldsymbol{\Sigma} \boldsymbol{\Sigma}_{0}^{-1}\right\| \leq 2 b_{2} / b_{1}$ by assumption $A_{3}(3)$. So we have

$$
\left\|\left(\boldsymbol{B}_{S}-\boldsymbol{B}_{\mathbf{0} S}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \geq\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F}-\left\|\boldsymbol{B}_{S^{c}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \geq 7 M \epsilon_{n} / 8
$$

since $\left\|\boldsymbol{B}_{S^{c}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \leq\left\|\boldsymbol{B}_{S^{c}} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}\left\|\boldsymbol{\Sigma} \boldsymbol{\Sigma}_{0}^{-1}\right\|^{1 / 2},\left\|\boldsymbol{B}_{S^{c}} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F} \leq a_{n} p_{n} \prec \epsilon_{n}$ and $\left\|\boldsymbol{\Sigma} \boldsymbol{\Sigma}_{0}^{-1}\right\|^{1 / 2}$ is bounded. In addition, we have

$$
\begin{aligned}
\left\|\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \boldsymbol{X}_{S}^{T} X_{S^{c}} \boldsymbol{B}_{S^{c}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} & \leq \sqrt{\lambda_{\max }\left(\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1}\right)}\left\|X_{S^{c}} \boldsymbol{B}_{S^{c}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \\
& \leq \sqrt{1 /\left(n \lambda_{0}\right)}\left\|\boldsymbol{\Sigma} \boldsymbol{\Sigma}_{0}^{-1}\right\|^{1 / 2}\left\|X_{S^{c}} \boldsymbol{B}_{S^{c}} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F} \\
& \leq M \epsilon_{n} / 8 \text { for } M \geq 8 \sqrt{b_{2} /\left(b_{1} \lambda_{0}\right)}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 2}} \mathbb{E}_{(\boldsymbol{B}, \boldsymbol{\Sigma})}\left(1-\phi_{n, S}^{(2)}\right) \\
= & \sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 2}} P\left(\left\|\left(\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \boldsymbol{X}_{S}^{T} \boldsymbol{Y}-\boldsymbol{B}_{\mathbf{0} S}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \leq M \epsilon_{n} / 2\right) \\
& \sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 2}} P\left(\|\left(\boldsymbol{B}_{S}-\boldsymbol{B}_{\mathbf{0} S}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}+\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \boldsymbol{X}_{S}^{T} X_{S^{c}} \boldsymbol{B}_{S^{c}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right. \\
& \left.+\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \boldsymbol{X}_{S}^{T} E \boldsymbol{\Sigma}^{1 / 2} \boldsymbol{\Sigma}_{0}^{-1 / 2} \|_{F} \leq M \epsilon_{n} / 2\right) \\
& \sup \quad P\left(\left\|\left(\boldsymbol{B}_{S}-\boldsymbol{B}_{\mathbf{0} S}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F}-\left\|\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \boldsymbol{X}_{S}^{T} X_{S^{c}} \boldsymbol{B}_{S^{c}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F}\right. \\
\leq & \left.-\left\|\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \boldsymbol{X}_{S}^{T} E \boldsymbol{\Sigma}^{1 / 2} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \leq M \epsilon_{n} / 2\right) \\
\leq & \sup _{(\boldsymbol{B}, \boldsymbol{\Sigma}) \in C_{n, 2}} P\left(\left\|\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \boldsymbol{X}_{S}^{T} E \boldsymbol{\Sigma}^{1 / 2} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \geq M \epsilon_{n} / 4\right) \\
\leq & P\left(\chi_{q_{n}|S|}^{2} \geq \lambda_{0} b_{1} M^{2} n \epsilon_{n}^{2} /\left(32 b_{2}\right)\right) \\
\leq & e^{-\lambda_{0} b_{1} M^{2} n \epsilon_{n}^{2} /\left(128 b_{2}\right)}
\end{aligned}
$$

The second last inequality holds since

$$
\begin{aligned}
\left\|\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \boldsymbol{X}_{S}^{T} E \boldsymbol{\Sigma}^{1 / 2} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F}^{2} & \leq\left\|\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1} \boldsymbol{X}_{S}^{T} E\right\|_{F}^{2}\left\|\boldsymbol{\Sigma} \boldsymbol{\Sigma}_{0}^{-1}\right\| \\
& \leq 2\left(b_{2} / b_{1}\right) \lambda_{\max }\left(\left(\boldsymbol{X}_{S}^{T} \boldsymbol{X}_{S}\right)^{-1}\right)\left\|H_{S} E\right\|_{F}^{2} \\
& \leq \frac{2 b_{2}}{b_{1} \lambda_{0} n} \chi_{q_{n}|S|}^{2}
\end{aligned}
$$

Let $k_{3}=\min \left(\frac{c M^{2}}{16 K^{2}}, \frac{\lambda_{0} b_{1} M^{2}}{128 b_{2}}\right), 12$ is proved.

Part II: Now we will show $\pi\left(B_{n}\right) \leq e^{-k_{1} n \epsilon_{n}^{2}}$.
Define $N=\left|\left\{j:\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} \geq a_{n}\right\}\right|$, following the proof in Part II of Theorem A. 1 in Song and Liang [28], we have $\pi\left(B_{n}\right) \leq e^{-t_{n}} /\left(2 \sqrt{\pi t_{n}}\right)$, where

$$
\begin{aligned}
t_{n} & =(\tilde{p}-1) \log (\tilde{p}-1)+(\tilde{p}-1) \log \frac{1}{p_{n} v_{n}}+\left(p_{n}-\tilde{p}+1\right) \log \frac{p_{n}-\tilde{p}+1}{p_{n}-p_{n} v_{n}} \\
& \left.\geq \frac{u}{2} \tilde{p} \log p_{n}-\left(p_{n}-\tilde{p}+1\right) \log \left(1+\frac{\tilde{p}-1-p_{n} v_{n}}{p_{n}-\tilde{p}+1}\right) \quad \text { (recall that } p_{n} v_{n} \leq p_{n}^{-u}\right) \\
& \geq \frac{u}{2} \tilde{p} \log p_{n}-\left(\tilde{p}-1-p_{n} v_{n}\right) \\
& \geq \frac{u}{4} \tilde{p} \log p_{n} \\
& \geq u k_{23} M n \epsilon_{n}^{2} / 16
\end{aligned}
$$

Thus $\pi\left(B_{n}\right) \leq e^{-k_{1} n \epsilon_{n}^{2}}$ is proved with $k_{1}=u k_{23} M / 16$.

Part III: At last, we will show (13).
By Part III in Song and Liang [28], it suffices to show that

$$
\begin{align*}
& P_{\theta_{0}}\left(\pi\left(\left\|(\boldsymbol{Y}-X \boldsymbol{B}) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2}-\left\|\left(\boldsymbol{Y}-X \boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F}^{2}+n \log \left(\frac{|\boldsymbol{\Sigma}|}{\left|\boldsymbol{\Sigma}_{0}\right|}\right) \leq k_{4} n \epsilon_{n}^{2} / 2\right)\right. \\
& \left.\quad \geq e^{-k_{4} n \epsilon_{n}^{2} / 2}\right) \geq 1-e^{-k_{5} n \epsilon_{n}^{2}} \tag{14}
\end{align*}
$$

Proof of (14) has three steps. The first step is to show event $\Omega:=\left\{\|\varepsilon\|_{F}^{2} \leq\right.$ $n q_{n}\left(1+c_{1}\right)$ and $\left.\left\|\boldsymbol{X}^{T} \varepsilon\right\|_{2, \infty} \leq c_{2} n \epsilon_{n}\right\}$ has large probability closing to 1 . In second step, we show that on event $\Omega$, and when the data $\boldsymbol{Y}$ is generated under true parameter $\left(\boldsymbol{B}_{\mathbf{0}}, \boldsymbol{\Sigma}_{0}\right),\left\{\left\|(\boldsymbol{Y}-X \boldsymbol{B}) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2}-\left\|\left(\boldsymbol{Y}-X \boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F}^{2}+\right.$ $\left.\left.n \log \left(|\boldsymbol{\Sigma}| /\left|\boldsymbol{\Sigma}_{0}\right|\right) \leq k_{4} n \epsilon_{n}^{2} / 2\right)\right\}$ is a super-set of $\left\{\cap_{i=1}^{d}\left\{\tilde{\boldsymbol{\Sigma}}: 1 \leq \lambda_{i}(\tilde{\boldsymbol{\Sigma}}) \leq 1+\right.\right.$ $\left.\epsilon_{n}^{2} / q_{n}\right\}$ and $\left.\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2,1} \leq \eta \epsilon_{n}\right\}$, where $\tilde{\boldsymbol{\Sigma}}:=\boldsymbol{\Sigma}_{0}^{-1 / 2} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_{0}^{-1 / 2}$, and $\eta$ is a constant satisfies $\eta \epsilon_{n} / p_{n} \geq a_{n}$. The last step is to get the lower bound of $\pi\left(\left\{\cap_{i=1}^{d}\left\{\tilde{\boldsymbol{\Sigma}}: 1 \leq \lambda_{i}(\tilde{\boldsymbol{\Sigma}}) \leq 1+\epsilon_{n}^{2} / q_{n}\right\}\right.\right.$ and $\left.\left.\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2,1} \leq \eta \epsilon_{n}\right\}\right)$.

Step 1: $P\left(\left\{\|\varepsilon\|_{F}^{2} \leq n q_{n}\left(1+c_{1}\right)\right.\right.$ and $\left.\left.\left\|\boldsymbol{X}^{T} \varepsilon\right\|_{2, \infty} \leq c_{2} n \epsilon_{n}\right\}\right) \geq 1-e^{-n \epsilon_{n}^{2}}$
It's easy to see that $P\left(\|\varepsilon\|_{F}^{2}>n q_{n}\left(1+c_{1}\right)\right) \leq e^{-c_{1}^{2} n q_{n} / 4} \leq e^{-c_{1}^{2} n \epsilon_{n}^{2} / 4}$ because $\|\varepsilon\|_{F}^{2} \sim \chi_{n q_{n}}^{2}$. And $\left(\boldsymbol{X}^{T} \varepsilon\right)_{i j}^{2} / n \leq\left(\boldsymbol{X}^{T} \varepsilon\right)_{i j}^{2} /\left(\boldsymbol{X}^{T} \boldsymbol{X}\right)_{i i}$ by the fact that $\boldsymbol{X}$ is uniformly bounded by 1 . And note that for fixed $i,\left(\boldsymbol{X}^{T} \boldsymbol{\varepsilon}\right)_{i j} / \sqrt{\left(\boldsymbol{X}^{T} \boldsymbol{X}\right)_{i i}} \sim$ $\mathcal{N}(0,1)$ independently, $j=1, \cdots, q_{n}$. Thus we have

$$
\begin{aligned}
P\left(\left\|\boldsymbol{X}^{T} \varepsilon\right\|_{2, \infty}>c_{2} n \epsilon_{n}\right) & =P\left(\max _{1 \leq i \leq n} \frac{1}{n} \sum_{j=1}^{q_{n}}\left(\boldsymbol{X}^{T} \varepsilon\right)_{i j}^{2}>c_{2}^{2} n \epsilon_{n}^{2}\right) \\
& \leq P\left(\max _{1 \leq i \leq n} Z_{i}>c_{2}^{2} n \epsilon_{n}^{2}\right) \quad \text { where } Z_{i} \sim \chi_{q_{n}}^{2} \\
& \leq p_{n} P\left(\chi_{q_{n}}^{2}>c_{2}^{2} n \epsilon_{n}^{2}\right) \\
& \leq p_{n} \exp \left\{-c_{2}^{2} n \epsilon_{n}^{2} / 4\right\} \\
& \leq \exp \left\{-\left(c_{2}^{2} / 4-1 / s_{0}\right) n \epsilon_{n}^{2}\right\} \\
& \leq e^{-3 n \epsilon_{n}^{2}}
\end{aligned}
$$

$$
P(\Omega) \geq 1-P\left(\|\varepsilon\|_{F}^{2}>n q_{n}\left(1+c_{1}\right)\right)-P\left(\left\|X^{T} E\right\|_{2, \infty}>c_{2} n \epsilon_{n}\right) \geq 1-e^{-n \epsilon_{n}^{2}} .
$$

Step 2: On event $\Omega$ and when the data $\boldsymbol{Y}$ is generated under true parameter $\left(\boldsymbol{B}_{0}, \boldsymbol{\Sigma}_{0}\right)$, if $(\boldsymbol{B}, \boldsymbol{\Sigma}) \in\left\{\cap_{i=1}^{d}\left\{\tilde{\boldsymbol{\Sigma}}: 1 \leq \lambda_{i}(\tilde{\boldsymbol{\Sigma}}) \leq 1+\epsilon_{n}^{2} / q_{n}\right\}\right.$ and $\|(\boldsymbol{B}-$ $\left.\left.\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1 / 2} \|_{2,1} \leq \eta \epsilon_{n}\right\}$, then the following four inequalities hold.

$$
\begin{align*}
n \log \left(|\boldsymbol{\Sigma}| /\left|\boldsymbol{\Sigma}_{0}\right|\right) & \leq n \epsilon_{n}^{2}  \tag{15}\\
\left|\left\|\boldsymbol{\varepsilon} \boldsymbol{\Sigma}_{0}^{1 / 2} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2}-\|\boldsymbol{\varepsilon}\|_{F}^{2}\right| & \leq\left(1+c_{1}\right) n \epsilon_{n}^{2}  \tag{16}\\
\left\|\boldsymbol{X}\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2} & \leq \eta^{2} n \epsilon_{n}^{2}  \tag{17}\\
\left|\operatorname{tr}\left(\boldsymbol{X}\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{0}^{1 / 2} E\right)\right| & \leq c_{2} \eta n \epsilon_{n}^{2} \tag{18}
\end{align*}
$$

(15) holds because

$$
n \log \left(|\boldsymbol{\Sigma}| /\left|\boldsymbol{\Sigma}_{0}\right|\right)=n \sum_{i=1}^{q_{n}} \log \lambda_{i}(\tilde{\boldsymbol{\Sigma}}) \leq n \sum_{i=1}^{q_{n}} \log \left(1+\epsilon_{n}^{2} / q_{n}\right) \leq n \epsilon_{n}^{2} .
$$

For (16), it's easy to see

$$
\|\varepsilon\|_{F}^{2} \lambda_{1}\left(\tilde{\boldsymbol{\Sigma}}^{-1}-I_{q_{n}}\right) \leq \operatorname{tr}\left(\varepsilon^{T} \varepsilon\left(\tilde{\boldsymbol{\Sigma}}^{-1}-I_{q_{n}}\right)\right) \leq\|\varepsilon\|_{F}^{2} \lambda_{q_{n}}\left(\tilde{\boldsymbol{\Sigma}}^{-1}-I_{q_{n}}\right) .
$$

And $1-\epsilon_{n}^{2} \leq \lambda_{1}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\right) \leq \lambda_{q_{n}}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\right) \leq 1$ because $1 \leq \lambda_{i}(\tilde{\boldsymbol{\Sigma}}) \leq 1+\epsilon_{n}^{2} / q_{n} \Rightarrow$ $1-\epsilon_{n}^{2} / q_{n} \leq \lambda_{q_{n}-i+1}(\tilde{\boldsymbol{\Sigma}}) \leq 1$. So we have

$$
\left|\left\|\varepsilon \boldsymbol{\Sigma}_{0}^{1 / 2} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2}-\|\varepsilon\|_{F}^{2}\right|=\left|\operatorname{tr}\left(\varepsilon^{T} \boldsymbol{\varepsilon}\left(\tilde{\boldsymbol{\Sigma}}^{-1}-I_{q_{n}}\right)\right)\right| \leq\|\varepsilon\|_{F}^{2} \epsilon_{n}^{2} / q_{n} \leq\left(1+c_{1}\right) n \epsilon_{n}^{2} .
$$

For (17), let $\boldsymbol{X}_{i}$ denote the $i$ th row of $\boldsymbol{X}$, we have

$$
\begin{aligned}
\left\|\boldsymbol{X}_{i}\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} & =\left\|\sum_{j=1}^{p_{n}} x_{i j}\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right)_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} \\
& \leq \sum_{j=1}^{p_{n}}\left|x_{i j}\right|\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right)_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} \\
& \leq\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2,1}
\end{aligned}
$$

So $\left\|\boldsymbol{X}\left(\boldsymbol{B}-\boldsymbol{B}_{0}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2}=\sum_{i=1}^{n}\left\|\boldsymbol{X}_{i}\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2}{ }^{2} \leq \eta^{2} n \epsilon_{n}^{2}$.
(18) is immediately by Lemma (2)

$$
\begin{aligned}
\left|\operatorname{tr}\left(\boldsymbol{X}\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{0}^{1 / 2} \boldsymbol{\varepsilon}\right)\right| & =\left|\operatorname{tr}\left(\boldsymbol{\varepsilon}^{T} \boldsymbol{X}\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{0}^{1 / 2}\right)\right| \\
& \leq\left\|\boldsymbol{X}^{T} \boldsymbol{\varepsilon}\right\|_{2, \infty}\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{0}^{1 / 2}\right\|_{2,1} \\
& \leq\left\|\boldsymbol{X}^{T} \boldsymbol{\varepsilon}\right\|_{2, \infty} \lambda_{q_{n}}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\right)^{1 / 2}\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2,1} \\
& \leq c_{2} \eta n \epsilon_{n}^{2}
\end{aligned}
$$

Combining (15)-18, we can get

$$
\begin{aligned}
& \left\|(\boldsymbol{Y}-\boldsymbol{X} \boldsymbol{B}) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2}-\left\|\left(\boldsymbol{Y}-\boldsymbol{X} \boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F}^{2}+n \log \left(|\boldsymbol{\Sigma}| /\left|\boldsymbol{\Sigma}_{0}\right|\right) \\
= & \left\|\boldsymbol{X}\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2}+2 \operatorname{tr}\left(X\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{0}^{1 / 2} E\right)+\left\|\boldsymbol{\varepsilon} \boldsymbol{\Sigma}_{0}^{1 / 2} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2}-\|\boldsymbol{\varepsilon}\|_{F}^{2}+n \log \left(|\boldsymbol{\Sigma}| /\left|\boldsymbol{\Sigma}_{0}\right|\right) \\
\leq & k_{4} n \epsilon_{n}^{2} \quad \text { when } k_{4} \geq 2+c_{1}+\eta^{2}+2 c_{2} \eta
\end{aligned}
$$

Step 3: We want to show
$\pi\left(\cap_{i=1}^{q_{n}}\left\{\tilde{\boldsymbol{\Sigma}}: 1 \leq \lambda_{i}(\tilde{\boldsymbol{\Sigma}}) \leq 1+\epsilon_{n}^{2} / q_{n}\right\}\right.$ and $\left.\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2,1} \leq \eta \epsilon_{n}\right) \geq e^{-k_{4} n \epsilon_{n}^{2} / 2}$.
By Lemma A. 4 in Ning and Ghosal [19],

$$
\begin{equation*}
-\log \left(\pi\left(\cap_{i=1}^{q_{n}}\left\{\tilde{\boldsymbol{\Sigma}}: 1 \leq \lambda_{i}(\tilde{\boldsymbol{\Sigma}}) \leq 1+\epsilon_{n}^{2} / q_{n}\right\}\right) \lesssim q_{n}^{2} \log q_{n}+q_{n}^{2} \log \frac{1}{\epsilon_{n}^{2}} \lesssim q_{n}^{2} \log n \lesssim n \epsilon_{n}^{2}\right. \tag{19}
\end{equation*}
$$

Now, we look at

$$
\pi\left(\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{0}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2,1} \leq \epsilon_{n} \mid \cap_{i=1}^{q_{n}}\left\{\tilde{\boldsymbol{\Sigma}}: 1 \leq \lambda_{i}(\tilde{\boldsymbol{\Sigma}}) \leq 1+\epsilon_{n}^{2} / q_{n}\right\}\right)
$$

It is easy to see that

$$
\begin{aligned}
\left\{\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2,1} \leq \eta \epsilon_{n}\right\} \supset & \left\{\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} \leq \eta \epsilon_{n} / p_{n} \text { for all } j \notin S\right\} \cap \\
& \left\{\left\|\left(\boldsymbol{B}_{j}-\boldsymbol{B}_{\mathbf{0}_{j}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} \leq \eta \epsilon_{n} / s_{0} \text { for all } j \in S_{0}\right\}
\end{aligned}
$$

Firstly,

$$
\begin{align*}
\pi\left(\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} \leq \eta \epsilon_{n} / p_{n} \text { for } j \notin S_{0}\right) & \geq \prod_{j \notin S_{0}} \pi\left(\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} \leq a_{n}\right)  \tag{20}\\
& \geq\left(1-p_{n}^{-(1+u)}\right)^{p_{n}} \rightarrow 1
\end{align*}
$$

For a $d$-dimensional vector $\boldsymbol{x}_{0}$, let $\mathcal{B}\left(\boldsymbol{x}_{0}, r\right)$ denotes a ball in $\mathbb{R}^{d}$ with center $\boldsymbol{x}_{0}$ and radius $r$. Note that conditioning on $\cap_{i=1}^{q_{n}}\left\{\tilde{\boldsymbol{\Sigma}}: 1 \leq \lambda_{i}(\tilde{\boldsymbol{\Sigma}}) \leq 1+\epsilon_{n}^{2} / q_{n}\right\}$, $\mathcal{B}\left(\boldsymbol{B}_{\mathbf{0}_{j}} \boldsymbol{\Sigma}^{-1 / 2}, \eta \epsilon_{n} / s_{0}\right) \subset \mathcal{B}\left(\mathbf{0},\left\|\boldsymbol{B}_{\mathbf{0}_{j}} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2}+\eta \epsilon_{n} / s_{0}\right) \subset \mathcal{B}\left(\mathbf{0}, \gamma\left\|\boldsymbol{B}_{\mathbf{0}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2, \infty}\right)$
because $\left\|\boldsymbol{B}_{\mathbf{0}_{j}} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} \leq\left\|\boldsymbol{B}_{\mathbf{0}_{j}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2} \sqrt{\lambda_{q_{n}}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\right)} \leq\left\|\boldsymbol{B}_{\mathbf{0}_{j}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2}$.
Therefore, we have

$$
\begin{aligned}
& \left.\pi\left(\left\|\left(\boldsymbol{B}_{j}-\boldsymbol{B}_{\mathbf{0}_{j}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} \leq \eta \epsilon_{n} / s_{0} \text { for all } j \in S_{0} \mid \cap_{i=1}^{d}\left\{\tilde{\boldsymbol{\Sigma}}: 1 \leq \lambda_{i}(\tilde{\boldsymbol{\Sigma}}) \leq 1+\epsilon_{n}^{2} / q_{n}\right\}\right)\right) \\
\geq & \left(\operatorname{Volumn}\left(\mathcal{B}\left(\boldsymbol{B}_{\mathbf{0}_{j}} \boldsymbol{\Sigma}^{-1 / 2}, \eta \epsilon_{n} / s_{0}\right)\right) \inf _{\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} \leq M_{0}} g_{\tau}\left(\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right)\right)^{s_{0}} \\
= & \left(\frac{\pi^{q_{n} / 2}}{\Gamma\left(q_{n} / 2+1\right)}\right)^{s_{0}}\left(\frac{\eta \epsilon_{n}}{s_{0}}\right)^{q_{n} s_{0}}\left(\inf _{\|\boldsymbol{x}\|_{2} \leq M_{0}} g_{\tau}(\boldsymbol{x})\right)^{s_{0}}
\end{aligned}
$$

Taking $\log$ and multiply by -1 on both sides,

$$
\begin{align*}
& \left.-\log \pi\left(\left\|\left(\boldsymbol{B}_{j}-\boldsymbol{B}_{\mathbf{0}_{j}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2} \leq \epsilon_{n} / s_{0} \text { for all } j \in S_{0} \mid \cap_{i=1}^{q_{n}}\left\{\tilde{\boldsymbol{\Sigma}}: 1 \leq \lambda_{i}(\tilde{\boldsymbol{\Sigma}}) \leq 1+\epsilon_{n}^{2} / q_{n}\right\}\right)\right) \\
\leq & -\frac{q_{n} s_{0}}{2} \log \eta^{2} \pi+s_{0} \log \Gamma\left(\frac{q_{n}}{2}+1\right)+q_{n} s_{0} \log s_{0}+\frac{q_{n} s_{0}}{2} \log \left(1 / \epsilon_{n}^{2}\right)-s_{0} \log \inf _{\|\boldsymbol{x}\|_{2} \leq M_{0}} g_{\tau}(\boldsymbol{x}) \\
\leq & q_{n} s_{0} \log \left(q_{n} s_{0}\right)+\frac{q_{n} s_{0}}{2} \log \left(1 / \epsilon_{n}^{2}\right)+s_{0}\left(-\log \inf _{\|\boldsymbol{x}\|_{2} \leq M_{0}} g_{\tau}(\boldsymbol{x})\right) \\
\lesssim & q_{n} s_{0} \log n+s_{0} \log p_{n} \lesssim n \epsilon_{n}^{2} \tag{21}
\end{align*}
$$

From (19), 20) and 21), we get
$\pi\left(\cap_{i=1}^{d}\left\{\tilde{\boldsymbol{\Sigma}}: 1 \leq \lambda_{i}(\tilde{\boldsymbol{\Sigma}}) \leq 1+\epsilon_{n}^{2} / q_{n}\right\}\right.$ and $\left.\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{\mathbf{0}}\right) \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2,1} \leq \eta \epsilon_{n}\right) \geq e^{-k_{4} n \epsilon_{n}^{2} / 2}$ for sufficiently large $k_{4}$.

Also note that by Lemma A. 4 in Song and Liang [28, $k_{4}$ needs to satisfy $k_{4}<$ $\min \left(k_{1}, k_{3}\right)$. Existence of such $k_{4}$ is ensured by setting $M$ be a sufficiently large constant so that $k_{1}$ and $k_{3}$ are large.

Proof. (Theorem 22 Let $A_{n}$ be the auxiliary set defined in Theorem 1, then

$$
\begin{aligned}
A_{n}^{c} & =\left\{\text { at most } \tilde{p} \text { entries }\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\| \text { is larger than } a_{n}\right\} \\
& \cap\left\{\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\| \leq M\left(\|\boldsymbol{\Sigma}\| \vee\left\|\boldsymbol{\Sigma}_{0}\right\|\right) \epsilon_{n}\right\} \\
& \cap\left\{\left\|\left(\boldsymbol{B}-\boldsymbol{B}_{0}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F} \leq M \epsilon_{n}\right\}
\end{aligned}
$$

From Theorem 1, we have $P_{0}\left\{\Pi\left(A_{n}^{c} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)>1-\exp \left(-c_{1} n \epsilon_{n}^{2}\right)\right\}>1-\exp \left(-c_{2} n \epsilon_{n}^{2}\right)$.
All the following analysis is conditioning on $\Omega_{n}=\Pi\left(A_{n}^{c} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)>1-\exp \left(-c_{1} n \epsilon_{n}^{2}\right)$.

The proof has three parts. Part 1 shows that only the case $S_{n} \supseteq S_{0}$ needs to be considered. Part 2 calculates $\pi\left(S=S_{0} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)$ and $\pi\left(S=S^{\prime} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)$ where $S^{\prime} \supseteq S_{0}$. Part 3 shows that $\sum_{S^{\prime} \supseteq S_{0},\left|S^{\prime} \backslash S_{0}\right| \geq 1} \frac{\pi\left(S=S^{\prime} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)}{\pi\left(S=S_{0} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)} \rightarrow 0$.

Part I: We first show that it suffices to consider $S_{n} \supseteq S_{0}$ because $\pi\left(S_{n} \nsupseteq\right.$ $\left.S_{0} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right) \rightarrow 0$. Note that

$$
\begin{aligned}
\pi\left(S_{n} \nsupseteq S_{0} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right) & =\pi\left(\exists j \text { satisfying } \boldsymbol{B}_{\mathbf{0}_{j}} \neq \boldsymbol{B}_{j},\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\| \leq a_{n} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right) \\
& \leq \pi\left(\left\|\left(\boldsymbol{B}_{j}-\boldsymbol{B}_{\mathbf{0}_{j}}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\| \geq M \epsilon_{n} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right) \rightarrow 0
\end{aligned}
$$

The inequality holds because for $j$ satisfying $\boldsymbol{B}_{\mathbf{0}_{j}} \neq \boldsymbol{B}_{j}$ and $\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\| \leq a_{n}$,

$$
\begin{aligned}
\left\|\left(\boldsymbol{B}_{j}-\boldsymbol{B}_{\mathbf{0}_{j}}\right) \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\| & \geq\left\|\boldsymbol{B}_{\mathbf{0}_{j}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|-\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\| \\
& \geq \sqrt{1 / b_{2}}\left\|\boldsymbol{B}_{\mathbf{0} j}\right\|-\sqrt{2 b_{2} / b_{1}}\left\|\boldsymbol{B}_{j} \boldsymbol{\Sigma}^{-1 / 2}\right\| \\
& \geq \sqrt{1 / b_{2}} M_{1} \epsilon_{n}-\sqrt{2 b_{2} / b_{1}} a_{n} \geq M_{2} \epsilon_{n}
\end{aligned}
$$

since $\left\|\boldsymbol{\Sigma}_{0}\right\| \leq b_{2}$ and $\left\|\boldsymbol{\Sigma} \boldsymbol{\Sigma}_{0}^{-1}\right\| \leq 2 b_{2} / b_{1}$ by previous results.

Part II: Let $E_{1}=\left\{(\boldsymbol{B}, \boldsymbol{\Sigma}):\left\|\boldsymbol{B}_{1}-\boldsymbol{B}_{01}\right\|_{F} \leq c_{1} \epsilon_{n},\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\| \leq c_{2} \epsilon_{n}\right\}$, and $\underline{\pi\left(\boldsymbol{B}_{1} \mid \boldsymbol{\Sigma}\right)}=\inf _{\left(\boldsymbol{B}_{1}, \boldsymbol{\Sigma}\right) \in E_{1}} \pi\left(\boldsymbol{B}_{1}, \boldsymbol{\Sigma}\right) / \pi(\boldsymbol{\Sigma}), \overline{\pi\left(\boldsymbol{B}_{1} \mid \boldsymbol{\Sigma}\right)}=\sup _{\left(\boldsymbol{B}_{1}, \boldsymbol{\Sigma}\right) \in E_{1}} \pi\left(\boldsymbol{B}_{1}, \boldsymbol{\Sigma}\right) / \pi(\boldsymbol{\Sigma})$.

Consider $\pi\left(S=S_{0} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)$, let subscripts "1" and "2" denote model $S_{0}$ and $S_{0}^{c}$ respectively. It is easy to check that $\min _{j}\left\|\boldsymbol{B}_{1 j} \boldsymbol{\Sigma}^{-1 / 2}\right\|>a_{n}$ for $S_{n} \supseteq S_{0}$, so we have

$$
\begin{aligned}
\pi\left(S=S_{0} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)= & \pi\left(\left\|\boldsymbol{B}_{2} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2, \infty} \leq a_{n} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right) \\
\propto & \int|\boldsymbol{\Sigma}|^{-n / 2} \exp \left\{-\frac{1}{2} \sum_{i=1}^{n}\left(Y_{i}-X_{i} \boldsymbol{B}\right) \Sigma^{-1}\left(Y_{i}-X_{i} \boldsymbol{B}\right)^{T}\right\} \pi(\boldsymbol{B}, \boldsymbol{\Sigma}) \\
& I\left(\left\|\boldsymbol{B}_{2} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}\right) d \boldsymbol{B} d \boldsymbol{\Sigma} \\
\geq & \pi\left(\left\|\boldsymbol{B}_{2} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}\right) \int_{E_{1}\left\|B_{2} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}}|\boldsymbol{\Sigma}|^{-n / 2} \times \\
& \exp \left\{-\frac{1}{2} \sum_{i=1}^{n}\left(Y_{i}-X_{i} \boldsymbol{B}\right) \Sigma^{-1}\left(Y_{i}-X_{i} \boldsymbol{B}\right)^{T}\right\} \pi\left(\boldsymbol{B}_{1}, \boldsymbol{\Sigma}\right) d \boldsymbol{B}_{1} d \boldsymbol{\Sigma}
\end{aligned}
$$

$$
\begin{aligned}
& \int_{E_{1}\left\|B_{2} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}}|\boldsymbol{\Sigma}|^{-n / 2} \exp \left\{-\frac{1}{2} \sum_{i=1}^{n}\left(Y_{i}-X_{i} \boldsymbol{B}\right) \Sigma^{-1}\left(Y_{i}-X_{i} \boldsymbol{B}\right)^{T}\right\} \pi\left(\boldsymbol{B}_{1}, \boldsymbol{\Sigma}\right) d \boldsymbol{B}_{1} d \boldsymbol{\Sigma} \\
\geq & \frac{\pi\left(\boldsymbol{B}_{1} \mid \boldsymbol{\Sigma}\right)}{} \int_{\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\| \leq c_{2} \epsilon_{n}}\left\|B_{2} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2, \infty} \leq a_{n} \\
& |\boldsymbol{\Sigma}|^{-n / 2} \pi(\boldsymbol{\Sigma}) \times \\
& \int_{\left\|\boldsymbol{B}_{1}-\boldsymbol{B}_{01}\right\|_{\boldsymbol{F}} \leq c_{1} \epsilon_{n}} \exp \left\{-\frac{1}{2} \sum_{i=1}^{n}\left(Y_{i}-X_{i 1} \boldsymbol{B}_{1}-X_{i 2} \boldsymbol{B}_{2}\right) \Sigma^{-1}\left(Y_{i}-X_{i 1} \boldsymbol{B}_{1}-X_{i 2} \boldsymbol{B}_{2}\right)^{T}\right\} d \boldsymbol{B}_{1} d \boldsymbol{\Sigma}
\end{aligned}
$$

Let

$$
\begin{aligned}
\operatorname{SSE}\left(\boldsymbol{B}_{2}, \boldsymbol{\Sigma}\right) & =\min _{\boldsymbol{B}_{1}}\left\|\left(\boldsymbol{Y}-\boldsymbol{X}_{1} \boldsymbol{B}_{1}-\boldsymbol{X}_{2} \boldsymbol{B}_{2}\right) \Sigma^{-1 / 2}\right\|_{F}^{2} \\
& =\left\|\left(\boldsymbol{Y}-\boldsymbol{X}_{1} \hat{\boldsymbol{B}}_{1}-\boldsymbol{X}_{2} \boldsymbol{B}_{2}\right) \Sigma^{-1 / 2}\right\|_{F}^{2}
\end{aligned}
$$

where $\hat{\boldsymbol{B}}_{1}=\left(\boldsymbol{X}_{1}^{T} \boldsymbol{X}_{1}\right)^{-1} \boldsymbol{X}_{1}^{T}\left(\boldsymbol{Y}-\boldsymbol{X}_{2} \boldsymbol{B}_{2}\right)$

Then we have

$$
\begin{aligned}
& \int_{\left\|\boldsymbol{B}_{1}-\boldsymbol{B}_{01}\right\|_{F} \leq c_{1} \epsilon_{n}} \exp \left\{-\frac{1}{2} \sum_{i=1}^{n}\left(Y_{i}-X_{i 1} \boldsymbol{B}_{1}-X_{i 2} \boldsymbol{B}_{2}\right) \Sigma^{-1}\left(Y_{i}-X_{i 1} \boldsymbol{B}_{1}-X_{i 2} \boldsymbol{B}_{2}\right)^{T}\right\} d \boldsymbol{B}_{1} \\
&= \exp \left\{-\frac{1}{2} S S E\left(\boldsymbol{B}_{2}, \boldsymbol{\Sigma}\right)\right\} \times \\
& \int_{\left\|\boldsymbol{B}_{1}-\boldsymbol{B}_{01}\right\|_{F} \leq c_{1} \epsilon_{n}} \exp \left\{-\frac{1}{2} \operatorname{tr}\left(\boldsymbol{X}_{1}^{T} \boldsymbol{X}_{1}\left(\boldsymbol{B}_{1}-\hat{\boldsymbol{B}}_{1}\right) \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{B}_{1}-\hat{\boldsymbol{B}}_{1}\right)^{T}\right)\right\} d \boldsymbol{B}_{1} \\
& \geq(2 \pi)^{-\left(s_{0} q_{n} / 2\right)}|\boldsymbol{\Sigma}|^{s_{0} / 2}\left|\boldsymbol{X}_{1}^{T} \boldsymbol{X}_{1}\right|^{-q_{n} / 2} \exp \left\{-\frac{1}{2} \operatorname{SSE}\left(\boldsymbol{B}_{2}, \boldsymbol{\Sigma}\right)\right\} \operatorname{Pr}\left(\left\|\boldsymbol{B}_{1}-\hat{\boldsymbol{B}}_{1}\right\|_{F} \leq c \sqrt{s_{0} q_{n} / n}\right)
\end{aligned}
$$

where $\boldsymbol{B}_{1} \sim \operatorname{MatrixNormal}\left(\hat{\boldsymbol{B}}_{1}, \boldsymbol{\Sigma},\left(\boldsymbol{X}_{1}^{T} \boldsymbol{X}_{1}\right)^{-1}\right)$.

Let $T_{\boldsymbol{B}} \sim \operatorname{MatrixNormal}\left(\mathbf{0}, I_{q_{n}}, I_{s_{0}}\right)$, then $\boldsymbol{B}_{1}-\hat{\boldsymbol{B}}_{1}=\Sigma^{1 / 2} T_{\boldsymbol{B}}\left(\boldsymbol{X}_{1}^{T} \boldsymbol{X}_{1}\right)^{-1 / 2}$.
Since $\|\Sigma\|$ is bounded and $\left\|\left(\boldsymbol{X}_{1}^{T} \boldsymbol{X}_{1}\right)^{-1}\right\| \leq 1 /\left(n \lambda_{0}\right)$, we have

$$
\begin{aligned}
\operatorname{Pr}\left(\left\|\boldsymbol{B}_{1}-\hat{\boldsymbol{B}}_{1}\right\|_{F} \geq c \sqrt{s_{0} q_{n} / n}\right) & =\operatorname{Pr}\left(\left\|\Sigma^{1 / 2} T_{\boldsymbol{B}}\left(\boldsymbol{X}_{1}^{T} \boldsymbol{X}_{1}\right)^{-1 / 2}\right\|_{F} \geq c \sqrt{s_{0} q_{n} / n}\right) \\
& =\operatorname{Pr}\left(\left\|T_{\boldsymbol{B}}\right\|_{F}^{2} \geq \tilde{c} s_{0} q_{n}\right) \\
& =\operatorname{Pr}\left(\chi_{s_{0} q_{n}}^{2} \geq \tilde{c} s_{0} q_{n}\right) \\
& \leq \exp \left(-2 s_{0} q_{n}\right) \text { if } \tilde{c} \geq 8
\end{aligned}
$$

So we have $\operatorname{Pr}\left(\left\|\boldsymbol{B}_{1}-\hat{\boldsymbol{B}}_{1}\right\|_{F} \lesssim \sqrt{s_{0} q_{n} / n}\right) \rightarrow 1$ for large $n$.

Let function $K_{n}\left(\boldsymbol{B}_{S}\right):=\left(\boldsymbol{Y}-\boldsymbol{X}_{S} \boldsymbol{B}_{S}\right)^{T}\left(I_{n}-H_{S^{c}}\right)\left(\boldsymbol{Y}-\boldsymbol{X}_{S} \boldsymbol{B}_{S}\right)$, where $H_{S^{c}}=\boldsymbol{X}_{S^{c}}\left(\boldsymbol{X}_{S^{c}}^{T} \boldsymbol{X}_{S^{c}}\right)^{-1} \boldsymbol{X}_{S^{c}}^{T}$ $\int_{\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\| \leq c_{2} \epsilon_{n}} \inf _{\left\|B_{2} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}}|\boldsymbol{\Sigma}|^{-\left(n-s_{0}\right) / 2} \exp \left\{-\frac{1}{2} \operatorname{SSE}\left(\boldsymbol{B}_{2}, \boldsymbol{\Sigma}\right)\right\} \pi(\boldsymbol{\Sigma}) d \boldsymbol{\Sigma}$ $\geq \inf _{\left\|B_{2} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}} \int_{\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\| \leq c_{2} \epsilon_{n}}|\boldsymbol{\Sigma}|^{-\left(n-s_{0}\right) / 2} \exp \left\{-\frac{1}{2} \operatorname{tr}\left(\left(K_{n}\left(\boldsymbol{B}_{2}\right)+\Phi\right) \boldsymbol{\Sigma}^{-1}\right)\right\} \pi(\boldsymbol{\Sigma}) d \boldsymbol{\Sigma}$ $=\int_{\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_{0}\right\| \leq c_{2} \epsilon_{n}} \frac{|\Phi|^{v / 2}}{2^{v q_{n} / 2} \Gamma_{q_{n}}(v / 2)}|\boldsymbol{\Sigma}|^{-\left(v+q_{n}+1+n-s_{0}\right) / 2} \exp \left\{-\frac{1}{2} \operatorname{tr}\left(\left(K_{n}\left(\boldsymbol{B}_{2}\right)+\Phi\right) \boldsymbol{\Sigma}^{-1}\right)\right\} d \boldsymbol{\Sigma}$ $=\frac{2^{\left(n-s_{0}\right) q_{n} / 2}|\Phi|^{v / 2} \Gamma_{q_{n}}\left(\left(v+n-s_{0}\right) / 2\right)}{\left|K_{n}\left(\boldsymbol{B}_{2}\right)+\Phi\right|^{\left(v+n-s_{0}\right) / 2} \Gamma_{q_{n}}(v / 2)} \operatorname{Pr}\left(\left\|T_{\boldsymbol{Y}_{n}}-\boldsymbol{\Sigma}_{0}\right\| \leq c_{2} \epsilon_{n}\right)$
where $T_{\boldsymbol{Y}_{n}} \sim \mathcal{I} \mathcal{W}\left(v+n-s_{0}, K_{n}\left(\boldsymbol{B}_{2}\right)+\Phi\right)$ and $\Gamma_{q_{n}}(\cdot)$ is multivariate gamma function.

For simplicity, let $T_{\boldsymbol{Y}_{n}} \sim \mathcal{I} \mathcal{W}\left(\nu_{n}, \Psi_{Y_{n}}\right)$, where $\nu_{n}=v+n-s_{0} \simeq n, \Psi_{\boldsymbol{Y}_{n}}=$ $K_{n}\left(\boldsymbol{B}_{2}\right)+\Phi$.

$$
\begin{aligned}
\operatorname{Pr}\left(\left\|T_{\boldsymbol{Y}_{\boldsymbol{n}}}-\boldsymbol{\Sigma}_{0}\right\| \geq c_{2} \epsilon_{n} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right) & \leq \mathbb{E}\left(\left\|T_{\boldsymbol{Y}_{\boldsymbol{n}}}-\boldsymbol{\Sigma}_{0}\right\|^{2} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right) /\left(c_{2}^{2} \epsilon_{n}^{2}\right) \\
& \leq \mathbb{E}\left(\left\|T_{\boldsymbol{Y}_{\boldsymbol{n}}}-\boldsymbol{\Sigma}_{0}\right\|_{F}^{2} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right) /\left(c_{2}^{2} \epsilon_{n}^{2}\right) \\
& =\left\|\mathbb{E}\left(T_{\boldsymbol{Y}_{\boldsymbol{n}}}\right)-\Sigma_{0}\right\|_{F}^{2} /\left(c_{2}^{2} \epsilon_{n}^{2}\right)+\sum_{i} \sum_{j} \operatorname{Var}\left(T_{\boldsymbol{Y}_{\boldsymbol{n}}, i j}\right) /\left(c_{2}^{2} \epsilon_{n}^{2}\right)
\end{aligned}
$$

We first show that if $\operatorname{Pr}_{\left(\boldsymbol{B}_{0}, \Sigma_{0}\right)}\left(\left\|\frac{\Psi_{\boldsymbol{Y}_{n}}}{\nu_{n}-q_{n}-1}-\Sigma_{0}\right\| \geq \frac{\epsilon_{n}}{M \sqrt{q_{n}}}\right) \leq \exp \left(-\tilde{c} n \epsilon_{n}^{2} / q_{n}\right)$ holds for sufficiently large $M$, then $\left\|\mathbb{E}\left(T_{\boldsymbol{Y}_{n}}\right)-\Sigma_{0}\right\|_{F}^{2} /\left(c_{2}^{2} \epsilon_{n}^{2}\right) \rightarrow 0$ in $P_{\left(\boldsymbol{B}_{0}, \Sigma_{0}\right)}$ and $\sum_{i} \sum_{j} \operatorname{Var}\left(T_{\boldsymbol{Y}_{n}, i j}\right) /\left(c_{2}^{2} \epsilon_{n}^{2}\right) \rightarrow 0$ in $P_{\left(\boldsymbol{B}_{0}, \Sigma_{0}\right)}$.

$$
\left\|\mathbb{E}\left(T_{\boldsymbol{Y}_{n}}\right)-\Sigma_{0}\right\|_{F}^{2} / \epsilon_{n}^{2}=\left\|\frac{\Psi_{\boldsymbol{Y}_{n}}}{\nu_{n}-q_{n}-1}-\Sigma_{0}\right\|_{F}^{2} / \epsilon_{n}^{2} \leq \frac{q_{n}}{\epsilon_{n}^{2}}\left\|\frac{\Psi_{\boldsymbol{Y}_{n}}}{\nu_{n}-q_{n}-1}-\Sigma_{0}\right\| \rightarrow
$$ 0 in $P_{\left(\boldsymbol{B}_{0}, \Sigma_{0}\right)}$

$$
\operatorname{Var}\left(T_{\boldsymbol{Y}_{n}}\right)= \begin{cases}\frac{\left(\nu_{n}-q_{n}+1\right) \Psi_{\boldsymbol{Y}_{n}, i j}^{2}+\left(\nu_{n}-q_{n}-1\right) \Psi_{\boldsymbol{Y}_{n}, i i} \Psi_{\boldsymbol{Y}_{n}, j j}}{\left(\nu_{n}-q_{n}\right)\left(\nu_{n}-q_{n}-1\right)^{2}\left(\nu_{n}-q_{n}-3\right)}, & i \neq j \\ \frac{2 \Psi_{\boldsymbol{Y}_{n}, i i}^{2}}{\left(\nu_{n}-q_{n}-1\right)^{2}\left(\nu_{n}-q_{n}-3\right)}, & i=j\end{cases}
$$

Because $\left\|\frac{\Psi_{\boldsymbol{Y}_{n}}}{\nu_{n}-q_{n}-1}-\Sigma_{0}\right\|_{F} \leq\left\|\frac{\Psi_{\boldsymbol{Y}_{n}}}{\nu_{n}-q_{n}-1}-\Sigma_{0}\right\| \rightarrow 0$ in $P_{\left(\boldsymbol{B}_{0}, \Sigma_{0}\right)}$, i.e. $\frac{\Psi_{\boldsymbol{Y}_{n}, i j}}{\nu_{n}-q_{n}-1} \rightarrow \Sigma_{0, i j}$ in $P_{\left(\boldsymbol{B}_{0}, \Sigma_{0}\right)}, \operatorname{Var}\left(T_{\boldsymbol{Y}_{n}, i j} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)=\mathcal{O}_{P_{\left(B_{0}, \Sigma_{0}\right)}}(1 / n)$. Hence $\sum_{i} \sum_{j} \operatorname{Var}\left(T_{\boldsymbol{Y}_{n}, i j}\right) /\left(c_{2}^{2} \epsilon_{n}^{2}\right) \leq \mathcal{O}_{P_{\left(\boldsymbol{B}_{0}, \Sigma_{0}\right)}}\left(\frac{q_{n}^{2}}{n \epsilon_{n}^{2}}\right) \rightarrow 0$ in $P_{\left(\boldsymbol{B}_{0}, \Sigma_{0}\right)}$.

Now, it suffices to show $\operatorname{Pr}_{\left(\boldsymbol{B}_{0}, \Sigma_{0}\right)}\left(\left\|\frac{\Psi_{\boldsymbol{Y}_{n}}}{\nu_{n}-q_{n}-1}-\Sigma_{0}\right\| \geq \frac{\epsilon_{n}}{M \sqrt{q_{n}}}\right) \leq \exp \left(-\tilde{c} n \epsilon_{n}^{2} / q_{n}\right)$ for sufficiently large $M$.

By Lemma 1, let $\tilde{\epsilon_{n}}=\frac{\epsilon_{n}}{M \sqrt{q_{n}}}, d=q_{n}, \Sigma=\Sigma_{0}, A=\varepsilon_{n} \Sigma_{0}^{1 / 2}, P=I_{n}-H_{1}$, $U=-X_{2} \boldsymbol{B}_{2}$, it is easy to verify that $\tilde{\epsilon_{n}} \rightarrow 0, n{\tilde{\epsilon_{n}}}^{2} \rightarrow \infty, q_{n} \prec n \tilde{\epsilon}_{n}^{2}, \nu_{n}-q_{n}-1 \simeq$ $n$ and since $a_{n} p_{n} \lesssim(\log n / n)^{1 / 2}$, we have

$$
\left\|\left(I_{n}-H_{1}\right) X_{2} \boldsymbol{B}_{2} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F}^{2} \leq\left\|X_{2} \boldsymbol{B}_{2} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{F}^{2} \leq n p_{n}^{2} a_{n}^{2} \lesssim \log n \lesssim n \tilde{\epsilon_{n}^{2}}
$$

Hence, $\operatorname{Pr}\left(\left\|\frac{S_{n}}{\nu_{n}-q_{n}-1}-\Sigma_{0}\right\| \geq \tilde{\epsilon_{n}}\right) \leq \exp \left(-\tilde{c} n{\tilde{\epsilon_{n}}}^{2}\right)$ by Lemma 1. And $\left\|\frac{\Phi}{\nu_{n}-q_{n}-1}\right\| \leq$ $\left\|\frac{\Phi}{\nu_{n}-q_{n}-1}\right\|_{F} \lesssim \frac{q_{n}}{n} \prec \tilde{\epsilon_{n}}$. We have $\operatorname{Pr}_{\left(\boldsymbol{B}_{0}, \Sigma_{0}\right)}\left(\left\|\frac{\Psi_{\boldsymbol{Y}_{n}}}{\nu_{n}-q_{n}-1}-\Sigma_{0}\right\| \geq \frac{\epsilon_{n}}{M \sqrt{q_{n}}}\right) \leq$ $\exp \left(-\tilde{c} n \epsilon_{n}^{2} / q_{n}\right)$ for sufficiently large $M$. Therefore, $\operatorname{Pr}\left(\left\|T_{\boldsymbol{Y}_{n}}-\boldsymbol{\Sigma}_{0}\right\| \lesssim \epsilon_{n}\right) \rightarrow 1$ for large $n$.

We now calculate $\pi\left(S=S^{\prime} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)$ where $S^{\prime} \supset S_{0}$. For notation simplicity, we abuse subscripts a little bit. Now let subscripts " 1 ", " 2 " and " 3 " denote model $S_{0}, S^{\prime} \backslash S_{0}$ and $\left(S^{\prime}\right)^{c}$ respectively.

$$
\begin{aligned}
& \pi\left(S=S^{\prime} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right) \propto \int|\boldsymbol{\Sigma}|^{-n / 2} \exp \left\{-\frac{1}{2} \sum_{i=1}^{n}\left(Y_{i}-X_{i} \boldsymbol{B}\right) \Sigma^{-1}\left(Y_{i}-X_{i} \boldsymbol{B}\right)^{T}\right\} \pi(\boldsymbol{B}, \boldsymbol{\Sigma}) \\
& I\left(\left\|B_{3} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}\right) I\left(\min _{j}\left\|B_{2 j} \boldsymbol{\Sigma}^{-1 / 2}\right\|>a_{n}\right) d \boldsymbol{B} d \boldsymbol{\Sigma} \\
& \lesssim \pi\left(\left\|B_{3} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}, \min _{j}\left\|B_{2 j} \boldsymbol{\Sigma}^{-1 / 2}\right\|>a_{n}\right) \times \\
& \sup _{\left\|B_{3} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2, \infty} \lesssim a_{n}} \int_{E_{1}}|\boldsymbol{\Sigma}|^{-n / 2} \exp \left\{-\frac{1}{2} \sum_{i=1}^{n}\left(Y_{i}-X_{i} \boldsymbol{B}\right) \Sigma^{-1}\left(Y_{i}-X_{i} \boldsymbol{B}\right)^{T}\right\} \pi\left(\boldsymbol{B}_{1}, \boldsymbol{\Sigma}\right) d \boldsymbol{B}_{1} d \boldsymbol{\Sigma}
\end{aligned}
$$

$$
\begin{aligned}
& \int_{E_{1}}|\boldsymbol{\Sigma}|^{-n / 2} \exp \left\{-\frac{1}{2} \sum_{i=1}^{n}\left(Y_{i}-X_{i} \boldsymbol{B}\right) \Sigma^{-1}\left(Y_{i}-X_{i} \boldsymbol{B}\right)^{T}\right\} \pi\left(\boldsymbol{B}_{1}, \boldsymbol{\Sigma}\right) d \boldsymbol{B}_{1} d \boldsymbol{\Sigma} \\
\leq & \overline{\pi\left(\boldsymbol{B}_{1} \mid \boldsymbol{\Sigma}\right)} \exp \left\{-\frac{1}{2} S S E\left(\boldsymbol{B}_{2}, \boldsymbol{B}_{3}, \boldsymbol{\Sigma}\right)\right\} \times \\
& \int|\boldsymbol{\Sigma}|^{-n / 2} \exp \left\{-\frac{1}{2} \operatorname{tr}\left(\boldsymbol{X}_{1}^{T} \boldsymbol{X}_{1}\left(\boldsymbol{B}_{1}-\tilde{\boldsymbol{B}}_{1}\right) \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{B}_{1}-\tilde{\boldsymbol{B}}_{1}\right)^{T}\right)\right\} d \boldsymbol{B}_{1} \\
= & \overline{\pi\left(\boldsymbol{B}_{1} \mid \boldsymbol{\Sigma}\right)}(2 \pi)^{-\left(s_{0} q_{n} / 2\right)}\left|\boldsymbol{X}_{1}^{T} \boldsymbol{X}_{1}\right|^{-q_{n} / 2} \int|\boldsymbol{\Sigma}|^{-\left(n-s_{0}\right) / 2} \exp \left\{-\frac{1}{2} S S E\left(\boldsymbol{B}_{2}, \boldsymbol{B}_{3}, \boldsymbol{\Sigma}\right)\right\} \pi(\boldsymbol{\Sigma}) d \boldsymbol{\Sigma} \\
= & \overline{\pi\left(\boldsymbol{B}_{1} \mid \boldsymbol{\Sigma}\right)}(2 \pi)^{-\left(s_{0} q_{n} / 2\right)}\left|\boldsymbol{X}_{1}^{T} \boldsymbol{X}_{1}\right|^{-q_{n} / 2} \frac{2^{\left(n-s_{0}\right) q_{n} / 2}|\Phi|^{v / 2} \Gamma_{q_{n}}\left(\left(v+n-s_{0}\right) / 2\right)}{\left|K_{n}\left(\boldsymbol{B}_{2}, \boldsymbol{B}_{3}\right)+\Phi\right|^{\left(v+n-s_{0}\right) / 2} \Gamma_{q_{n}}(v / 2)}
\end{aligned}
$$

Part III: Now we want to show $\sum_{S^{\prime} \supseteq S_{0},\left|S^{\prime} \backslash S_{0}\right| \geq 1} \frac{\pi\left(S=S^{\prime} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)}{\pi\left(S=S_{0} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)} \rightarrow 0$.

$\frac{\sup _{\left\|B_{S_{0}^{c}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}}\left|K_{n}\left(\boldsymbol{B}_{S_{0}^{c}}\right)+\Phi\right|^{\left(v+n-s_{0}\right) / 2}}{\inf _{\left\|B_{S^{\prime} c} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}}\left|K_{n}\left(\boldsymbol{B}_{S^{\prime} \backslash S_{0}}, \boldsymbol{B}_{S^{\prime c}}\right)+\Phi\right|^{\left(v+n-s_{0}\right) / 2}}$
Firstly,

$$
\frac{\overline{\frac{\pi\left(\boldsymbol{B}_{1} \mid \boldsymbol{\Sigma}\right)}{\left(\boldsymbol{B}_{1} \mid \boldsymbol{\Sigma}\right)}}}{\underline{\pi( })} l_{n}^{s_{0}} \text { by } \quad(7)
$$

And we also have

$$
\begin{aligned}
& \frac{\pi\left(\left\|B_{\left(S^{\prime}\right)^{c}} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}, \min _{j}\left\|\left(B_{S^{\prime} \backslash S_{0}} \boldsymbol{\Sigma}^{-1 / 2}\right)_{j}\right\|>a_{n}\right)}{\pi\left(\left\|B_{S_{0}^{c}} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}\right)} \\
= & \frac{\pi\left(\min _{j}\left\|\left(B_{S^{\prime} \backslash S_{0}} \boldsymbol{\Sigma}^{-1 / 2}\right)_{j}\right\|>a_{n}\right)}{\pi\left(\left\|B_{\left(S^{\prime} \backslash S_{0}\right)^{c}} \boldsymbol{\Sigma}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}\right)} \\
\leq & \left(\frac{p_{n}^{-(1+u)}}{1-p_{n}^{-(1+u)}}\right)^{\left|S^{\prime} \backslash S_{0}\right|}
\end{aligned}
$$

Next, we examine the behavior of $\frac{\left|K_{n}\left(\boldsymbol{B}_{S_{0}^{c}}\right)+\Phi\right|}{\left|K_{n}\left(\boldsymbol{B}_{S^{\prime} c}\right)+\Phi\right|}$ under constraints $\left\|B_{S_{0}^{c}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2, \infty} \leq$ $a_{n}$ and $\left\|B_{S^{\prime} c} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}$. It is easy to verify that

$$
\frac{\left|K_{n}\left(\boldsymbol{B}_{S_{0}^{c}}\right)+\Phi\right|}{\left|K_{n}\left(\boldsymbol{B}_{S^{\prime} c}\right)+\Phi\right|} \leq \frac{\left|K_{n}\left(\boldsymbol{B}_{S_{0}^{c}}\right)+\Phi\right|}{\left|K_{n}\left(\boldsymbol{B}_{S^{\prime} c}\right)\right|}
$$

where

$$
\begin{aligned}
K_{n}\left(\boldsymbol{B}_{S_{0}^{c}}\right)= & \left(\boldsymbol{Y}_{\boldsymbol{n}}-\boldsymbol{X}_{S_{0}^{c}} \boldsymbol{B}_{S_{0}^{c}}\right)^{T}\left(I_{n}-H_{S_{0}}\right)\left(\boldsymbol{Y}_{\boldsymbol{n}}-\boldsymbol{X}_{S_{0}^{c}} \boldsymbol{B}_{S_{0}^{c}}\right) \\
= & \Sigma_{0}^{1 / 2} \varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S_{0}}\right) \varepsilon_{\boldsymbol{n}} \Sigma_{0}^{1 / 2}+\left(\boldsymbol{X}_{S_{0}^{c}} \boldsymbol{B}_{S_{0}^{c}}\right)^{T}\left(I_{n}-H_{S_{0}}\right)\left(\boldsymbol{X}_{S_{0}^{c}} \boldsymbol{B}_{S_{0}^{c}}\right) \\
& -2 \Sigma_{0}^{1 / 2} \varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S_{0}}\right)\left(\boldsymbol{X}_{S_{0}^{c}} \boldsymbol{B}_{S_{0}^{c}}\right) \\
K_{n}\left(\boldsymbol{B}_{S^{\prime c}}\right)= & \left(\boldsymbol{Y}_{\boldsymbol{n}}-\boldsymbol{X}_{S^{\prime}} \boldsymbol{B}_{S^{\prime} c}\right)^{T}\left(I_{n}-H_{S^{\prime}}\right)\left(\boldsymbol{Y}_{\boldsymbol{n}}-\boldsymbol{X}_{S^{\prime c}} \boldsymbol{B}_{S^{\prime c}}\right) \\
= & \Sigma_{0}^{1 / 2} \varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S^{\prime}}\right) \varepsilon_{\boldsymbol{n}} \Sigma_{0}^{1 / 2} \\
& +\left(\boldsymbol{X}_{S_{0}} \boldsymbol{B}_{S_{0}}-\boldsymbol{X}_{S^{\prime c}} \boldsymbol{B}_{S^{\prime} c}\right)^{T}\left(I_{n}-H_{S^{\prime}}\right)\left(\boldsymbol{X}_{S_{0}} \boldsymbol{B}_{S_{0}}-\boldsymbol{X}_{S^{\prime} c} \boldsymbol{B}_{S^{\prime c}}\right) \\
& -2 \Sigma_{0}^{1 / 2} \varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S^{\prime}}\right)\left(\boldsymbol{X}_{S_{0}} \boldsymbol{B}_{S_{0}}-\boldsymbol{X}_{S^{\prime} c} \boldsymbol{B}_{S^{\prime c}}\right)
\end{aligned}
$$

Because $\left\|\boldsymbol{X}_{S_{0}^{c}} \boldsymbol{B}_{S_{0}^{c}}\right\|_{F}^{2} \leq\left\|\boldsymbol{X}_{S_{0}^{c}}\right\|_{F}^{2}\left\|\boldsymbol{B}_{S_{0}^{c}}\right\|_{F}^{2} \lesssim n p_{n}^{2} a_{n}^{2} \lesssim n \epsilon_{n}^{2} \prec n$ and

$$
\begin{aligned}
\left\|\Sigma_{0}^{1 / 2} \varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S_{0}}\right)\left(\boldsymbol{X}_{S_{0}^{c}} \boldsymbol{B}_{S_{0}^{c}}\right)\right\|_{F} & \lesssim\left\|\varepsilon_{\boldsymbol{n}}\right\|\left\|\boldsymbol{X}_{S_{0}^{c}} \boldsymbol{B}_{S_{0}^{c}}\right\|_{F} \\
& \lesssim \sqrt{n} \sqrt{n \epsilon_{n}^{2}} \text { with probability } \rightarrow 1 \\
& \prec n
\end{aligned}
$$

we have $\left|K_{n}\left(\boldsymbol{B}_{S_{0}^{c}}\right)\right| \simeq\left|\varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S_{0}}\right) \varepsilon_{\boldsymbol{n}}\right|\left|\Sigma_{0}\right|$.
Similarly, $\left\|\left(\boldsymbol{X}_{S_{0}} \boldsymbol{B}_{S_{0}}-\boldsymbol{X}_{S^{\prime}} \boldsymbol{B}_{S^{\prime c}}\right)\right\|_{F}^{2} \prec n$ and $\| \Sigma_{0}^{1 / 2} \varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S^{\prime}}\right)\left(\boldsymbol{X}_{S_{0}} \boldsymbol{B}_{S_{0}}-\right.$ $\left.\boldsymbol{X}_{S^{\prime}} \boldsymbol{B}_{S^{\prime c}}\right) \|_{F} \prec n$, so we have $\left|K_{n}\left(\boldsymbol{B}_{S^{\prime c}}\right)\right| \simeq\left|\varepsilon_{\boldsymbol{n}}{ }^{T}\left(I_{n}-H_{S^{\prime}}\right) \varepsilon_{\boldsymbol{n}} \| \Sigma_{0}\right|$.

By the bounds for eigenvalues in Vershynin [32, with probability going to 1,

$$
\begin{aligned}
& n-s_{0} \lesssim \lambda_{\min }\left(\varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S_{0}}\right) \varepsilon_{\boldsymbol{n}}\right) \leq \lambda_{\max }\left(\varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S_{0}}\right) \varepsilon_{\boldsymbol{n}}\right) \lesssim n-s_{0} \\
& n-s_{0} \lesssim \lambda_{\min }\left(\varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S^{\prime}}\right) \varepsilon_{\boldsymbol{n}}\right) \leq \lambda_{\max }\left(\varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S^{\prime}}\right) \varepsilon_{\boldsymbol{n}}\right) \lesssim n-s_{0}
\end{aligned}
$$

Therefore,

$$
\frac{\left|K_{n}\left(\boldsymbol{B}_{S_{0}^{c}}\right)+\Phi\right|}{\left|K_{n}\left(\boldsymbol{B}_{S^{\prime}}\right)\right|} \simeq \frac{\left|\varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S_{0}}\right) \varepsilon_{\boldsymbol{n}}\right|}{\left|\varepsilon_{\boldsymbol{n}}^{T}\left(I_{n}-H_{S^{\prime}}\right) \varepsilon_{\boldsymbol{n}}\right|} \lesssim\left(\frac{n-s_{0}}{n-s^{\prime}}\right)^{q_{n}} .
$$

Note that $\left(\frac{n-s^{\prime}}{n-s_{0}}\right)^{q_{n}\left(v+n-s_{0}\right)} \simeq\left(\left(1-\frac{s^{\prime}-s_{0}}{n-s_{0}}\right)^{n-s_{0}}\right)^{q_{n}} \geq \exp \left(-c\left(s^{\prime}-s_{0}\right) q_{n}\right) \geq$ $\exp \left(-c\left(s^{\prime}-s_{0}\right) \log p_{n}\right)$ because $\left(1-\frac{s^{\prime}-s_{0}}{n-s_{0}}\right)^{n-s_{0}} \geq \exp \left(-c\left(s^{\prime}-s_{0}\right)\right)$ for large $n, c>1$ and $q_{n} \lesssim \log p_{n}$.

Therefore,

$$
\frac{\sup _{\left\|\boldsymbol{B}_{S_{0}^{c}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}}\left|K_{n}\left(\boldsymbol{B}_{S_{0}^{c}}\right)+\Phi\right|^{\left(v+n-s_{0}\right) / 2}}{\inf _{\left\|\boldsymbol{B}_{S^{\prime}} \boldsymbol{\Sigma}_{0}^{-1 / 2}\right\|_{2, \infty} \leq a_{n}}\left|K_{n}\left(\boldsymbol{B}_{S^{\prime c}}\right)+\Phi\right|^{\left(v+n-s_{0}\right) / 2}} \leq \exp \left(c\left|S^{\prime} \backslash S_{0}\right| \log p_{n}\right)
$$

Combining the above parts, we have $\frac{\pi\left(S=S^{\prime} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)}{\pi\left(S=S_{0} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)} \leq l_{n}^{s_{0}}\left(1+p_{n}^{c-(1+u)}\right)^{\left|S^{\prime} \backslash S_{0}\right|}$ and
$\sum_{S^{\prime} \supseteq S_{0},\left|S^{\prime} \backslash S_{0}\right| \geq 1} \frac{\pi\left(S=S^{\prime} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)}{\pi\left(S=S_{0} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)} \leq l_{n}^{s_{0}}\left(\left(1+p_{n}^{-(1+u-c)}\right)^{p_{n}}-1\right) \simeq l_{n}^{s_{0}} p_{n}^{-(u-c)}$.
By (7), we can get $\sum_{S^{\prime} \supseteq S_{0},\left|S^{\prime} \backslash S_{0}\right| \geq 1} \frac{\pi\left(S=S^{\prime} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)}{\pi\left(S=S_{0} \mid \boldsymbol{Y}_{\boldsymbol{n}}\right)} \rightarrow 1$ for $1 \leq c<u$.

Proof. (Theorem 3) It is easy to verify that when $\boldsymbol{B}$ follows the given distribution, $g_{\tau}(\cdot)$ takes the form $g_{\tau}(\cdot)=g(\cdot / \sqrt{\tau}) / \tau^{q_{n} / 2}$, where $\tau=\tau_{n}$ and

$$
g(\boldsymbol{x})=\int_{0}^{\infty}(2 \pi)^{-q_{n} / 2} \xi^{-q_{n} / 2} \exp \left(-\frac{\|\boldsymbol{x}\|_{2}^{2}}{2 \xi}\right) \cdot K \xi^{-r} L(\xi) d \xi
$$

We first show that (4) holds when $L(\xi)$ has upper bound. When $L(\xi) \leq C_{12}$, we have

$$
g(\boldsymbol{x}) \leq C_{12} K 2^{r-1} \pi^{-q_{n} / 2} \Gamma\left(\frac{d}{2}+r-1\right)\|\boldsymbol{x}\|_{2}^{-(d+2 r-2)}
$$

And by Lemma (3),

$$
\begin{aligned}
\int_{\|\boldsymbol{x}\|_{2} \geq a_{n}} g_{\tau}(\boldsymbol{x}) d \boldsymbol{x} & =\int_{\|\boldsymbol{x}\|_{2} \geq a_{n}} g\left(\boldsymbol{x} / \sqrt{\tau}_{n}\right) / \tau_{n}^{q_{n} / 2} d \boldsymbol{x} \\
& =\int_{\|\boldsymbol{z}\|_{2} \geq a_{n} / \sqrt{\tau_{n}}} g(\boldsymbol{z}) d \boldsymbol{z} \quad\left(\boldsymbol{z}=\boldsymbol{x} / \sqrt{\tau} \bar{n}_{n}\right) \\
& \leq C_{12} K 2^{r-1} \pi^{-q_{n} / 2} \Gamma\left(\frac{q_{n}}{2}+r-1\right) \int_{\|\boldsymbol{z}\|_{2} \geq a_{n} / \sqrt{\tau_{n}}}\|\boldsymbol{z}\|_{2}^{-(d+2 r-2)} d \boldsymbol{z} \\
& \leq \frac{C_{12} K 2^{r-1}}{2(r-1)}\left(\frac{d}{2}+r-1\right)^{r}\left(\frac{\tau_{n}}{a_{n}^{2}}\right)^{r-1} \\
& \leq p_{n}^{-(1+u)} \quad\left(0<u<u^{\prime}\right)
\end{aligned}
$$

since $q_{n} / 2+r-1 \lesssim \log p_{n} \prec p_{n}^{u^{\prime}-u}$ for any $0<u<u^{\prime}$ and $\left(\tau_{n} / a_{n}^{2}\right)^{r-1} \lesssim p_{n}^{-\left(1+u^{\prime}\right)}$. $L(\xi) \leq 1$ is a special case of $L(\xi) \leq C_{12}$.

Now, we want to show that (5) holds with $\log \tau_{n} \gtrsim-\log p_{n}$. It is obvious that $g(\boldsymbol{x})$ is a decreasing function of $\|\boldsymbol{x}\|_{2}$, so $\inf _{\|\boldsymbol{x}\|_{2} \leq M_{0}} g_{\tau_{n}}(\boldsymbol{x}) \geq\left. g_{\tau_{n}}(\boldsymbol{x})\right|_{\|\boldsymbol{x}\|_{2}=M_{0}}$. When $L(\xi) \geq 1-C_{11} \xi^{-t}$,
$g(\boldsymbol{x}) \geq K 2^{r-1} \pi^{-q_{n} / 2} \Gamma\left(\frac{q_{n}}{2}+r-1\right)\|\boldsymbol{x}\|_{2}^{-(d+2 r-2)}\left(1-C_{11} 2^{t} \frac{\Gamma\left(\frac{d}{2}+r+t-1\right)}{\Gamma\left(\frac{d}{2}+r-1\right)}\|\boldsymbol{x}\|_{2}^{-2 t}\right)$.
Since $\tau_{n} \lesssim a_{n}^{2} p_{n}^{-\left(1+u^{\prime}\right) /(r-1)}$,

$$
C_{11} 2^{t} \frac{\Gamma\left(\frac{q_{n}}{2}+r+t-1\right)}{\Gamma\left(\frac{q_{n}}{2}+r-1\right)}\left\|M_{0} / \sqrt{\tau_{n}}\right\|_{2}^{-2 t} \leq C_{11} M_{0}^{-2 t}\left[(d+2(r+t-1)) \tau_{n}\right]^{t} \rightarrow 0 .
$$

So we have

$$
\begin{aligned}
\left.\log g_{\tau_{n}}(\boldsymbol{x})\right|_{\|\boldsymbol{x}\|_{2}=M_{0}} & \geq \log \left(\tau_{n}^{-q_{n} / 2} \tilde{K} 2^{r-1} \pi^{-q_{n} / 2} \Gamma\left(\frac{q_{n}}{2}+r-1\right)\left\|M_{0} / \sqrt{\tau_{n}}\right\|_{2}^{-\left(q_{n}+2 r-2\right)}\right) \\
& \geq \mathrm{constant}-d \log \left(\sqrt{\pi} M_{0}\right)+(r-1) \log \tau_{n} \\
& \gtrsim-\log p_{n}
\end{aligned}
$$

because $d \lesssim \log p_{n}$ and $\log \tau_{n} \gtrsim-\log p_{n}$.
When $L(\xi) \geq C_{21} \xi^{-t_{1}}, g(\boldsymbol{x}) \geq K C_{21} 2^{t_{1}+r-1} \pi^{-d / 2} \Gamma\left(\frac{d}{2}+t_{1}+r-1\right)\|\boldsymbol{x}\|_{2}^{-\left(d+2\left(t_{1}+r-1\right)\right)}$, the rest follows the above inequalities.

Corollary 1. Polynomial-tailed distribution: Student's t-distribution, TPBN, HIB, GDP and Horseshoe+ satisfy either condition (1) or (2) in Theorem (3).

Proof. For t-distribution, let $L(\xi)=\exp (-a / \xi), a>0$, and $1-a \xi^{-1} \leq$ $\exp (-a / \xi) \leq 1$ for $\xi>0$.
For TPBN distribution, let $L(\xi)=(\xi /(1+\xi))^{a+u} \leq 1, a, u>0$, and by Bernoulli inequality, $(\xi /(1+\xi))^{a+u}=(1+1 / \xi)^{-(a+u)} \geq 1-(a+u) \xi^{-1}$ for $\xi>0$. Note that Horseshoe and NEG are special cases of TPBN with $a=u=1 / 2$ and $u=1$, respectively.

For HIB distribution, let
$L(\xi)=\left(1 \vee \phi^{2}\right) e^{s}(\xi /(1+\xi))^{a+u} \exp \left(-\frac{s}{1+\xi}\right)\left(\phi^{2}+\frac{1-\phi^{2}}{1+\xi}\right)^{-1}, \quad a, u, \phi^{2}>0, s \in \mathbb{R}$.
Because $1 \wedge \frac{1}{\phi^{2}} \leq\left(\frac{1-\phi^{2}}{1+\xi}\right)^{-1} \leq 1 \vee \frac{1}{\phi^{2}}$, we have $1-(a+u) \xi^{-1} \leq L(\xi) \leq\left(\phi^{2} \vee \frac{1}{\phi^{2}}\right) e^{s}$.
For GDP distribution, $L(\xi)=\int_{0}^{\infty} t^{a} \exp (-t-\eta \sqrt{2 t / \xi}) d t$. Because $1-\sqrt{2} \eta t^{1 / 2} \xi^{-1 / 2} \leq$ $\exp (-\eta \sqrt{2 t / \xi}) \leq 1$, we have $1-\sqrt{2} \eta \frac{\Gamma(a+3 / 2)}{\Gamma(a+1)} \xi^{-1 / 2} \leq L(\xi) / \Gamma(a+1) \leq 1$.

For Horseshoe+ distribution, $L(\xi)=\xi^{3 / 4}(\xi-1)^{-1} \log \xi / 4, \xi>1$. Note that $\log \xi / 4=\log \xi^{1 / 4} \leq \xi^{1 / 4}-1$, so $L(\xi) \leq 1$. And $\log \xi \geq 1-\xi^{-1}>0$, so $L(\xi) \geq \xi^{-1 / 4} / 4$.

Lemma 1. Let $A$ be a $n \times d$ random matrix with independent rows $A_{i} \sim$ $\mathcal{N}_{d}(0, \boldsymbol{\Sigma})$. If $\epsilon_{n} \rightarrow 0$ and $n \epsilon_{n}^{2} \rightarrow \infty$ as $n \rightarrow \infty$ and $d \prec n \epsilon_{n}^{2}$, then for any $n \times n$ projection matrix $P$ with rank $r \simeq n$ and any $n \times d$ fixed matrix $U$ with $\left\|P U \Sigma^{-1 / 2}\right\|_{F}^{2} \lesssim n \epsilon_{n}^{2}$, we have the following inequalities for some constants $K, c>0$ and sufficiently large $n$,

$$
\begin{aligned}
& \operatorname{Pr}\left(\left\|\frac{1}{r} A^{T} P A-\boldsymbol{\Sigma}\right\| \geq K\|\boldsymbol{\Sigma}\| \epsilon_{n}\right) \leq e^{-c n \epsilon_{n}^{2}} \\
& \operatorname{Pr}\left(\left\|\frac{1}{r}(A+U)^{T} P(A+U)-\boldsymbol{\Sigma}\right\| \geq 2 K\|\boldsymbol{\Sigma}\| \epsilon_{n}\right) \leq e^{-c n \epsilon_{n}^{2} / 2}
\end{aligned}
$$

Proof. Let $P=Q^{T} \Lambda Q$ be a spectral decomposition of P , where the first $r$ diagonal elements of $\Lambda$ are 1 and the rest $n-r$ elements are 0 . Because $Q$ is orthogonal, $Z=Q A$ is a $n \times d$ matrix with independent rows $Z_{i} \sim \mathcal{N}_{d}(0, \boldsymbol{\Sigma})$ and $A^{T} P A=Z^{T} \Lambda Z=\sum_{i=1}^{r} Z_{i} Z_{i}^{T}$, where $Z_{i}^{T}$ is the $i$ th row of $Z$. By Theorem 5.39 and Remark 5.40 in Vershynin [32],

$$
\operatorname{Pr}\left(\left\|\frac{1}{r} \sum_{i=1}^{r} Z_{i} Z_{i}^{T}-\boldsymbol{\Sigma}\right\| \geq \max \left(\delta, \delta^{2}\right)\|\boldsymbol{\Sigma}\|\right) \leq 2 e^{-c t^{2}} \quad \text { where } \delta=C \sqrt{d / r}+t / \sqrt{r}
$$

Let $t=\sqrt{n} \epsilon_{n}$, the first part is proved because $\max \left(\delta, \delta^{2}\right) \leq K \epsilon_{n}$ for large $K$.
Now we prove the second inequality. Let $E=A \boldsymbol{\Sigma}^{-1 / 2}$, then $E$ has iid standard normal entries. It suffices to show

$$
\operatorname{Pr}\left(\left\|\frac{1}{r}\left(E+U \boldsymbol{\Sigma}^{-1 / 2}\right)^{T} P\left(E+U \boldsymbol{\Sigma}^{-1 / 2}\right)-I_{d}\right\| \geq 2 K \epsilon_{n}\right) \leq e^{-c n \epsilon_{n}^{2} / 2}
$$

since $\left\|\frac{1}{r}(A+U)^{T} P(A+U)-\boldsymbol{\Sigma}\right\| \leq\left\|\frac{1}{r}\left(E+U \boldsymbol{\Sigma}^{-1 / 2}\right)^{T} P\left(E+U \boldsymbol{\Sigma}^{-1 / 2}\right)-I_{q_{n}}\right\|\|\boldsymbol{\Sigma}\|$
Because $\left\|P U \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}^{2} \lesssim n \epsilon_{n}^{2}$, we get $\frac{1}{r}\left\|P U \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F} \prec \epsilon_{n}$. By first part of Lemma (1), $\operatorname{Pr}\left(\left\|\frac{1}{n} E^{T} E-I_{d}\right\| \geq K \epsilon_{n}\right) \leq e^{-c n \epsilon_{n}^{2}}$ and by the triangle inequality $\left\|\frac{1}{n} E^{T} E-I_{d}\right\| \geq \frac{1}{n}\left\|E^{T} E\right\|-1$, we have $\operatorname{Pr}\left(\left\|E^{T} E\right\| \geq 2 n\right) \leq e^{-c n \epsilon_{n}^{2}}$. Conditioned on $\left\|E^{T} E\right\| \leq 2 n, \frac{2}{r}\left\|E^{T} P U \boldsymbol{\Sigma}^{-1 / 2}\right\| \leq \frac{2}{r}\left\|E^{T} E\right\|^{1 / 2}\left\|P U \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F} \leq K \epsilon_{n} / 2$ for
sufficiently large $K$. So we have

$$
\begin{aligned}
& \operatorname{Pr}\left(\left\|\frac{1}{r}\left(E+U \boldsymbol{\Sigma}^{-1 / 2}\right)^{T} P\left(E+U \boldsymbol{\Sigma}^{-1 / 2}\right)-I_{d}\right\| \geq 2 K \epsilon_{n}\right) \\
\leq & \operatorname{Pr}\left(\left.\frac{1}{r}\left\|P U \boldsymbol{\Sigma}^{-1 / 2}\right\|_{F}+\frac{2}{r}\left\|E^{T} P U \boldsymbol{\Sigma}^{-1 / 2}\right\|+\left\|\frac{1}{r} E^{T} P E-I_{d}\right\| \geq 2 K \epsilon_{n} \right\rvert\,\left\|E^{T} E\right\| \leq 2 n\right) \\
& +\operatorname{Pr}\left(\left\|E^{T} E\right\| \geq 2 n\right) \\
\leq & \operatorname{Pr}\left(\left\|\frac{1}{r} E^{T} P E-I_{d}\right\| \geq K \epsilon_{n}\right)+e^{-c n \epsilon_{n}^{2}} \\
\leq & e^{-c n \epsilon_{n}^{2} / 2} \text { for large } n
\end{aligned}
$$

Lemma 2. For compatible matrices $A$ and $B$, we have $|\operatorname{tr}(A B)| \leq\left\|A^{T}\right\|_{2, \infty}\|B\|_{2,1}$.
Further, if $B$ is a square matrix, then $\|A B\|_{2,1} \leq \sqrt{\lambda_{\max }\left(B B^{T}\right)}\|A\|_{2,1}$
Proof.

$$
\begin{gathered}
\operatorname{tr}(A B) \leq \sum_{j=1}^{m} \sum_{i=1}^{n}\left|a_{i j} b_{j i}\right| \leq \sum_{j=1}^{m} \sqrt{\sum_{i=1}^{n} a_{i j}^{2} \sum_{i=1}^{n} b_{j i}^{2}} \leq\left\|A^{T}\right\|_{2, \infty} \sum_{i=1}^{n}\|B\|_{2,1} \\
\|A B\|_{2,1}=\sum_{i=1}^{n} \sqrt{A_{i} B B^{T} A_{i}^{T}} \leq \sum_{i=1}^{n} \sqrt{\lambda_{\max }\left(B B^{T}\right)}\left\|A_{i}\right\|_{2}=\sqrt{\lambda_{\max }\left(B B^{T}\right)}\|A\|_{2,1}
\end{gathered}
$$

Lemma 3. For a d-dimentional vector $\boldsymbol{x}$,

$$
\int_{\|\boldsymbol{x}\|_{2} \geq a}\|\boldsymbol{x}\|_{2}^{-(d+k)} d \boldsymbol{x}=\frac{2 \pi^{d / 2} a^{-k}}{k \Gamma(d / 2)} \quad k, a>0
$$

Proof. The result is immediate by polar coordinate transformation in Scott [27],

$$
\int_{\|\boldsymbol{x}\|_{2} \geq a}\|\boldsymbol{x}\|_{2}^{-(d+k)} d \boldsymbol{x}=\left(\int_{a}^{\infty} r^{-k-1} d r\right)\left(\prod_{i=1}^{d-2} \int_{-\pi / 2}^{\pi / 2} \cos ^{d-i-1} \theta_{i} d \theta_{i}\right)\left(\int_{0}^{2 \pi} 1 d \theta_{d-1}\right)
$$

and $\int_{-\pi / 2}^{\pi / 2} \cos ^{d-i-1} \theta d \theta=\pi^{1 / 2} \frac{\Gamma((d-i) / 2)}{\Gamma((d-i+1) / 2)}$.


[^0]:    Email addresses: njiandan@ufl.edu (Ruoyang Zhang), ghoshm@ufl.edu (Malay Ghosh)

