Electronic Journal of Statistics
Vol. 18 (2024) 5198–5257
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2326

Spectral norm posterior contraction in Bayesian sparse spiked covariance matrix model

Fangzheng Xie[1]

[1] *Department of Statistics, Indiana University,*
e-mail: fxie@iu.edu

Abstract: Posterior contraction rates with regard to non-intrinsic metrics have been a long-standing challenge in the Bayesian analysis of high-dimensional models. This paper establishes the minimax-optimal posterior contraction rates of the Bayesian sparse spiked covariance matrix model under the spectral norm, a non-intrinsic metric for the Gaussian covariance matrix model. Our proof technique relies on the recent advance in the geometric properties of Euclidean representation for subspaces and low-rank matrices, a local asymptotic normality argument, and the distributional approximation to the asymptotic posterior distribution of the sparse spiked covariance matrix.

MSC2020 subject classifications: Primary 62C10, 62F15; secondary 62H25.
Keywords and phrases: Cayley parameterization, non-intrinsic metric, posterior contraction, sparse spiked covariance matrix, spectral norm.

Received November 2023.

Contents

1. Introduction

1.1. Background

Estimating covariance structures in the presence of high-dimensional variables has attracted attention in the past decades. High-dimensional covariance matrix estimation naturally connects to latent factor models [20, 26], principal component analysis [44], graphical models [43], discriminant analysis [18], and clustering analysis [12, 15, 49, 60], among others. It also finds applications in a broad spectrum of real-world problems, including genomics [37], computer vision [23, 33], and econometrics [17], to name a selected few. There has also been substantial development in high-dimensional structured covariance/precision matrix estimation. Examples of structured covariance matrices include spiked(low-rank) covariance matrices [7, 29, 31, 32], sparse covariance matrices [10], and sparse precision matrices [19]. The readers are referred to the survey paper [9] and the reference therein.

1.2. Main contribution

Consider the spiked covariance matrix estimation problem [8, 14, 30, 42] where independent observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are generated from $\mathrm{N}_p(\mathbf{0}_p, \mathbf{\Omega})$, where $\mathbf{0}_p$ is the p-dimensional zero vector, $\mathbf{\Omega}$ has the spiked form

$$\mathbf{\Omega} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}} + \mathbf{I}_p, \tag{1}$$

$\mathbf{V}$ is a $p \times r$ matrix with orthonormal vectors, and $\mathbf{\Lambda}$ is a diagonal matrix with diagonals being $\lambda_1 \geq \ldots \geq \lambda_r > 0$, and $r \ll p$. We focus on the case where the leading eigenvector matrix $\mathbf{V}$ exhibits the row sparsity, namely, only s rows of $\mathbf{V}$ are nonzero and $s \ll p$.

This work focuses on studying the rate-optimality of the posterior contraction under the above sampling model with a certain hierarchical prior distribution for the sparse spiked covariance matrix (1). Leveraging the Cayley parameterization in [28, 47] and the accompanying technical tools developed in [55], we constructed a hierarchical prior model for (1), and obtain the minimax-optimal posterior contraction rate in the spectral norm. This is a nontrivial result and does not follow directly from the so-called prior-concentration-and-testing framework pioneered by [24, 25] because the spectral norm is not an intrinsic metric (*i.e.*, not equivalent to the Fisher information metric of the Gaussian covariance model). It has been observed in [27] that the framework of [24] does not apply directly to non-intrinsic metrics. Leveraging the non-singularity of the Fréchet derivative (Jacobian matrix) of the Cayley parameterization studied in [55], we establish the local asymptotic normality under the Gaussian spiked covariance model and derive the shape of the asymptotic posterior distribution, which leads to the minimax-optimal posterior contraction with regard to the spectral norm.

1.3. Related work

Due to the structural convenience, the spiked covariance model has been used as a natural probabilistic model for principal component analysis (PCA). In the high-dimensional regime where the model dimension p far exceeds the number of samples n, the authors of [30] showed that the classical PCA might lead to inconsistent estimators and certain structural assumptions are needed, e.g., sparse structures [8, 30] or effective rank constraints [31, 32]. Correspondingly, the sparse structure of $\mathbf{V}$ motivates the development of sparse PCA methods. For an incomplete list of works related to the sparse spiked covariance model and sparse PCA, see [3, 5, 7, 8, 30, 34, 35, 52, 53, 59] and the reference therein. In the case where $p/n \to \gamma$ for some constant $\gamma \in (0, 1]$, the authors of [14, 42] studied the asymptotics of the sample eigenstructure. In [31, 32], the so-called effective rank assumption is made, in the sense that $(\sum_{k=1}^{r} \lambda_k + p)/(\lambda_1 + 1) = o(n)$, which in turn requires that λ_1 needs to diverge to ∞ sufficiently fast (recall that $\lambda_1 \geq \ldots \geq \lambda_r > 0$ are the eigenvalues in $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$ appearing in model (1)).

When the parameter of interest is the principal subspace $\mathrm{Span}(\mathbf{V})$ (*i.e.*, the column space of $\mathbf{V}$) and the eigenvalues of $\mathbf{\Omega}$ are bounded away from 0 and ∞, the authors of [53] established the minimax rate $\sqrt{(rs + s\log p)/n}$ under the so-called Frobenius sine-theta distance. Furthermore, the authors of [7] derived the minimax rate $\sqrt{(s\log p)/n}$ for the principle subspace under the spectral sine-theta distance. We defer the formal definitions of the sine-theta distances to Section 2.1. The minimax rate with regard to the spectral sine-theta distance is sharper than that with regard to the Frobenius sine-theta distance when $r \gg \log p$.

In contrast to the well-developed theory from the frequentist side, the literature regarding Bayesian approaches for sparse spiked covariance model is slightly underexplored. In [41], the authors first studied the minimax-optimal posterior contraction of $\mathbf{\Omega}$ with sparse priors, assuming the rank r is bounded. Under a more general assumption that the number of nonzero rows of $\mathbf{V}$ is no greater than rs, the authors of [21] established the rate-optimal posterior contraction of Bayesian sparse PCA under the Frobenius sine-theta distance. Recently, the authors of [39, 56] focused on the posterior contraction rate under the spectral norm, assuming that $r\log n \lesssim \log p$. Under the same low-rank assumption, in [38, 39], the author studied the contraction rate under the spectral norm for both the exact posterior distribution and the variational distribution and proposed computationally efficient algorithms for Bayesian inference with sparse priors. We remark that the assumption $r\log n \lesssim \log p$ greatly simplifies the problem of posterior contraction in spectral norm since the minimax rate $\sqrt{(s\log p)/n}$ under the spectral sine-theta distance coincides with the minimax rate $\sqrt{(rs + s\log p)/n}$ under the Frobenius sine-theta distance. Namely, rate-optimal posterior contraction under the Frobenius $\sin\Theta$ distance directly implies rate-optimal posterior contraction under the spectral $\sin\Theta$ distance when $r\log n \lesssim \log p$. In this work, we are particularly interested in the posterior contraction under the spectral sine-theta distance between principal subspaces when $r \gg \log p$, in which the phase transition phenomenon between the two minimax rates occurs.

There have been several works in the literature that deal with the parameterization of subspaces in the Grassmannian using Euclidean vectors, including [1, 2, 4, 16, 45, 54]. For example, in [1], the authors proposed the cross-section mapping and derived the canonical metric. The cross-section mapping induces a parameterization of subspaces that has also been discussed in [2, 4, 16, 45, 54]. In our Bayesian sparse spiked covariance matrix estimation context, these parameterization methods may lead to extra complications when sparsity needs to be enforced over the rows of $\mathbf{V}$, whereas it is easier to deal with sparsity via the Cayley parameterization.

1.4. Organization

We provide the necessary notations and definitions about this work and briefly review the necessary ingredients of Cayley parameterization of subspaces and Euclidean representation of low-rank matrices in Section 2. Section 3 elaborates on the prior specification for the sparse spiked covariance matrix model, establishes the spectral norm posterior contraction, and provides the proof sketch of the main result. Additional discussion is provided in Section 4. The technical proofs are deferred to Section 5.

2. Preliminaries

2.1. Notations and definitions

We use the symbol $:=$ to assign mathematical definitions of quantities. For $a, b \in \mathbb{R}$, let $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$. For a positive integer p, let $[p] := \{1, \ldots, p\}$. For two non-negative sequences $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$, we use the symbol $a_n \lesssim b_n$ or $a_n = O(b_n)$ ($a_n \gtrsim b_n$, resp.) to mean that $a_n \leq Cb_n$ ($a_n \geq Cb_n$, resp.) for some constant $C > 0$, and we use the notation $a_n \asymp b_n$ to indicate that $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. The notation $\mathbf{A}^\dagger$ denotes the Moore-Penrose pseudoinverse of an arbitrary matrix $\mathbf{A}$. For a general matrix $\mathbf{A}$ with p rows, we define the support of $\mathbf{A}$ as $\mathrm{supp}(\mathbf{A}) := \{j \in [p] : [\mathbf{A}]_{j*} \neq \mathbf{0}\}$, where $[\mathbf{A}]_{j*}$ denotes the jth row of $\mathbf{A}$. We use $C, C_0, C_1, C_2, c, c_0, c_1, c_2, \ldots$ to denote generic constants that may change from line to line unless otherwise stated. The $r \times r$ identity matrix is denoted by $\mathbf{I}_r$, the p-dimensional zero vector is denoted by $\mathbf{0}_p$, and $\mathbf{0}_{p\times q}$ denotes the $p \times q$ zero matrix. We reserve the symbol $\mathbf{I}(\cdot)$ without a subscript for the Fisher information matrix of a (regular) statistical model, and it should not be confused with the identity matrix. A $p \times r$ Stiefel matrix $\mathbf{U}$ (with $r \leq n$) is a $p \times r$ matrix satisfying $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_r$. Given two positive integers p, r, we denote by $\mathbb{O}(p, r) := \{\mathbf{U} \in \mathbb{R}^{p\times r} : \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_r\}$ the collection of all $p \times r$ Stiefel matrices and write $\mathbb{O}(r) := \mathbb{O}(r, r)$. For any $\mathbf{U} \in \mathbb{O}(p, r)$, we use $\mathrm{Span}(\mathbf{U})$ to denote the r-dimensional subspace in $\mathbb{R}^p$ spanned by the columns of $\mathbf{U}$. The collection of all $r \times r$ symmetric matrices is denoted by $\mathbb{M}(r)$, and the collection of all $r \times r$ symmetric positive definite matrices is denoted by $\mathbb{M}_+(r)$. For a matrix $\mathbf{\Sigma} \in \mathbb{R}^{p_1\times p_2}$ and indices $i \in [p_1], j \in [p_2]$, let $[\mathbf{\Sigma}]_{ij}$ denote the element on the ith row and jth column of $\mathbf{\Sigma}$, $[\mathbf{\Sigma}]_{i*}$ denote the ith row of $\mathbf{\Sigma}$, and $[\mathbf{\Sigma}]_{j*}$ denote the jth column of $\mathbf{\Sigma}$. Furthermore, we use $\sigma_1(\mathbf{\Sigma}), \ldots, \sigma_{p_1\wedge p_2}(\mathbf{\Sigma})$ to denote the singular values of $\mathbf{\Sigma}$ sorted in the non-increasing order, i.e., $\sigma_1(\mathbf{\Sigma}) \geq \ldots \geq \sigma_{p_1\wedge p_2}(\mathbf{\Sigma})$. When $\mathbf{\Sigma}$ is a $p \times p$ symmetric square matrix, $\lambda_1(\mathbf{\Sigma}), \ldots, \lambda_p(\mathbf{\Sigma})$ denote the eigenvalues of $\mathbf{\Sigma}$ sorted in the non-increasing order in magnitude, namely, $|\lambda_1(\mathbf{\Sigma})| \geq \ldots \geq |\lambda_p(\mathbf{\Sigma})|$. The spectral norm of a general matrix $\mathbf{\Sigma}$, denoted by $\|\mathbf{\Sigma}\|_2$, is the largest singular value of $\mathbf{\Sigma}$, and the Frobenius norm of $\mathbf{\Sigma}$, denoted by $\|\mathbf{\Sigma}\|_{\mathrm{F}}$, is defined to be $\|\mathbf{\Sigma}\|_{\mathrm{F}} := (\sum_{i=1}^{p_1}\sum_{j=1}^{p_2}[\mathbf{\Sigma}]_{ij}^2)^{1/2}$. For a Euclidean vector $\mathbf{x} = [x_1, \ldots, x_p]^{\mathrm{T}} \in \mathbb{R}^p$, we denote $[\mathbf{x}]_i := x_i$, $\|\mathbf{x}\|_2$ the usual Euclidean norm $\|\mathbf{x}\|_2 = (\sum_i x_i^2)^{1/2}$, let $B_2(\mathbf{x}, \epsilon) := \{\mathbf{y} \in \mathbb{R}^p : \|\mathbf{y} - \mathbf{x}\|_2 < \epsilon\}$, and let $\mathrm{diag}(\mathbf{x})$ be the $p \times p$ diagonal matrix with x_i being the element on its ith row and ith column.

For a $p_1 \times p_2$ matrix $\mathbf{M}$, the operator $\mathrm{vec}(\cdot)$ converts $\mathbf{M}$ to a p_1p_2-dimensional Euclidean vector by stacking the columns of $\mathbf{\Sigma}$ consecutively, i.e.,

$$\begin{aligned}\mathrm{vec}(\mathbf{M}) &:= [[\mathbf{M}]_{*1}^{\mathrm{T}}, [\mathbf{M}]_{*2}^{\mathrm{T}}, \ldots, [\mathbf{M}]_{*p_2}^{\mathrm{T}}]^{\mathrm{T}} \\ &= [[\mathbf{M}]_{11}, \ldots, [\mathbf{M}]_{p_11}, [\mathbf{M}]_{12}, \ldots, [\mathbf{M}]_{p_12}, \ldots, [\mathbf{M}]_{1p_2}, \ldots, [\mathbf{M}]_{p_1p_2}]^{\mathrm{T}}.\end{aligned}$$

The operator $\mathrm{vech}(\cdot)$ transforms an $r \times r$ square symmetric matrix $\mathbf{M}$ to an $r(r+1)/2$-dimensional Euclidean vector by eliminating all its super-diagonal elements, i.e.,

$$\mathrm{vech}(\mathbf{M}) := [[\mathbf{M}]_{11}, [\mathbf{M}]_{21}, \ldots, [\mathbf{M}]_{r1}, [\mathbf{M}]_{22}, \ldots, [\mathbf{M}]_{r2}, \ldots, [\mathbf{M}]_{rr}]^{\mathrm{T}}.$$

For any two positive integers p, q, we denote $\mathbf{K}_{pq}$ the $pq\times pq$ commutation matrix such that $\mathrm{vec}(\mathbf{M}^{\mathrm{T}}) = \mathbf{K}_{pq}\mathrm{vec}(\mathbf{M})$ for any $\mathbf{M} \in \mathbb{R}^{p\times q}$, and denote $\mathbb{D}_p$ the duplication matrix such that $\mathrm{vec}(\mathbf{M}) = \mathbb{D}_p\mathrm{vech}(\mathbf{M})$ for any symmetric $\mathbf{M} \in \mathbb{R}^{p\times p}$. We refer the readers to [36] for a review of the properties of the commutation matrix $\mathbf{K}_{pq}$ and the duplication matrix $\mathbb{D}_p$. For two matrices $\mathbf{A} \in \mathbb{R}^{p\times q}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$, we use $\mathbf{A} \otimes \mathbf{B}$ to denote the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$, defined to be the $pm \times qn$ matrix of the form

$$\mathbf{A} \otimes \mathbf{B} := \begin{bmatrix}[\mathbf{A}]_{11}\mathbf{B} & [\mathbf{A}]_{12}\mathbf{B} & \ldots & [\mathbf{A}]_{1q}\mathbf{B} \\ [\mathbf{A}]_{21}\mathbf{B} & [\mathbf{A}]_{22}\mathbf{B} & \ldots & [\mathbf{A}]_{2q}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ [\mathbf{A}]_{p1}\mathbf{B} & [\mathbf{A}]_{p2}\mathbf{B} & \ldots & [\mathbf{A}]_{pq}\mathbf{B}\end{bmatrix}.$$

The distance between linear subspaces can be measured in terms of the canonical angles, formally defined as follows. Given two Stiefel matrices $\mathbf{U}, \mathbf{U}_0 \in \mathbb{O}(p, r)$, let $\sigma_1(\mathbf{U}_0^{\mathrm{T}}\mathbf{U}) \geq \ldots \geq \sigma_r(\mathbf{U}_0^{\mathrm{T}}\mathbf{U}) \geq 0$ be the singular values of $\mathbf{U}_0^{\mathrm{T}}\mathbf{U}$. Note the singular values of $\mathbf{U}_0^{\mathrm{T}}\mathbf{U}$ are unitarily invariant and only depend on $\mathrm{Span}(\mathbf{U})$ and $\mathrm{Span}(\mathbf{U}_0)$. The canonical angles between $\mathbf{U}_0$ and $\mathbf{U}$ are defined to be the diagonal entries of

$$\Theta(\mathbf{U}_0, \mathbf{U}) := \mathrm{diag}\left[\cos^{-1}\left\{\sigma_1(\mathbf{U}_0^{\mathrm{T}}\mathbf{U})\right\}, \ldots, \cos^{-1}\left\{\sigma_r(\mathbf{U}_0^{\mathrm{T}}\mathbf{U})\right\}\right] \in \mathbb{R}^{r\times r}.$$

Then, the spectral sine-theta distance and the Frobenius sine-theta distance between two subspaces $\mathrm{Span}(\mathbf{U}_0)$ and $\mathrm{Span}(\mathbf{U})$ are defined by $\|\sin\Theta(\mathbf{U}_0, \mathbf{U})\|_2$ and $\|\sin\Theta(\mathbf{U}_0, \mathbf{U})\|_{\mathrm{F}}$, respectively.

2.2. Review of Cayley parameterization and Euclidean representation of low-rank matrices

Cayley transform of subspaces [28, 47] and Euclidean representation of low-rank matrices [55], which we briefly review here, are the key tools for us to derive rate-optimal spectral norm posterior contraction in Bayesian sparse spiked covariance matrix model. We first provide the heuristics of Cayley parameterization and Euclidean representation of low-rank matrices before diving into the formal definitions. Let $\mathcal{G}(p, r)$ denote the Grassmannian, *i.e.*, the collection of all r-dimensional linear subspaces in $\mathbb{R}^p$ equipped with the canonical metric (see [1] for details). On a high level, Cayley parameterization constructs a diffeomorphism (a differentiable one-to-one function), denoted by $\boldsymbol{\varphi} \mapsto \mathrm{Span}\{\mathbf{U}(\boldsymbol{\varphi})\}$, between an open subset of $\mathbb{R}^{(p-r)r}$ and $\mathcal{G}(p, r)$, such that the intrinsic dimension of $\mathcal{G}(p, r)$ can be fully captured by low-dimensional Euclidean vectors. Cayley parameterization of subspaces further leads to the Euclidean representation of low-rank matrices developed in [58]. A crucial result there is that the smallest singular value of the Jacobian matrix of this Euclidean representation function is strictly positive, implying the non-singularity of the Fisher information matrix of the spiked covariance model under such an Euclidean representation. It turns out that the non-singularity of the Fisher information matrix is the key to our distributional approximation to the posterior distribution result, which is of fundamental interest for obtaining the rate-optimal posterior contraction of the principal subspace under the spectral sine-theta distance.

We now formally introduce the Cayley parameterization and the Euclidean representation of low-rank matrices. Suppose $\mathbb{S} \subset \mathbb{R}^p$ is an r-dimensional linear subspace in the p-dimensional Euclidean space. In [28], the authors show that, with respect to the uniform probability distribution over $\mathcal{G}(p, r)$, almost every r-dimensional subspace in $\mathbb{R}^p$ can be represented as the column span of a unique orthonormal matrix $\mathbf{U} \in \mathbb{O}_+(p, r)$, where

$$\mathbb{O}_+(p, r) := \{\mathbf{U} = [\mathbf{Q}_1^{\mathrm{T}}\ \mathbf{Q}_2^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{O}(p, r) : \mathbf{Q}_1 \in \mathbb{M}_+(r)\}.$$

Let $\mathbf{A} \in \mathbb{R}^{(p-r)\times r}$ with $\|\mathbf{A}\|_2 < 1$, and let $\boldsymbol{\varphi} := \mathrm{vec}(\mathbf{A})$. Then, the Cayley parameterization [28, 47] is defined as the function

$$\mathbf{U} : \boldsymbol{\varphi} \in \mathbb{R}^{(p-r)r} \mapsto \mathbf{U}(\boldsymbol{\varphi}) := (\mathbf{I}_p + \mathbf{X}_{\boldsymbol{\varphi}})(\mathbf{I}_p - \mathbf{X}_{\boldsymbol{\varphi}})^{-1}\mathbf{I}_{p\times r} \in \mathbb{O}_+(p,r), \tag{2}$$

where

$$\mathbf{X}_{\boldsymbol{\varphi}} := \begin{bmatrix} \mathbf{0}_{r\times r} & -\mathbf{A}^{\mathrm{T}} \\ \mathbf{A} & \mathbf{0}_{(p-r)\times(p-r)} \end{bmatrix}, \quad \text{and} \quad \mathbf{I}_{p\times r} := \begin{bmatrix} \mathbf{I}_r \\ \mathbf{0}_{(p-r)\times r} \end{bmatrix}. \tag{3}$$

Equivalently, $\mathbf{U}(\boldsymbol{\varphi})$ can be written in the following block form:

$$\mathbf{U} = \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{bmatrix} = \begin{bmatrix} (\mathbf{I}_r - \mathbf{A}^{\mathrm{T}}\mathbf{A})(\mathbf{I}_r + \mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1} \\ 2\mathbf{A}(\mathbf{I}_r + \mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1} \end{bmatrix}.$$

In [28], the authors show that the Cayley parameterization $\mathbf{U}(\cdot)$ has the following two properties: It is a bijective function onto $\mathbb{O}_+$ with inverse given by $\mathbf{U}(\boldsymbol{\varphi}) \mapsto \mathrm{vec}\{\mathbf{Q}_2(\mathbf{I}_r + \mathbf{Q}_1)^{-1}\}$, and it is a differentiable function whose Fréchet derivative is

$$D\mathbf{U}(\boldsymbol{\varphi}) = 2[\mathbf{I}_{p\times r}^{\mathrm{T}}(\mathbf{I}_p - \mathbf{X}_{\boldsymbol{\varphi}})^{-\mathrm{T}} \otimes (\mathbf{I}_p - \mathbf{X}_{\boldsymbol{\varphi}})^{-1}]\boldsymbol{\Gamma}_{\boldsymbol{\varphi}}, \tag{4}$$

where $\boldsymbol{\Gamma}_{\boldsymbol{\varphi}} := (\mathbf{I}_{p^2} - \mathbf{K}_{pp})(\boldsymbol{\Theta}_1^{\mathrm{T}} \otimes \boldsymbol{\Theta}_2^{\mathrm{T}})$, $\boldsymbol{\Theta}_1 := \mathbf{I}_{p\times r}^{\mathrm{T}}$, and $\boldsymbol{\Theta}_2 := [\mathbf{0}_{(p-r)\times r}, \mathbf{I}_{p-r}]$.

Leveraging the Cayley parameterization of subspaces, the author of [55] establishes a Euclidean representation framework for symmetric low-rank matrices as follows. Specifically, let $\boldsymbol{\mu} := \mathrm{vech}(\mathbf{M})$ and denote by $\mathbf{M}(\boldsymbol{\mu}) := \mathbf{M}$ the inverse of the function $\mathbf{M} \mapsto \mathrm{vech}(\mathbf{M})$. Then, with $\boldsymbol{\theta} := [\boldsymbol{\varphi}^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{T}}]^{\mathrm{T}}$, the Euclidean representation of low-rank matrices is defined as the following function

$$\boldsymbol{\Sigma}(\cdot) : \mathscr{D}(p,r) \to \mathscr{S}(p,r), \quad \boldsymbol{\theta} \mapsto \mathbf{U}(\boldsymbol{\varphi})\mathbf{M}(\boldsymbol{\mu})\mathbf{U}(\boldsymbol{\varphi})^{\mathrm{T}}, \tag{5}$$

where

$$\mathscr{D}(p,r) := \left\{ \boldsymbol{\theta} = \begin{bmatrix} \mathrm{vec}(\mathbf{A}) \\ \boldsymbol{\mu} \end{bmatrix} \in \mathbb{R}^{(p-r)r} \times \mathbb{R}^{r(r+1)/2} : \|\mathbf{A}\|_2 < 1 \right\}, \tag{6}$$

$$\mathscr{S}(p,r) := \{\boldsymbol{\Sigma} = \mathbf{U}\mathbf{M}\mathbf{U}^{\mathrm{T}} : \mathbf{U} \in \mathbb{O}_+(p,r), \mathbf{M} \in \mathbb{M}(r)\}. \tag{7}$$

Conversely, for any symmetric positive semidefinite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ with $\mathrm{rank}(\boldsymbol{\Sigma}) = r \le p$, it yields spectral decomposition $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{V} \in \mathbb{O}(p,r)$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$ with $\lambda_1 \ge \ldots \ge \lambda_r > 0$. Then, for almost every $\mathrm{Span}(\mathbf{V}) \in \mathcal{G}(p,r)$, there exists $\mathbf{U} \in \mathbb{O}_+(p,r)$ such that $\mathrm{Span}(\mathbf{V}) = \mathrm{Span}(\mathbf{U})$, which entails that $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{M}\mathbf{U}^{\mathrm{T}}$ for some $r\times r$ symmetric positive definite matrix $\mathbf{M}$. By the one-to-one property of $\mathbf{U}(\cdot)$, there exists a unique $\mathbf{A} \in \mathbb{R}^{(p-r)\times r}$, $\|\mathbf{A}\|_2 < 1$, such that $\mathbf{U} = \mathbf{U}(\boldsymbol{\varphi})$, where $\boldsymbol{\varphi} = \mathrm{vec}(\mathbf{A})$. This shows that almost every $p\times p$ rank-r matrix is in the range of $\boldsymbol{\Sigma}(\cdot)$, which is the collection of low-rank matrices $\mathscr{S}(p,r)$ defined in (7).

In [55], the author shows that the function $\boldsymbol{\Sigma}(\cdot)$ is also differentiable with the Fréchet derivative $D\boldsymbol{\Sigma}(\boldsymbol{\theta}) = [D_{\boldsymbol{\varphi}}\boldsymbol{\Sigma}(\boldsymbol{\theta}), D_{\boldsymbol{\mu}}\boldsymbol{\Sigma}(\boldsymbol{\theta})]$, where

$$\begin{aligned} D_{\boldsymbol{\varphi}}\boldsymbol{\Sigma}(\boldsymbol{\theta}) &:= (\mathbf{I}_{p^2} + \mathbf{K}_{pp})\{\mathbf{U}(\boldsymbol{\varphi})\mathbf{M}(\boldsymbol{\mu}) \otimes \mathbf{I}_p\}D\mathbf{U}(\boldsymbol{\varphi}), \\ D_{\boldsymbol{\mu}}\boldsymbol{\Sigma}(\boldsymbol{\theta}) &:= \{\mathbf{U}(\boldsymbol{\varphi}) \otimes \mathbf{U}(\boldsymbol{\varphi})\}\mathbb{D}_r, \end{aligned} \tag{8}$$

where $D\mathbf{U}(\boldsymbol{\varphi})$ is defined in (4). In what follows, $\boldsymbol{\Sigma}$ plays the role of the low-rank component $\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$ in the spiked covariance model (1). For $\boldsymbol{\Sigma}_0$ being the true value of $\boldsymbol{\Sigma}$, we denote by $\boldsymbol{\theta}_0 := [\boldsymbol{\varphi}_0^{\mathrm{T}}, \boldsymbol{\mu}_0^{\mathrm{T}}]^{\mathrm{T}}$ be the vector such that $\boldsymbol{\Sigma}_0 := \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = \mathbf{U}(\boldsymbol{\varphi}_0)\mathbf{M}(\boldsymbol{\mu}_0)\mathbf{U}(\boldsymbol{\varphi}_0)^{\mathrm{T}}$, in which case we let $\mathbf{U}_0 := \mathbf{U}(\boldsymbol{\varphi}_0)$ and $\mathbf{M}_0 := \mathbf{M}(\boldsymbol{\mu}_0)$. The key technical result in [55] is the following lower bound on the smallest singular value of the Fréchet derivative matrix $D_{\boldsymbol{\mu}}\boldsymbol{\Sigma}(\cdot)$.

Theorem 2.1. *Under the above setup, if $r \geq 2$, then*

$$\sigma_{\min}\{D_{\boldsymbol{\varphi}}\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\} \geq \frac{2\sqrt{2}\sigma_r(\mathbf{M}_0)(1-\|\mathbf{A}_0\|_2^2)}{(1+\|\mathbf{A}_0\|_2^2)^2},$$

$$\|\{D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\}^{-1}\|_2 \leq 1 + \frac{(1+64\|\mathbf{M}_0\|_2^2)(1+\|\mathbf{A}_0\|_2^2)^4}{8\lambda_r^2(\mathbf{M}_0)(1-\|\mathbf{A}_0\|_2^2)^2}.$$

Remark 1. Theorem 2.1 above guarantees that the Fisher information matrix of the spiked covariance model (1) is non-degenerate. Formally, let $\mathscr{C}(p)$ be a collection of $p \times p$ symmetric positive definite matrices and $\mathcal{P} := \{p_{\boldsymbol{\Omega}}(\mathbf{y}) : \boldsymbol{\Omega} \in \mathscr{C}(p)\}$, where $p_{\boldsymbol{\Omega}}(\mathbf{x}) := \det(2\pi\boldsymbol{\Omega})^{-1/2}e^{-(1/2)\mathbf{x}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{x}}$ denotes the density function of $\mathrm{N}_p(\mathbf{0}_p, \boldsymbol{\Omega})$. Denote by $\dot{\boldsymbol{\ell}}_{\boldsymbol{\Omega}}(\mathbf{x}_i) := \nabla_{\mathrm{vech}(\boldsymbol{\Omega})}\log p_{\boldsymbol{\Omega}}(\mathbf{x}_i)$ the score function with regard to $\mathrm{vech}(\boldsymbol{\Omega})$, and $\mathbf{I}(\boldsymbol{\Omega}) := \mathbb{E}\dot{\boldsymbol{\ell}}_{\boldsymbol{\Omega}}(\mathbf{x})\dot{\boldsymbol{\ell}}_{\boldsymbol{\Omega}}(\mathbf{x})^{\mathrm{T}}$ the corresponding Fisher information matrix. It can be shown that (see, e.g., Chapter 10 in [36])

$$\mathbf{I}(\boldsymbol{\Omega}_0) = \frac{1}{2}\mathbb{D}_p^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1})\mathbb{D}_p.$$

However, if one instead restricts $\boldsymbol{\Omega}$ to the spiked matrix class

$$\mathscr{C}(p,r) := \{\boldsymbol{\Omega} := \boldsymbol{\Sigma} + \mathbf{I}_p : \boldsymbol{\Sigma} \in \mathscr{S}(p,r)\}$$

and assumes $\boldsymbol{\Omega}_0 = \boldsymbol{\Sigma}_0 + \mathbf{I}_p$ for some $\boldsymbol{\Sigma}_0 \in \mathscr{S}(p,r)$, then the spiked matrix $\boldsymbol{\Omega}$ can be represented by a lower dimensional Euclidean vector $\boldsymbol{\theta} \in \mathscr{D}(p,r)$. Thus, the statistical submodel under the $\boldsymbol{\theta}$-parameterization can be written as

$$\mathscr{F}(p,r) = \{p_{\boldsymbol{\Omega}(\boldsymbol{\theta})}(\cdot) : \boldsymbol{\Omega}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}) + \mathbf{I}_p, \boldsymbol{\theta} \in \mathscr{D}(p,r)\}.$$

Denote by $\boldsymbol{\theta}_0 \in \mathscr{D}(p,r)$ the vector such that $\boldsymbol{\Omega}_0 = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) + \mathbf{I}_p$. By the chain rule, the Fisher information matrix with respect to the $\boldsymbol{\theta}$-parameterization in the submodel $\mathscr{F}(p,r)$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is given by

$$\mathbf{I}(\boldsymbol{\theta}_0) = D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}(\mathbb{D}_p^{\dagger})^{\mathrm{T}}\mathbb{D}_p^{\mathrm{T}}\boldsymbol{\Psi}(\boldsymbol{\Omega}_0)\mathbb{D}_p\mathbb{D}_p^{\dagger}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0).$$

Using the fact that $\mathbb{D}_p\mathbb{D}_p^{\dagger} = (1/2)(\mathbf{I}_{p^2} + \mathbf{K}_{pp})$, $(\mathbb{D}_p\mathbb{D}_p^{\dagger})(\mathbf{I}_{p^2} + \mathbf{K}_{pp})^2 = (\mathbf{I}_{p^2} + \mathbf{K}_{pp})$, and $\mathbb{D}_p\mathbb{D}_p^{\dagger}(\mathbf{U}_0 \otimes \mathbf{U}_0)\mathbb{D}_p = (\mathbf{U}_0 \otimes \mathbf{U}_0)\mathbb{D}_p$ (see, e.g., Chapter 4 in [36]), we further conclude that

$$\mathbf{I}(\boldsymbol{\theta}_0) = \mathbf{I}(\boldsymbol{\theta}_0) = \frac{1}{2}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0).$$

Hence, by Theorem 2.1, the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta}_0)$ with respect to the $\boldsymbol{\theta}$-parameterization in the submodel $\mathscr{F}(p,r)$ is also non-degenerate. This observation will be indispensable in investigating the spectral norm posterior contraction in the Bayesian sparse spiked covariance model in Section 3 below.

2.3. Review of minimax rates for sparse spiked covariance matrices

In this subsection, we take a moment to formally review the minimax estimation results for $\mathrm{Span}(\mathbf{V})$ established in [7, 53] before diving into the proposed hierarchical Bayesian model and the posterior contraction results. In [53], the authors showed that, when $\lambda_r(\mathbf{\Omega}) - \lambda_{r+1}(\mathbf{\Omega})$ is bounded away from 0 and $\mathbf{V} \in \mathbb{O}(p, r)$ is the eigenvector matrix of $\mathbf{\Omega}$ corresponding to $\lambda_1(\mathbf{\Omega}), \dots, \lambda_r(\mathbf{\Omega})$, the minimax rate for estimating $\mathrm{Span}(\mathbf{V})$ under the Frobenius sine-theta distance is given by

$$\inf_{\widehat{\mathbf{V}}} \sup_{\mathbf{V} \in \mathbb{O}(p,r), |\mathrm{supp}(\mathbf{V})| \leq s} \mathbb{E}_{\mathbf{V}} \{ \| \sin \Theta(\widehat{\mathbf{V}}, \mathbf{V}) \|_{\mathrm{F}}^2 \} \asymp \frac{rs + s \log p}{n}. \tag{9}$$

Here, $\inf_{\widehat{\mathbf{V}}}$ denotes the infimum over all possible estimators $\widehat{\mathbf{V}}$ (measurable functions of $\mathbf{y}_1, \dots, \mathbf{y}_n$). In [7], the authors further showed that the minimax rate for estimating $\mathrm{Span}(\mathbf{V})$ under the spectral sine-theta distance is given by

$$\inf_{\widehat{\mathbf{V}}} \sup_{\mathbf{\Omega} \in \Theta_0(s,p,r,\lambda,\tau)} \mathbb{E}_{\mathbf{\Omega}} \{ \| \sin \Theta(\widehat{\mathbf{V}}, \mathbf{V}) \|_2^2 \} \asymp \frac{s \log p}{n}, \tag{10}$$

where

$$\begin{aligned} \Theta_0(s, p, r, \lambda, \tau) := \{ \mathbf{\Omega} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^{\mathrm{T}} + \mathbf{I}_p : \mathbf{V} \in \mathbb{O}(p, r), |\mathrm{supp}(\mathbf{V})| \leq s, \\ \lambda / \tau \leq \lambda_r(\mathbf{\Lambda}) \leq \lambda_1(\mathbf{\Lambda}) \leq \lambda \}, \end{aligned}$$

and λ, τ are bounded away from 0 and ∞. Note that the minimax rate (9) under the Frobenius sine-theta distance has an extra term $\sqrt{rs/n}$ compared to (10) when $r \gg \log p$. Below, we will design a hierarchical prior model for $\mathbf{\Omega}$ by leveraging the Cayley parameterization and show that the posterior contraction rate with regard to the spectral sine-theta distance is exactly $\sqrt{(s \log p)/n}$.

3. Bayesian sparse spiked covariance model

3.1. Prior specification

Recall that in the spiked covariance model (1), the leading eigenvector matrix $\mathbf{V}$ can only be identified up to an orthogonal matrix in $\mathbb{O}(r)$ in the presence of eigenvalue multiplicity. Because, for any covariance matrix $\mathbf{\Omega}$ of the form (1), there exists some permutation matrix $\mathbf{\Pi}$ such that $\mathbf{I}_{p \times r}^{\mathrm{T}}(\mathbf{\Pi} \mathbf{V})$ is non-singular, and the leading eigenvector matrix of $\mathbf{\Pi} \mathbf{\Omega} \mathbf{\Pi}^{\mathrm{T}}$ is exactly $\mathbf{\Pi} \mathbf{V}$. Therefore, without loss of generality, we assume that $\mathbf{I}_{p \times r}^{\mathrm{T}} \mathbf{V}$ is invertible. By the construction in Section 2.2, $\mathbf{\Omega}$ can be written as $\mathbf{\Omega} = \mathbf{\Sigma} + \mathbf{I}_p$ for some $\mathbf{\Sigma} \in \mathscr{S}(p, r)$, and there exist some $\mathbf{U} \in \mathbb{O}_+(p, r)$ and $\mathbf{M} \in \mathbb{M}_+(r)$, such that

$$\mathbf{\Omega} = \mathbf{\Sigma} + \mathbf{I}_p = \mathbf{U} \mathbf{M} \mathbf{U}^{\mathrm{T}} + \mathbf{I}_p.$$

We follow the setup and notations in Section 2.2. Since $\mathbf{\Sigma}$ can be parameterized by a Euclidean vector $\boldsymbol{\theta} = [\boldsymbol{\varphi}^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{T}}]^{\mathrm{T}} \in \mathscr{D}(p, r)$ through the map

$\boldsymbol{\Sigma}(\cdot) : \mathscr{D}(p,r) \to \mathscr{S}(p,r), \boldsymbol{\theta} \mapsto \boldsymbol{\Sigma}(\boldsymbol{\theta})$, we then use $\boldsymbol{\Omega}(\cdot)$ to generically denote the induced map $\boldsymbol{\theta} \mapsto \boldsymbol{\Omega}(\boldsymbol{\theta}) := \boldsymbol{\Sigma}(\boldsymbol{\theta}) + \mathbf{I}_p$. Furthermore, let $\boldsymbol{\Omega}_0$ denote the true value of the covariance $\boldsymbol{\Omega}$ corresponding to the distribution of $\mathbf{y}_1, \ldots, \mathbf{y}_n$, $\boldsymbol{\theta}_0 \in \mathscr{D}(p,r)$ be the inverse image of $\boldsymbol{\Omega}_0$ under the map $\boldsymbol{\Omega}(\cdot)$, and $\boldsymbol{\Sigma}_0 := \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$. Let $\boldsymbol{\varphi}, \boldsymbol{\varphi}_0 \in \mathbb{R}^{(p-r)r}$ be vectors and $\mathbf{A}, \mathbf{A}_0 \in \mathbb{R}^{(p-r)\times r}$ be matrices such that $\boldsymbol{\varphi} = \mathrm{vec}(\mathbf{A}), \boldsymbol{\varphi}_0 = \mathrm{vec}(\mathbf{A}_0)$, where $\|\mathbf{A}\|_2 < 1$, $\|\mathbf{A}_0\|_2 < 1$, and let $\mathbf{M}$, $\mathbf{M}_0 \in \mathbb{M}_+(r)$ be positive definite matrices such that $\boldsymbol{\mu} = \mathrm{vech}(\mathbf{M})$ and $\boldsymbol{\mu}_0 = \mathrm{vech}(\mathbf{M}_0)$, respectively.

The advantage of the Cayley parameterization is that the row sparsity of $\mathbf{U}(\boldsymbol{\varphi})$ can be directly incorporated into the rows of $\mathbf{A}$. By the construction of the Cayley parameterization, $\mathbf{U}(\boldsymbol{\varphi})$ can be written as

$$\mathbf{U}(\boldsymbol{\varphi}) = \begin{bmatrix} (\mathbf{I}_r - \mathbf{A}^{\mathrm{T}}\mathbf{A})(\mathbf{I}_r + \mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1} \\ 2\mathbf{A}(\mathbf{I}_r + \mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1} \end{bmatrix}.$$

It follows that $[\mathbf{U}(\boldsymbol{\varphi})]_{j*} = \mathbf{0}$ if and only if $[\mathbf{A}]_{(j-r)*}$ for any $j \in [p]\backslash\{r\}$. Furthermore, $\mathbf{U}(\boldsymbol{\varphi})$ is subject to the orthonormal constraint $\mathbf{U}(\boldsymbol{\varphi})^{\mathrm{T}}\mathbf{U}(\boldsymbol{\varphi}) = \mathbf{I}_r$, whereas working with $\mathbf{A}$ is more convenient. Hence, we consider the following sparsity-inducing prior distribution on $\mathbf{A}$. We assume throughout that the rank r of the spiked component is known. When r is unknown, our theory developed in this work can be applied by replacing r with a consistent estimator $\widehat{r}$ (*i.e.*, $\lim_{n\to\infty} \mathbb{P}_0(\widehat{r} = r) = 1$). Such a consistent estimator can be obtained, *e.g.*, using the approach proposed in [7]. Let π_p be the density of a discrete distribution supported on $\{0, 1, 2, \ldots, p-r\}$ of the form

$$\pi_p(t) = \frac{1}{z_n} n^{-rt}(p-r)^{-at}, \quad t = 0, \ldots, p-r \tag{11}$$

for some constants $a, c > 0$, where

$$z_n := \sum_{t=0}^{p-r} \left\{ \frac{1}{n^r(p-r)^a} \right\}^t = \frac{1 - \{n^{-r}(p-r)^{-a}\}^{p-r+1}}{1 - n^{-r}(p-r)^{-a}}$$

is the normalizing constant. Based on π_p, a subset $S \subset [p-r]$ representing the support of $\mathbf{A}$ is drawn from the following distribution:

$$\pi_S(S) := \frac{\pi_p(|S|)}{\binom{p-r}{|S|}}, \quad S \subset [p-r], \tag{12}$$

where $|S|$ denotes the cardinality of a finite set S. Given $S = \{j_1, j_2, \ldots, j_{|S|}\} \subset [p-r]$, suppose $S^c := [p-r]\backslash S$ can be written as $S^c = \{k_1, k_2, \ldots, k_{|S^c|}\}$ and denote by

$$\mathbf{A}_S := \begin{bmatrix} [\mathbf{A}]_{j_1 1} & [\mathbf{A}]_{j_1 2} & \ldots & [\mathbf{A}]_{j_1 r} \\ [\mathbf{A}]_{j_2 1} & [\mathbf{A}]_{j_2 2} & \ldots & [\mathbf{A}]_{j_2 r} \\ \vdots & \vdots & & \vdots \\ [\mathbf{A}]_{j_{|S|} 1} & [\mathbf{A}]_{j_{|S|} 2} & \ldots & [\mathbf{A}]_{j_{|S|} r} \end{bmatrix},$$

$$\mathbf{A}_{S^c} := \begin{bmatrix} [\mathbf{A}]_{k_1 1} & [\mathbf{A}]_{k_1 2} & \dots & [\mathbf{A}]_{k_1 r} \\ [\mathbf{A}]_{k_2 1} & [\mathbf{A}]_{k_2 2} & \dots & [\mathbf{A}]_{k_2 r} \\ \vdots & \vdots & & \vdots \\ [\mathbf{A}]_{k_{|S^c|} 1} & [\mathbf{A}]_{k_{|S^c|} 2} & \dots & [\mathbf{A}]_{k_{|S^c|} r} \end{bmatrix}.$$

We then define the prior distribution of $\mathbf{A}$ by

$$\begin{aligned} \Pi_{\mathbf{A}}(\mathrm{d}\mathbf{A}) &:= \sum_{S \subset [p-r]} \pi_S(S) \left\{ \pi_{\mathbf{A}_S}(\mathbf{A}_S) \mathrm{d}\mathbf{A}_S \right\} \left\{ \delta_{\mathbf{0}_{|S^c| \times r}}(\mathrm{d}\mathbf{A}_{S^c}) \right\}, \\ \pi_{\mathbf{A}_S}(\mathbf{A}_S)\mathrm{d}\mathbf{A}_S &:= \frac{\exp\{-2\|\mathrm{vec}(\mathbf{A}_S)\|_1\} \mathbb{1}(\|\mathbf{A}_S\|_2 < 1) \mathrm{d}\mathbf{A}_S}{\int_{\|\mathbf{A}_S\|_2 < 1} \exp(-2\|\mathrm{vec}(\mathbf{A}_S)\|_1) \mathrm{d}\mathbf{A}_S}. \end{aligned} \tag{13}$$

The prior distribution on the entire covariance matrix $\mathbf{\Omega}$ through $\boldsymbol{\theta}$ is completed by assigning the following prior distribution to $\mathbf{M}$, which is independent of $\Pi_{\mathbf{A}}(\mathrm{d}\mathbf{A})$:

$$\pi_{\boldsymbol{\mu}}(\boldsymbol{\mu}) \propto \exp\left(-2\|\boldsymbol{\mu}\|_1\right) \mathbb{1}\{\mathbf{M}(\boldsymbol{\mu}) \in \mathbb{M}_+(r)\}. \tag{14}$$

Then the joint prior distribution on $\boldsymbol{\theta} = [\boldsymbol{\varphi}^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{T}}]^{\mathrm{T}} = [\mathrm{vec}(\mathbf{A})^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{T}}]^{\mathrm{T}} \in \mathscr{D}(p, r)$ is defined as the product of the sparsity inducing prior (13) on $\mathbf{A}$ and the prior distribution (14) on $\boldsymbol{\mu}$:

$$\Pi_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{\theta}) := \Pi_{\mathbf{A}}(\mathrm{d}\mathbf{A}) \pi_{\boldsymbol{\mu}}(\boldsymbol{\mu}) \mathrm{d}\boldsymbol{\mu}. \tag{15}$$

Denote by $\mathbf{Y}_n := [\mathbf{y}_1, \dots, \mathbf{y}_n] \in \mathbb{R}^{p \times n}$ the data matrix concatenated by $\mathbf{y}_1, \dots, \mathbf{y}_n$ and

$$\ell(\mathbf{\Omega}) := -\frac{n}{2} \log\det(2\pi\mathbf{\Omega}) - \frac{n}{2} \mathrm{tr}(\widehat{\mathbf{\Omega}}\mathbf{\Omega}^{-1})$$

the log-likelihood function of $\mathbf{\Omega}$, where $\widehat{\mathbf{\Omega}} := (1/n) \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^{\mathrm{T}}$. We assume the high-dimensionality setup $p/n \to \infty$ so that the sample covariance matrix $\widehat{\mathbf{\Omega}}$ is no longer invertible. Then, the posterior distribution of interest given the data matrix $\mathbf{Y}_n$ can be written using the Bayes formula:

$$\Pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in A \mid \mathbf{Y}_n) = \frac{\int_A \exp\{\ell(\mathbf{\Omega}(\boldsymbol{\theta})) - \ell(\mathbf{\Omega}_0)\} \Pi_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{\theta})}{\int \exp\{\ell(\mathbf{\Omega}(\boldsymbol{\theta})) - \ell(\mathbf{\Omega}_0)\} \Pi_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{\theta})},$$

where A is any measurable subset of $\mathscr{D}(p, r)$.

3.2. Rate-optimal posterior contraction in spectral norm

The main result of this section is Theorem 3.1 below. It asserts that the posterior contraction rate under $\|\sin\Theta\{\mathbf{U}(\boldsymbol{\varphi}), \mathbf{U}_0\}\|_2$ is minimax optimal under additional necessary conditions, which we present below first:

A1 (Minimum row support) $s := |\mathrm{supp}(\mathbf{U}_0)| \geq 2r$.
A2 (Regularity) $\sup_n \|\mathbf{A}_0\|_2 < 1$.

A3 (Bounded spectra) There exists some constants $\underline{\lambda}, \overline{\lambda} > 0$ such that

$$\underline{\lambda} \leq \lambda_r(\mathbf{M}_0) \leq \lambda_1(\mathbf{M}_0) \leq \overline{\lambda}.$$

A4 (Fast convergence rate) $(r^2 s^2 \log n + rs^2 \log p)^3/n \to 0$.

A5 (Minimum row signal strength) The non-zero rows of $\mathbf{A}_0$ satisfy

$$\lim_{n\to\infty} \frac{\min_{j\in\mathrm{supp}(\mathbf{A}_0)} \|[\mathbf{A}_0]_{j*}\|_2}{\sqrt{(rs\log n + s\log p)/n}} = \infty.$$

Remark 2. Some remarks regarding conditions A1–A5 are in order. Conditions A1 and A3 are standard conditions for sparse spiked covariance model (see, for example, condition 2 in [53] for A1 and the parameter space in [21] for A3). Condition A2 requires that the spectral norm of $\mathbf{A}_0$ is bounded away from 1. In [55], the author shows that, locally around $\mathbf{U}_0$, the Frobenius sine-theta distance between $\mathbf{U}$ and $\mathbf{U}_0$ is equivalent to the Frobenius norm $\|\mathbf{U} - \mathbf{U}_0\|_{\mathrm{F}}$, and hence, $\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_0\|_2$, up to a constant factor. Furthermore, Theorem 2.1 and Remark 1 indicate that the Fisher information matrix with regard to the $\boldsymbol{\theta}$-parameterization evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, given by

$$\mathbf{I}(\boldsymbol{\theta}_0) = \frac{1}{2} D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}} (\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1}) D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0), \tag{16}$$

is asymptotically non-singular, i.e., $\|\mathbf{I}(\boldsymbol{\theta}_0)^{-1}\|_2$ is bounded away from ∞ when $n \to \infty$.

Condition A4 claims that the posterior contraction rate with regard to $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1$ is sufficiently fast. Furthermore, roughly speaking, using the Fréchet derivative formula $D\boldsymbol{\Sigma}(\cdot)$, we are able to derive a local asymptotic normality expansion of the log-likelihood function $\ell(\boldsymbol{\Omega}(\boldsymbol{\theta}))$ as follows:

$$\begin{aligned} \ell(\boldsymbol{\Omega}(\boldsymbol{\theta})) - \ell(\boldsymbol{\Omega}_0) = {} & \frac{n}{2} \mathrm{vec}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)^{\mathrm{T}} (\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1}) D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ & - \frac{n}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{I}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0), \end{aligned} \tag{17}$$

where condition A4 guarantees that remainder R_n is negligible.

Condition A5 requires that the minimum of the Euclidean norms of the non-zero rows of $\mathbf{A}_0$ cannot be too small. It is similar to the so-called β-min condition in the sparse linear regression model (see, e.g., 6). In [34], a similar condition is also required for the exact recovery of $\mathrm{supp}(\mathbf{U}_0)$ (*i.e.*, the notion of sparsistency) using the Fantope projection and selection method.

It is quite clear that condition A3 can be satisfied by a broad collection of positive definite matrices. Below, we provide a class of examples of $\mathbf{A}_0$ satisfying conditions A1, A2, A4, and A5.

Example. Let $\gamma_1, \gamma_2 > 0$ be constants such that $s = 2r = O(n^{\gamma_1})$, $\log p = O(n^{\gamma_2})$, and they satisfy $12\gamma_1 < 1$, $9\gamma_1 + 3\gamma_2 < 1$. These constraints guarantee that condition A4 holds. Let $\alpha_n, \beta_n > 0$ be n-dependent quantities with $\alpha_n^2 + \beta_n^2 = 1$, and $\sup_{n\geq 1}\beta_n < 1$. Consider

$$\mathcal{U} = \left\{\operatorname{diag}(\mathbf{I}_r, \mathbf{P})\begin{bmatrix}\alpha_n\mathbf{I}_r\\ \beta_n\mathbf{H}_0\\ \mathbf{0}_{(p-2r)\times r}\end{bmatrix} : \mathbf{P} \text{ is a permutation matrix, } \mathbf{H}_0 \in \mathbb{O}(r)\right\}.$$

Clearly, any $\mathbf{U}_0 \in \mathcal{U}$ satisfies condition A1. Now let

$$\mathbf{A}_0 = \frac{\beta_n}{\alpha_n + 1}\mathbf{P}\begin{bmatrix}\mathbf{H}_0\\ \mathbf{0}_{(p-2r)\times r}\end{bmatrix},$$

where $\mathbf{P}$ is a $(p-r)\times(p-r)$ permutation matrix and $\mathbf{H}_0 \in \mathbb{O}(r)$. Clearly, $\sup_{n\geq 1}\|\mathbf{A}_0\|_2 \leq \sup_{n\geq 1}\beta_n < 1$, so that condition A2 holds. A simple algebra shows that

$$\mathbf{U}(\operatorname{vec}(\mathbf{A}_0)) = \operatorname{diag}(\mathbf{I}_r, \mathbf{P})\begin{bmatrix}\alpha_n\mathbf{I}_r\\ \beta_n\mathbf{H}_0\\ \mathbf{0}_{(p-2r)\times r}\end{bmatrix} \in \mathcal{U}.$$

Finally, because $(rs\log n + s\log p)/n = o(1)$ and

$$\min_{j\in\operatorname{supp}(\mathbf{A}_0)}\|[\mathbf{A}_{0j}\|_2 = \min_{j=1,\ldots,r}\frac{\beta_n}{\alpha_n+1}\|[\mathbf{H}_0]_{j*}\|_2 = \frac{\sqrt{1-\beta_n^2}}{\beta_n+1}$$

is bounded away from 0, condition A5 holds automatically.

Theorem 3.1. *Assume the setup and the prior specification in Section 3.1 and suppose conditions A1–A5 hold. There exists some large constant $M > 0$ independent of r, s, p, n, such that*

$$\mathbb{E}_0\Pi_{\boldsymbol{\theta}}\left\{\|\sin\Theta(\mathbf{U}(\boldsymbol{\varphi}), \mathbf{U}_0)\|_2 > M\sqrt{\frac{s\log p}{n}}\;\middle|\;\mathbf{Y}_n\right\} \lesssim \sqrt{\frac{(r^2s^2\log n + rs^2\log p)^3}{n}}.$$

In particular,

$$\lim_{n\to\infty}\mathbb{E}_0\Pi_{\boldsymbol{\theta}}\left\{\|\sin\Theta(\mathbf{U}(\boldsymbol{\varphi}), \mathbf{U}_0)\|_2 > M\sqrt{\frac{s\log p}{n}}\;\middle|\;\mathbf{Y}_n\right\} = 0.$$

Theorem 3.1 is a non-trivial result. It relies on the asymptotic characterization of the shape of the posterior distribution $\Pi_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{\theta} \mid \mathbf{Y}_n)$, which is summarized in Theorem 3.2 below.

Theorem 3.2. *Assume the setup and the prior specification in Section 3.1 and suppose conditions A1–A5 hold. For any index set*

$$S \in \mathcal{S}_0 := \{S \subset [p-r] : \operatorname{supp}(\mathbf{A}_0) \subset S, |S| \leq \kappa_0(s-r)\},$$

let $\mathbf{F}_S$ be the matrix such that

$$\boldsymbol{\theta}_S := \begin{bmatrix}\operatorname{vec}(\mathbf{A}_S)\\ \boldsymbol{\mu}\end{bmatrix} = \mathbf{F}_S^{\mathrm{T}}\boldsymbol{\theta} \quad \textit{for any vector } \boldsymbol{\theta} = \begin{bmatrix}\operatorname{vec}(\mathbf{A})\\ \boldsymbol{\mu}\end{bmatrix}.$$

Furthermore, define the following quantities:

$$\mathbf{I}_S(\boldsymbol{\theta}_0) := \mathbf{F}_S^{\mathrm{T}}\mathbf{I}(\boldsymbol{\theta}_0)\mathbf{F}_S,$$
$$\widehat{\boldsymbol{\theta}}_S := \boldsymbol{\theta}_{0S} + (1/2)\mathbf{I}_S(\boldsymbol{\theta}_0)^{-1}\mathbf{F}_S^{\mathrm{T}} D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}\mathrm{vec}\{\boldsymbol{\Omega}_0^{-1}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\},$$
$$\widehat{w}_S \propto \frac{\pi_p(|S|)|2\pi\mathbf{I}_S(\boldsymbol{\theta}_0)^{-1}/n|^{1/2}\exp\{(n/2)\widehat{\boldsymbol{\theta}}_S^{\mathrm{T}}\mathbf{I}_S(\boldsymbol{\theta}_0)\widehat{\boldsymbol{\theta}}_S\}}{\binom{p-r}{|S|}\int_{\{\|\mathbf{A}_S\|_2<1\}}\exp(-2\|\mathrm{vec}(\mathbf{A}_S)\|_1)\mathrm{d}\mathbf{A}_S}\ \textit{s.t.}\ \sum_{S\in\mathcal{S}_0}\widehat{w}_S = 1,$$

where $\mathbf{I}(\theta_0)$ *is defined in* (16)*. Let* $\Pi_{\boldsymbol{\theta}}^{\infty}(\boldsymbol{\theta} \in \cdot \mid \mathbf{Y}_n)$ *be the following random mixture of normals*

$$\Pi_{\boldsymbol{\theta}}^{\infty}(\mathrm{d}\boldsymbol{\theta} \mid \mathbf{Y}_n) := \sum_{S\in\mathcal{S}_0}\widehat{w}_S\{\phi(\boldsymbol{\theta}_S \mid \widehat{\boldsymbol{\theta}}_S, \mathbf{I}_S(\boldsymbol{\theta}_0)^{-1}/n)\mathrm{d}\boldsymbol{\theta}_S\}\{\delta_{\mathbf{0}_{|S^c|\times r}}(\mathrm{d}\boldsymbol{\theta}_{S^c})\}, \quad (18)$$

Here, $\phi(\mathbf{x} \mid \mathbf{u}, \boldsymbol{\Omega}) := \det(2\pi\boldsymbol{\Omega})^{-1/2}e^{-(\mathbf{x}-\mathbf{u})^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\mathbf{x}-\mathbf{u})/2}$ *and* $\boldsymbol{\theta}_{S^c} := \mathrm{vec}(\mathbf{A}_{S^c})$*. Then there exists some constant* $\kappa_0 \geq 1$ *such that*

$$\mathbb{E}_0\|\Pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in \cdot \mid \mathbf{Y}_n) - \Pi_{\boldsymbol{\theta}}^{\infty}(\boldsymbol{\theta} \in \cdot \mid \mathbf{Y}_n)\|_{\mathrm{TV}} \lesssim \sqrt{\frac{(r^2s^2\log n + rs^2\log p)^3}{n}}.$$

Remark 3. By Theorem 2.1 and Remark 1, the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta}_0)$ is strictly positive definite, implying that the submatrix $\mathbf{I}_S(\boldsymbol{\theta}_0) := \mathbf{F}_S^{\mathrm{T}}\mathbf{I}(\boldsymbol{\theta}_0)\mathbf{F}_S$ is also strictly positive definite. Hence, leveraging the intrinsic perturbation tools developed in [55] and Theorem 3.2, we can further study the behavior of $\|\sin\Theta\{\mathbf{U}(\boldsymbol{\varphi}), \mathbf{U}_0\}\|_2$ under the posterior distribution $\Pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in \cdot \mid \mathbf{Y}_n)$ through the behavior of $\boldsymbol{\theta} - \boldsymbol{\theta}_0$ under the limiting distribution $\Pi_{\boldsymbol{\theta}}^{\infty}(\boldsymbol{\theta} \in \cdot \mid \mathbf{Y}_n)$. Theorem 3.2 may be of independent interest as well.

3.3. Proof sketch

We now discuss the proof sketch for Section 3.2, namely, Theorem 3.1. The proof is lengthy and is partitioned into several subsections. The sketch of the proof can be loosely summarized as the following steps:

1. Prior concentration (Section 5.1). We provide a lower bound for the prior probability that $\boldsymbol{\Omega}(\boldsymbol{\theta})$ is inside a small neighborhood of $\boldsymbol{\Omega}_0$, i.e.,

$$\Pi_{\boldsymbol{\theta}}\{\|\boldsymbol{\Omega}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_0\|_{\mathrm{F}} \leq \eta_n\},$$

   where $(\eta_n)_{n=1}^{\infty}$ is a sequence converging to 0.
2. Posterior sparsity (Section 5.2). We prove that with posterior probability going to 1, the intrinsic dimension cannot be too large, namely,

$$\lim_{n\to\infty}\mathbb{E}_0\Pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} : |S_{\mathbf{A}}| \leq \kappa_0 s \mid \mathbf{Y}_n) = 0$$

   for some constant $\kappa_0 > 0$.

3. Construction of certain test functions (Section 5.3). This step is needed to obtain the rate-optimal posterior contraction under the Frobenius norm following the general framework of [25].
4. Posterior contraction under the Frobenius norm (Section 5.4). This is immediate once the previous steps are completed but also serves as an intermediate step to the posterior contraction under the spectral norm.
5. Local asymptotic normality (Section 5.5). We expand the log-likelihood function locally at $\boldsymbol{\theta}_0$ under the posterior sparsity restriction via a Taylor expansion argument, which can be viewed as a variant of the local asymptotic normality (see, e.g., Chapter 7 in [50]). The neighborhood radius of the local asymptotic normality is determined by the posterior contraction rate under the Frobenius norm established in Section 5.4.
6. Distributional approximation (Section 5.6). Leveraging the local asymptotic normality result established in Section 5.5, we prove Theorem 3.2, i.e., the asymptotic characterization of the shape of the posterior distribution $\Pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in \cdot \mid \mathbf{Y}_n)$ using a random mixture of normal distributions.
7. Posterior contraction under the spectral norm (Section 5.7). Finally, we prove the rate-optimal posterior contraction of $\mathrm{Span}(\mathbf{U})$ under the spectral sine-theta distance using the asymptotic distributional approximation result obtained in Theorem 3.2.

We now briefly explain the purposes of the above technical steps. As mentioned earlier, the key technical result is Theorem 3.2. In steps 1–4, we follow the classical prior-concentration-and-testing framework established in [24, 25] to obtain the posterior contraction rate under the Frobenius norm. Such a preliminary convergence result is necessary because it enables us to obtain the local asymptotic normality (step 5) in a shrinking neighborhood of $\boldsymbol{\theta}_0$ corresponding to the Frobenius norm contraction rate. The local asymptotic normality established via step 5 is also of fundamental interest for step 6, which completes the proof of Theorem 3.2. Finally, our main spectral norm posterior contraction result in Theorem 3.1 is obtained from Theorem 3.2 using a discretization trick.

3.4. The challenge with the testing approach

We briefly sketch the argument for why minimax-optimal posterior contraction under the spectral sine-theta distance is challenging in the current sparse spiked covariance model context if one follows the prior-concentration-and-testing framework pioneered by [24, 25]. Let $\boldsymbol{\theta}_0 = [\boldsymbol{\varphi}_0^{\mathrm{T}}, \boldsymbol{\mu}_0^{\mathrm{T}}]^{\mathrm{T}} \in \mathscr{D}(p, r)$. The key step there is the construction of a test function Ψ_n for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_A : \|\sin\Theta(\mathbf{U}(\boldsymbol{\varphi}), \mathbf{U}(\boldsymbol{\varphi}_0)\|_2 > M\sqrt{(s\log p)/n}$, such that

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\theta}_0}\Psi_n = 0, \qquad \sup_{\boldsymbol{\theta}:\|\sin\Theta(\mathbf{U}(\boldsymbol{\varphi}),\mathbf{U}(\boldsymbol{\varphi}_0)\|_2 > M\sqrt{(s\log p)/n}} \mathbb{E}_{\boldsymbol{\theta}}(1 - \Psi_n) \lesssim e^{-cn\eta_n^2}, \tag{19}$$

where $\eta_n = \sqrt{(rs\log n + s\log p)/n}$ corresponds to the contraction rate under the Frobenius norm loss (intrinsic metric), and $c > 0$ is a constant that cannot be easily controlled. See, for example, Equation (5.6) in [27] for a detailed derivation. Now let $\boldsymbol{\theta}_1 = [\boldsymbol{\varphi}_1^{\mathrm{T}}, \boldsymbol{\mu}_0^{\mathrm{T}}]^{\mathrm{T}} \in \mathscr{D}(p, r)$ be such that $\|\sin\Theta(\mathbf{U}(\boldsymbol{\varphi}_1), \mathbf{U}(\boldsymbol{\varphi}_0))\|_{\mathrm{F}} = 2M\sqrt{(s\log p)/n}$, $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}(\boldsymbol{\theta}_1)$, and consider the likelihood ratio test

$$\begin{aligned}\psi_{n,\boldsymbol{\theta}_1} = \mathbb{1}\big\{\ell(\boldsymbol{\Omega}_1) - \ell(\boldsymbol{\Omega}_0) > \sqrt{n}u_n\|\boldsymbol{\Omega}_1^{-1/2}(\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_1^{-1/2}\|_{\mathrm{F}} \\ + n\mathrm{tr}(\mathbf{I}_p - \boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_0)\big\}\end{aligned}$$

such that $\mathbb{E}_{\boldsymbol{\theta}_0}\Psi_n = \mathbb{E}_{\boldsymbol{\theta}_0}\psi_{n,\boldsymbol{\theta}_1}$, where $u_n \to \infty$ in order to obtain a vanishing type-I error probability. Because $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Omega}_1$ have the same set of eigenvalues, it is clear that $\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}\|_{\mathrm{F}} \geq CM\sqrt{(s\log p)/n}$ by Davis-Kahan theorem, where $C > 0$ is a constant. Note that $\ell(\boldsymbol{\Omega}_1) - \ell(\boldsymbol{\Omega}_0) = (1/2)\sum_{i=1}^n \mathbf{y}_i^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Omega}_1^{-1})\mathbf{y}_i$ since $\det(\boldsymbol{\Omega}_0^{-1}) = \det(\boldsymbol{\Omega}_1^{-1})$. Then by Neyman-Pearson lemma,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}_1}(1 - \Psi_n) \\ &\quad\geq \mathbb{E}_{\boldsymbol{\theta}_1}(1 - \psi_{n,\boldsymbol{\theta}_1}) \\ &\quad= \mathbb{P}_{\boldsymbol{\theta}_1}\Bigg\{\sum_{i=1}^n \mathbf{y}_i^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Omega}_1^{-1})\mathbf{y}_i - n\mathrm{tr}(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Omega}_1 - \mathbf{I}_p) \leq \\ &\qquad\qquad \sqrt{n}u_n\|\boldsymbol{\Omega}_1^{-1/2}(\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_1^{-1/2}\|_{\mathrm{F}} + n\mathrm{tr}(2\mathbf{I}_p - \boldsymbol{\Omega}_0^{-1}\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_0)\Bigg\}.\end{aligned}$$

In order to obtain a consistent likelihood ratio test, the right-hand side of the above inequality needs to be lower bounded by

$$\mathbb{P}_{\boldsymbol{\theta}_1}\Bigg\{\sum_{i=1}^n \mathbf{y}_i^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Omega}_1^{-1})\mathbf{y}_i - n\mathrm{tr}(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Omega}_1 - \mathbf{I}_p) \leq -c's\log p\Bigg\} = O(e^{-c''s\log p})$$

by Hanson-Wright inequality [46], where c', c'' are constants that cannot be well controlled. Since $s\log p = o(n\eta_n^2)$ in situations where $\log p = o(r\log n)$, we conjecture that it is impossible to construct a test function satisfying (19). For a more thorough discussion on the testing approach and the posterior contraction rates under non-intrinsic metrics in more general contexts, see Section 5.3 of [27].

4. Discussion

In this paper, we leverage the techniques of Cayley parameterization of subspaces from [28, 47] and Euclidean representation of low-rank matrices from [55] to address the spectral norm posterior contraction rate in Bayesian sparse spiked covariance matrix estimation problem. The non-trivial part of this issue is that the spectral norm is not equivalent to the intrinsic Fisher information metric on the statistical manifold of the Gaussian spiked covariance model, which prevents us from using the so-called "master theorems" established in [24, 25] to derive the associated posterior contraction rate. See [27] for a discussion on the posterior contraction rates with regard to non-intrinsic metrics. Our proof technique relies on a complete characterization of the shape of the posterior distribution through a non-trivial local asymptotic normality argument as well as the recent advance in the singular values of the Fréchet derivative of Cayley parameterization due to [55].

There are several potential extensions of this work. Our conditions (conditions A2, A4, and A5) are slightly stronger than those required in the frequentist minimax-optimal estimator literature [7], and it is currently unclear whether these conditions can be relaxed following the framework established in this work. We conjecture that new techniques may be required to relax these conditions. The computation strategy for the posterior distribution derived in this work is a non-trivial task due to both the high-dimensional nature of the problem and the lack of a closed-form Gibbs sampler associated with the spike-and-slab prior specification through the Cayley parameterization. A plausible computation strategy is to leverage the proximal mapping technique developed in [57, 58] together with the tools from Hamiltonian Monte Carlo. Specifically, the main idea in [57, 58] is to construct an auxiliary variable $\boldsymbol{\beta}$ with an absolutely continuous prior distribution $\pi_{\boldsymbol{\beta}}$, such that the induced prior distribution defined by the proximal mapping $\boldsymbol{\varphi} = \arg\min_{\|\boldsymbol{\psi}\|_1 \leq \xi} \|\boldsymbol{\beta} - \boldsymbol{\psi}\|_2$ has the same prior distribution as the sparsity-inducing prior (13), where the hyperparameter ξ is also assigned an appropriate hyperprior. The mechanism of the sparsity-inducing effect of the proximal mapping resembles the shrinkage effect of LASSO [48]. By doing so, the authors of [57, 58] argued that the posterior computation for $\boldsymbol{\varphi}$ can be implemented tractably by sampling $\boldsymbol{\beta}$ using Hamiltonian Monte Carlo. Finally, we currently assume that the rank r of the spiked component of the true covariance matrix is either known or estimated using a consistent estimator (*e.g.*, [7]). It is also possible to relax this assumption by first assigning a prior distribution for r that is supported on $\{1, \ldots, r_{\max}\}$, where $r_{\max} > 0$ is a conservative upper bound for r, and then construct the entire hierarchical prior distribution by treating (11)–(15) as the conditional prior distribution of the remaining parameters given r. This may require a nontrivial extension of the distributional approximation result in Theorem 3.2 because it is unclear how to formulate the local asymptotic normality with a varying dimensional parameter space. We defer these research directions to future work.

5. Proofs of the main results

Denote by $S_0 := \mathrm{supp}(\mathbf{A}_0)$, $s_0 := |S_0|$, and

$$\gamma(|S|) := \int_{\|\mathbf{A}_S\|_2 < 1} \exp(-2\|\mathrm{vec}(\mathbf{A}_S)\|_1) \mathrm{d}\mathbf{A}_S.$$

We begin the proof with the following upper and lower bounds for $\gamma(|S|)$:

$$\gamma(|S|) \leq \int_{\mathbb{R}^{|S| \times r}} \exp(-2\|\mathrm{vec}(\mathbf{A}_S)\|_1) \mathrm{d}\mathbf{A}_S = 1, \tag{20}$$

$$\gamma(|S|) \geq \exp\left\{-\frac{1}{2} r|S| \log(r|S|) - (2 - \log 2) r|S|\right\}, \quad |S| \geq 1, \tag{21}$$

where the lower bound can be derived as follows:

$$\begin{aligned}\gamma(|S|) &\geq \int_{\|\mathbf{A}_S\|_{\mathrm{F}}<1} \exp(-2\|\mathrm{vec}(\mathbf{A}_S)\|_1)\mathrm{d}\mathbf{A}_S\\ &\geq \int_{\|\mathrm{vec}(\mathbf{A}_S)\|_\infty \leq (r|S|)^{-1/2}} \exp(-2\|\mathrm{vec}(\mathbf{A}_S)\|_1)\mathrm{d}\mathbf{A}_S\\ &= \left\{\int_{-(r|S|)^{-1/2}}^{(r|S|)^{-1/2}} e^{-2x}\mathrm{d}x\right\}^{r|S|} \geq \left\{\frac{2e^{-2}}{(r|S|)^{1/2}}\right\}^{r|S|}\\ &= \exp\left\{-\frac{1}{2}r|S|\log(r|S|) - (2-\log 2)r|S|\right\}.\end{aligned}$$

Also, observe that for sufficiently large n, $z_n \in [1/2, 2]$.

5.1. Prior concentration

This subsection focuses on proving the following lemma that describes the prior concentration behavior of $\Pi_{\boldsymbol{\theta}}(\cdot)$:

Lemma 5.1. *Assume the prior specification in Section 3.1 and conditions A1–A5 hold. If $(\eta_n)_{n=1}^\infty$ is a sequence such that $\eta_n/\|\boldsymbol{\Omega}_0\|_2 \to 0$ and $n\eta_n^2 \to \infty$, then*

$$\Pi_{\boldsymbol{\theta}}\{\|\boldsymbol{\Omega}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_0\|_{\mathrm{F}} < \eta_n\} \geq \exp\left(-C_0 r s_0 \log n - C s_0 \log p\right)$$

for some constant $C_0 = C(\|\boldsymbol{\Omega}_0\|_2) > 0$ that only depends on $\|\boldsymbol{\Omega}_0\|_2$, and some absolute constant $C > 0$.

Before proving Lemma 5.1, we need the following auxiliary lemma from [41].

Lemma 5.2 (Lemma C.1 in the Supplement of [56])**.** *Let $(\eta_n)_{n=1}^\infty$ be a sequence converging to 0 with $n\eta_n^2 \to \infty$. Then there exist a constant $c > 0$ and a sequence of events $(\Xi_n)_{n=1}^\infty$ with $\mathbb{P}_0(\Xi_n) \leq 2\exp(-cn\eta_n^2/\|\boldsymbol{\Omega}_0\|_2^2)$ such that over the event Ξ_n,*

$$D_n \geq \exp\left\{-Cn\eta_n^2\log(2\|\boldsymbol{\Omega}_0\|_2)\right\}\Pi_{\boldsymbol{\theta}}\left\{\boldsymbol{\theta} : \|\boldsymbol{\Omega}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_0\|_{\mathrm{F}} < \eta_n\right\},$$

where $D_n := \int \exp\{\ell(\boldsymbol{\Omega}(\boldsymbol{\theta})) - \ell(\boldsymbol{\Omega}_0)\}\Pi_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{\theta})$.

Proof of Lemma 5.1. Let $\boldsymbol{\varphi}_0 := \mathrm{vec}(\mathbf{A}_0) = \mathbf{U}^{-1}(\mathbf{U}_0)$ be such that $\mathbf{U}(\boldsymbol{\varphi}_0) = \mathbf{U}_0$, where $\mathbf{U}(\boldsymbol{\varphi})$ is the Cayley transform of $\boldsymbol{\varphi} \in \mathbb{R}^d$, $\boldsymbol{\mu}_0$ is the vector formed by taking the upper diagonal entries of $\mathbf{M}_0$, and $\boldsymbol{\theta}_0 := [\boldsymbol{\varphi}_0^{\mathrm{T}}, \boldsymbol{\mu}_0^{\mathrm{T}}]^{\mathrm{T}}$. By Theorem 2.1 in [55], for any $\boldsymbol{\varphi} \in B_2(\boldsymbol{\varphi}_0, \epsilon)$ with sufficiently small $\epsilon > 0$,

$$\|\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\|_{\mathrm{F}} \leq 2\|D\mathbf{U}(\boldsymbol{\varphi}_0)\|_2\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_0\|_2 \leq 4\sqrt{2}\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_0\|_2.$$

Therefore, for $\boldsymbol{\varphi} \in B_2(\boldsymbol{\varphi}_0, \eta_n/[32\|\boldsymbol{\Omega}_0\|_2])$ and $\boldsymbol{\mu} \in B_2(\boldsymbol{\mu}_0, \eta_n/8)$ with sufficiently large n, we have

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq \|\boldsymbol{\varphi} - \boldsymbol{\varphi}_0\|_2 + \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|_2 < \frac{\eta_n}{32\|\boldsymbol{\Omega}_0\|_2} + \frac{\eta_n}{8} \leq \frac{\eta_n}{4},$$

and then, for sufficiently large n, we obtain

$$\begin{aligned}\|\boldsymbol{\Omega}(\boldsymbol{\theta})-\boldsymbol{\Omega}_0\|_{\mathrm{F}} &= \|\boldsymbol{\Sigma}(\boldsymbol{\theta})-\boldsymbol{\Sigma}_0\|_{\mathrm{F}}\\ &\leq 2\|\mathbf{M}_0\|_2\|\mathbf{U}(\boldsymbol{\varphi})-\mathbf{U}_0\|_{\mathrm{F}}+\|\mathbf{M}-\mathbf{M}_0\|_{\mathrm{F}}\\ &\quad+2\|\mathbf{M}-\mathbf{M}_0\|_{\mathrm{F}}\|\mathbf{U}(\boldsymbol{\varphi})-\mathbf{U}_0\|_{\mathrm{F}}+\|\mathbf{M}_0\|_2\|\mathbf{U}(\boldsymbol{\varphi})-\mathbf{U}_0\|_{\mathrm{F}}^2\\ &\leq 8\sqrt{2}\|\boldsymbol{\Omega}_0\|_2\|\boldsymbol{\varphi}-\boldsymbol{\varphi}_0\|_2+\sqrt{2}\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|_2+32\|\boldsymbol{\theta}-\boldsymbol{\theta}\|_2^2\\ &\quad+\|\boldsymbol{\Omega}_0\|_2\|\mathbf{U}(\boldsymbol{\varphi})-\mathbf{U}_0\|_{\mathrm{F}}^2\\ &\leq 16\|\boldsymbol{\Omega}_0\|_2\|\mathbf{A}-\mathbf{A}_0\|_2+2\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|_2<\eta_n.\end{aligned}$$

Now we can estimate the prior mass $\Pi_{\boldsymbol{\theta}}\{\|\boldsymbol{\Omega}(\boldsymbol{\theta})-\boldsymbol{\Omega}_0\|_{\mathrm{F}}<\eta_n\}$ from below:

$$\begin{aligned}\Pi_{\boldsymbol{\theta}}\{\|\boldsymbol{\Omega}(\boldsymbol{\theta})-\boldsymbol{\Omega}_0\|_{\mathrm{F}}<\eta_n\} \geq \Pi_{\boldsymbol{\mu}}&\left(\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|_2<\frac{\eta_n}{8}\right)\\ &\times\Pi_{\mathbf{A}}\left(\|\mathbf{A}-\mathbf{A}_0\|_{\mathrm{F}}<\frac{\eta_n}{32\|\boldsymbol{\Omega}_0\|_2}\right),\end{aligned}$$

where $\Pi_{\boldsymbol{\mu}}(\mathrm{d}\boldsymbol{\mu})=\pi_{\boldsymbol{\mu}}(\boldsymbol{\mu})\mathrm{d}\boldsymbol{\mu}$. Denote $\Pi_{\boldsymbol{\mu}}^L(\mathrm{d}\boldsymbol{\mu})$ the Laplace distribution on $\boldsymbol{\mu}$ by

$$\Pi_{\boldsymbol{\mu}}^L(\mathrm{d}\boldsymbol{\mu}):=\exp(-2\|\boldsymbol{\mu}\|_1)\mathrm{d}\boldsymbol{\mu},\quad \boldsymbol{\mu}\in\mathbb{R}^{r(r+1)/2}.$$

Clearly, $\Pi_{\boldsymbol{\mu}}(\mathrm{d}\boldsymbol{\mu})$ is the normalized restriction of $\Pi_{\boldsymbol{\mu}}^L(\mathrm{d}\boldsymbol{\mu})$ on $\mathbf{M}(\boldsymbol{\mu})\in\mathbb{M}_+(r)$, where $\boldsymbol{\mu}=\mathrm{vech}\{\mathbf{M}(\boldsymbol{\mu})\}$. Now let $\boldsymbol{\mu}\in B_2(\boldsymbol{\mu}_0,\eta_n/8)$. Then for sufficiently large n,

$$\|\mathbf{M}(\boldsymbol{\mu})-\mathbf{M}(\boldsymbol{\mu}_0)\|_{\mathrm{F}}\leq\sqrt{2}\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|_2<\frac{\sqrt{2}\eta_n}{8}\to 0.$$

Since $\mathbf{M}_0$ is already strictly positive definite with $\lambda_r(\mathbf{M}_0)$ bounded away from 0, it follows that $\mathbf{M}(\boldsymbol{\mu})$ is also positive definite. Now we proceed to provide a lower bound the first factor as follows for sufficiently large n:

$$\begin{aligned}\Pi_{\boldsymbol{\mu}}\left(\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|_2<\frac{\eta_n}{8}\right) &= \frac{\Pi_{\boldsymbol{\mu}}^L\{\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|_2<\eta_n/8,\mathbf{M}(\boldsymbol{\mu})\in\mathbb{M}_+(r)\}}{\Pi_{\boldsymbol{\mu}}^L\{\mathbf{M}(\boldsymbol{\mu})\in\mathbb{M}_+(r)\}}\\ &= \frac{\Pi_{\boldsymbol{\mu}}^L(\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|_2<\eta_n/8)}{\Pi_{\boldsymbol{\mu}}^L\{\mathbf{M}(\boldsymbol{\mu})\in\mathbb{M}_+(r)\}}\\ &\geq \Pi_{\boldsymbol{\mu}}^L\left(\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|_2<\frac{\eta_n}{8}\right)\\ &\geq \mathrm{vol}\left\{B_2\left(\boldsymbol{\mu}_0,\frac{\eta_n}{8}\right)\right\}\\ &\quad\times\exp\left\{-2\max_{\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|_2\leq 1}(\|\boldsymbol{\mu}_0\|_1+\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|_1)\right\}\\ &\geq \mathrm{vol}\left\{B_2\left(\mathbf{0}_{r(r+1)/2},1\right)\right\}\left(\frac{\eta_n}{8}\right)^{r(r+1)/2}\\ &\quad\times\exp\left\{-2\sqrt{\frac{r(r+1)}{2}}(\|\boldsymbol{\mu}_0\|_2+1)\right\}\end{aligned}$$

$$
\begin{aligned}
&\gtrsim \frac{1}{\sqrt{\pi r(r+1)/2}}\left(\frac{\sqrt{2\pi e}\eta_n}{8\sqrt{r(r+1)/2}}\right)^{r(r+1)/2} \\
&\quad\times \exp\left(-2r\|\mathbf{M}_0\|_{\mathrm{F}}\right) \\
&\geq \exp\left(-Cr^2\left|\log\frac{\eta_n}{r}\right| - 2r^{3/2}\|\boldsymbol{\Omega}_0\|_2\right) \\
&\geq \exp\left(-Crs\left|\log\frac{\eta_n}{r}\right| - 2rs\|\boldsymbol{\Omega}_0\|_2\right).
\end{aligned}
$$

We now focus on the first factor. Note that for any row index $j \in [p]$, $j > r$, $[\mathbf{U}_0]_{j*} = \mathbf{0}_r$ if and only if $[\mathbf{A}(\mathbf{U}_0)]_{(j-r)*} = \mathbf{0}_r$. Given S drawn from $\pi_S(S)$, denote $\Pi^L_{\mathbf{A}_S}(\mathrm{d}\mathbf{A}_S)$ the Laplace distribution on $\mathrm{vec}(\mathbf{A}_S)$, i.e.,

$$
\Pi^L_{\mathbf{A}_S}(\mathrm{d}\mathbf{A}_S) := \exp(-2\|\mathrm{vec}(\mathbf{A}_S)\|_1)\mathrm{d}\mathbf{A}_S.
$$

Clearly, $\Pi_{\mathbf{A}_S}$ is the normalized restriction of $\Pi^L_{\mathbf{A}_S}$ on $\{\mathbf{A}_S \in \mathbb{R}^{|S|\times r} : \|\mathbf{A}_S\|_2 < 1\}$. Furthermore, given $S = S_0$ drawn from $\pi_S(S)$, for any $\mathbf{A}_{S_0} \in \{\|\mathbf{A}_{S_0} - \mathbf{A}_{0S_0}\|_{\mathrm{F}} < \eta_n/(32\|\boldsymbol{\Omega}_0\|_2)\}$ with $\eta_n \to 0$, we have

$$
\|\mathbf{A}_{S_0}\|_2 \leq \|\mathbf{A}_{0S_0}\|_2 + \|\mathbf{A}_{S_0} - \mathbf{A}_{0S_0}\|_{\mathrm{F}} \leq \sup_{n\geq 1}\|\mathbf{A}_0\|_2 + o(1) < 1.
$$

This implies that

$$
\left\{\mathbf{A}_{S_0} : \|\mathbf{A}_{S_0} - \mathbf{A}_{0S_0}\|_{\mathrm{F}} < \frac{\eta_n}{32\|\boldsymbol{\Omega}_0\|_2}\right\} \subset \{\mathbf{A}_{S_0} : \|\mathbf{A}_{S_0}\|_2 < 1\}.
$$

Then for sufficiently large n, we provide the following lower bound the first factor by restricting S to be S_0:

$$
\begin{aligned}
&\Pi_{\mathbf{A}}\left(\|\mathbf{A} - \mathbf{A}_0\|_{\mathrm{F}} < \frac{\eta_n}{32\|\boldsymbol{\Omega}_0\|_2}\right) \\
&\quad\geq \Pi_S(S_0)\Pi_{\mathbf{A}_{S_0}}\left(\|\mathbf{A}_{S_0} - \mathbf{A}_{0S_0}\|_{\mathrm{F}} < \frac{\eta_n}{32\|\boldsymbol{\Omega}_0\|_2}\right) \\
&\quad= \Pi_S(S_0)\frac{\Pi^L_{\mathbf{A}_{S_0}}\{\|\mathbf{A}_{S_0} - \mathbf{A}_{0S_0}\|_{\mathrm{F}} < \eta_n/(32\|\boldsymbol{\Omega}_0\|_2), \|\mathbf{A}_{S_0}\|_2 < 1\}}{\Pi^L_{\mathbf{A}_{S_0}}(\mathbf{A}_{S_0} : \|\mathbf{A}_{S_0}\|_2 < 1)} \\
&\quad\geq \Pi_S(S_0)\Pi^L_{\mathbf{A}_{S_0}}\left(\|\mathbf{A}_{S_0} - \mathbf{A}_{0S_0}\|_{\mathrm{F}} < \frac{\eta_n}{32\|\boldsymbol{\Omega}_0\|_2}\right) \\
&\quad\geq \frac{\pi_p(s_0)}{\binom{p-r}{s_0}}\mathrm{vol}\{B_2(\mathbf{0}_{s_0 r}, 1)\}\left(\frac{\eta_n}{32\|\boldsymbol{\Omega}_0\|_2}\right)^{s_0 r} \\
&\quad\quad\times \exp\left\{-2\max_{\|\mathbf{A}_{S_0} - \mathbf{A}_{0S_0}\|_{\mathrm{F}} < 1}(\|\mathrm{vec}(\mathbf{A}_{0S_0})\|_1 + \|\mathrm{vec}(\mathbf{A}_{S_0} - \mathbf{A}_{0S_0})\|_1\right\} \\
&\quad\geq \frac{\pi_p(s_0)}{\binom{p-r}{s_0}}\frac{1}{2\sqrt{s_0 r\pi}}\left(\frac{\sqrt{2\pi e}\eta_n}{32\|\boldsymbol{\Omega}_0\|_2\sqrt{s_0 r}}\right)^{s_0 r}\exp\left\{-2\sqrt{s_0 r}\left(\|\mathbf{A}_0\|_{\mathrm{F}} + 1\right)\right\} \\
&\quad\geq \frac{\pi_p(s_0)}{\binom{p-r}{s_0}}\frac{1}{2\sqrt{s_0 r\pi}}\left(\frac{\sqrt{2\pi e}\eta_n}{32\|\boldsymbol{\Omega}_0\|_2\sqrt{s_0 r}}\right)^{s_0 r}\exp\left(-4\sqrt{s_0 r^2}\right)
\end{aligned}
$$

$$\geq \frac{\pi_p(s_0)}{\binom{p-r}{s_0}} \exp\left\{-Cs_0 r \left|\log \frac{\eta_n}{\|\mathbf{\Omega}_0\|_2\sqrt{s_0 r}}\right| - 4s_0 r\right\}.$$

Since for sufficiently large n, $z_n \geq 1/2$ and

$$\begin{aligned}\frac{\pi_p(s_0)}{\binom{p-r}{s_0}} &\geq \frac{1}{2} n^{-rs_0}(p-r)^{-as_0}\left(\frac{s_0}{p-r}\right)^{2s_0} \geq \frac{1}{2} n^{-rs_0}(p-r)^{-as_0}(p-r)^{-2s_0} \\ &\geq \frac{1}{2}\exp(-rs_0\log n - cs_0\log p)\end{aligned}$$

for some constant $c > 0$, it follows that

$$\begin{aligned}&\Pi_{\mathbf{A}}\left(\|\mathbf{A}-\mathbf{A}_0\|_{\mathrm{F}} < \frac{\eta_n}{32\|\mathbf{\Omega}_0\|_2}\right) \\ &\quad\geq \exp\left\{-Crs_0\left|\log\frac{\eta_n}{\|\mathbf{\Omega}_0\|_2\sqrt{s_0 r}}\right| - C(rs_0\log n + s_0\log p)\right\}.\end{aligned}$$

Hence, using the fact that $n\eta_n^2 \to \infty$, we conclude that

$$\Pi_{\boldsymbol{\theta}}\left\{\|\mathbf{\Omega}(\boldsymbol{\theta}) - \mathbf{\Omega}_0\|_{\mathrm{F}} < \eta_n\right\} \geq \exp\left\{-C(\|\mathbf{\Omega}_0\|_2) rs_0 - Cs_0\log p\right\}$$

for some constant $C(\|\mathbf{\Omega}_0\|_2) > 0$ that only depends on $\|\mathbf{\Omega}_0\|_2$. The proof is thus completed. □

5.2. *Posterior sparsity*

In this subsection, we aim at establishing Lemma 5.4 regarding the posterior sparsity of $\mathbf{A}$ given the observed data, which in turn depends on Lemma 5.3 that characterizes the prior sparsity of $\mathbf{A}$.

Lemma 5.3. *Assume the prior specification in Section 3.1 and conditions A1–A5 hold. Then, for any constant $\kappa \geq 1$, there exists some constant $C > 0$ such that*

$$\Pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} : |\mathrm{supp}(\mathbf{A})| > \kappa s_0) \lesssim \exp\{-C\kappa(rs_0\log n + s_0\log p)\}.$$

Proof of Lemma 5.3. Write

$$\begin{aligned}&\Pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} : |\mathrm{supp}(\mathbf{A})| > \kappa s_0) \\ &\quad= \sum_{|S|=\lfloor\kappa s_0\rfloor}^{p-r} \pi_p(|S|) = \frac{1}{z_n}\sum_{t=\lfloor\kappa s_0\rfloor}^{p-r}\exp\{-rt\log n - at\log(p-r)\} \\ &\quad\leq 2\sum_{t=\lfloor\kappa s_0\rfloor}^{p-r}\exp(-rs_0\log n - Cs_0\log p) \\ &\quad\lesssim \exp\left(-\kappa rs_0\log n - \frac{C}{2}\kappa s_0\log p\right)\sum_{t=\lfloor\kappa s_0\rfloor}^{p-r}\exp\left(-\frac{C}{2}s_0\log p\right)\end{aligned}$$

$$\lesssim \exp\{-C\kappa(rs_0 \log n + s_0 \log p)\}$$

for some absolute constant $C > 0$. The proof is thus completed. □

Lemma 5.4. *Assume the prior specification in Section 3.1 and conditions A1–A5 hold. Then there exist some constants $\kappa_0 \geq 1$, $c > 0$ depending on $\|\mathbf{\Omega}_0\|_2$, such that*

$$\mathbb{E}_0\Pi_{\boldsymbol{\theta}}\left\{\boldsymbol{\theta} : |\mathrm{supp}(\mathbf{A})| > \kappa_0 s_0 \;\middle|\; \mathbf{Y}_n\right\} \leq 3e^{-c(rs\log n + s\log p)}.$$

Proof of Lemma 5.4. Let

$$\Xi_n := \{D_n \geq \exp\{-Cn\eta_n^2 \log(2\|\mathbf{\Omega}_0\|_2)\}\Pi_{\boldsymbol{\theta}}\{\boldsymbol{\theta} : \|\mathbf{\Omega}(\boldsymbol{\theta}) - \mathbf{\Omega}_0\|_{\mathrm{F}} < \eta_n\}\},$$

where $\eta_n := \sqrt{(rs\log n + s\log p)/n}$. By Lemma 5.2, $\mathbb{P}_0(\Xi_n^c) \leq 2e^{-cn\eta_n^2}$ for some constant $c > 0$ depending on $\|\mathbf{\Omega}_0\|_2$, and by Lemma 5.1,

$$\Pi_{\boldsymbol{\theta}}\{\|\mathbf{\Omega}(\boldsymbol{\theta}) - \mathbf{\Omega}_0\|_{\mathrm{F}} < \eta_n\} \geq \exp(-C_0 rs_0 \log n + Cs_0 \log p)$$

for some constant C_0 depending on $\|\mathbf{\Omega}_0\|_2$. Therefore, over the event Ξ_n, we have

$$\begin{aligned} D_n &\geq \exp\{-Cn\eta_n^2 \log(2\|\mathbf{\Omega}_0\|_2) - C_0 rs_0 \log n - Cs_0 \log p\} \\ &\geq \exp\{-C_0(rs_0 \log n + s_0 \log p)\} = \exp(-C_0 n\eta_n^2) \end{aligned}$$

for some constant C_0 depending on $\|\mathbf{\Omega}_0\|_2$. Hence, by Lemma 5.3 and the Fubini's theorem, we have,

$$\begin{aligned} &\mathbb{E}_0\Pi_{\boldsymbol{\theta}}\left\{\boldsymbol{\theta} : |\mathrm{supp}(\mathbf{A})| > \kappa_0 s_0 \;\middle|\; \mathbf{Y}_n\right\} \\ &\quad\leq \mathbb{E}_0\Pi_{\boldsymbol{\theta}}\left\{\boldsymbol{\theta} : |\mathrm{supp}(\mathbf{A})| > \kappa_0 s_0 \;\middle|\; \mathbf{Y}_n\right\}\mathbb{1}(\Xi_n) + \mathbb{P}_0(\Xi_n^c) \\ &\quad\leq e^{C_0 n\eta_n^2}\mathbb{E}_0\int_{\{\boldsymbol{\theta}:|\mathrm{supp}(\mathbf{A})|>\kappa_0 s_0\}} \exp\{\ell(\mathbf{\Omega}(\boldsymbol{\theta})) - \ell(\mathbf{\Omega}_0)\}\Pi_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{\theta}) + \mathbb{P}_0(\Xi_n^c) \\ &\quad= e^{C_0 n\eta_n^2}\int_{\{\boldsymbol{\theta}:|\mathrm{supp}(\mathbf{A})|>\kappa_0 s_0\}} \mathbb{E}_0\exp\{\ell(\mathbf{\Omega}(\boldsymbol{\theta})) - \ell(\mathbf{\Omega}_0)\}\Pi_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{\theta}) + \mathbb{P}_0(\Xi_n^c) \\ &\quad= e^{C_0 n\eta_n^2}\Pi_{\boldsymbol{\theta}}\left\{\boldsymbol{\theta} : |\mathrm{supp}(\mathbf{A})| > \kappa_0 s_0\right\} + \mathbb{P}_0(\Xi_n^c) \\ &\quad\lesssim \exp\left\{C_0 n\eta_n^2 - C\kappa_0\left(rs_0 \log n + s_0 \log p\right)\right\} + \mathbb{P}_0(\Xi_n^c) \\ &\quad\leq 2\exp\left(C_0 n\eta_n^2 - C\kappa_0 n\eta_n^2\right) + \exp(-crs\log n - cs\log p). \end{aligned}$$

Therefore, we conclude that

$$\mathbb{E}_0\Pi_{\boldsymbol{\theta}}\left\{\boldsymbol{\theta} : |\mathrm{supp}(\mathbf{A})| > \kappa_0 s_0 \;\middle|\; \mathbf{Y}_n\right\} \leq 3e^{-c(rs\log n + s\log p)}$$

by taking κ_0, possibly depending on $\|\mathbf{\Omega}_0\|_2$, to be sufficiently large. □

5.3. Construction of test functions

In this section, we construct a test function that will be useful for deriving posterior contraction under the Frobenius norm through Lemma 5.5 below.

Lemma 5.5. *Assume the random vectors $\mathbf{y}_1, \dots, \mathbf{y}_n$ follows $\mathrm{N}_p(\mathbf{0}_p, \boldsymbol{\Omega})$ independently, where $\boldsymbol{\Omega} := \mathbf{U}\mathbf{M}\mathbf{U}^{\mathrm{T}} + \mathbf{I}_p$, $\mathbf{U} \in \mathbb{O}_+(p, r)$, and $\mathbf{M} \in \mathbb{M}_+(r)$. Let $\boldsymbol{\Omega}_0 := \mathbf{U}_0\mathbf{M}_0\mathbf{U}_0^{\mathrm{T}} + \mathbf{I}_p$, where $\mathbf{U}_0 \in \mathbb{O}_+(p, r)$ with $|\mathrm{supp}\{\mathbf{A}(\mathbf{U}_0)\}| \leq s_0$ and $\mathbf{M}_0 \in \mathbb{M}_+(r)$. If $(\epsilon_n)_{n=1}^{\infty}$ is a sequence converging to 0, then for any $\kappa \geq 1$ and $M > 4$, there exists a sequence of test functions $(\phi_n)_{n=1}^{\infty}$ such that*

$$\mathbb{E}_{\boldsymbol{\Omega}_0}\phi_n \leq 3\exp\left\{(2C+4)\kappa s_0 \log p - \frac{CM^2 n\epsilon_n^2}{4\|\boldsymbol{\Omega}_0\|_2^2}\right\},$$
$$\sup_{\boldsymbol{\Omega}\in H_1} \mathbb{E}_{\boldsymbol{\Omega}}(1-\phi_n) \leq \exp\left(2C\kappa s_0 - \frac{CMn\epsilon_n^2}{8\|\boldsymbol{\Omega}_0\|_2^2}\right),$$

where

$$H_1 := \{\boldsymbol{\Omega} = \mathbf{U}\mathbf{M}\mathbf{U}^{\mathrm{T}} + \mathbf{I}_p : \|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_{\mathrm{F}} > M\epsilon_n, \mathbf{U} \in \mathbb{O}_+(p,r), \mathbf{M} \in \mathbb{M}_+(r), |\mathrm{supp}\{\mathbf{A}(\mathbf{U})\}| \leq \kappa s\}$$

and C is some absolute constant.

The proof of Lemma 5.5 relies on the oracle testing lemma from [21] below.

Lemma 5.6 (Lemma 5.7 in [21])**.** *Let the random vectors $\mathbf{y}_1, \dots, \mathbf{y}_n$ follow $\mathrm{N}_d(\mathbf{0}_d, \boldsymbol{\Omega})$ independently, where $\boldsymbol{\Omega} \in \mathbb{R}^{d\times d}$. If $(\epsilon_n)_{n=1}^{\infty}$ is a sequence converging to 0, then for any $M > 0$ and $d \times d$ covariance matrices $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$, there exists a test function ϕ_n such that*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\Omega}^{(1)}}\phi_n &\leq \exp\left(Cd - \frac{CM^2 n\epsilon_n^2}{4\|\boldsymbol{\Omega}^{(1)}\|_2^2}\right) + 2\exp\left(Cd - C\sqrt{M}n\right),\\ \sup_{\{\boldsymbol{\Omega}^{(2)}:\|\boldsymbol{\Omega}^{(2)}-\boldsymbol{\Omega}^{(1)}\|_2 > M\epsilon_n\}} &\mathbb{E}_{\boldsymbol{\Omega}^{(2)}}(1-\phi_n)\\ &\leq \exp\left[Cd - \frac{CMn\epsilon_n^2}{4}\left\{1 \vee \frac{M}{(\sqrt{M}+2)^2\|\boldsymbol{\Omega}^{(1)}\|_2^2}\right\}\right]\end{aligned} \tag{22}$$

with some absolute constant $C > 0$.

Proof of Lemma 5.5. The proof of Lemma 5.5 is very similar to that of Lemma 5.4 in [21] and is included here for completeness. Decompose H_1 by

$$H_1 \subset \bigcup_{S:|S|\leq\kappa s_0} H_{1S},$$

where

$$H_{1S} := \{\boldsymbol{\Omega} = \mathbf{U}\mathbf{M}\mathbf{U}^{\mathrm{T}} + \mathbf{I}_p : \mathbf{U} \in \mathbb{O}_+(p,r), \mathbf{M} \in \mathbb{M}_+(r), \|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_{\mathrm{F}} > M\epsilon_n,$$

$$S = \text{supp}\{\mathbf{A}(\mathbf{U})\}\}.$$

Let $\bar{S} := S \cup S_0$, where $S_0 := \text{supp}\{\mathbf{A}(\mathbf{U}_0)\}$, and let $\bar{s} := |\bar{S}|$. Clearly, $\bar{s} \leq (\kappa + 1)s_0$ and

$$\|\mathbf{\Omega} - \mathbf{\Omega}_0\|_{\text{F}} = \|\bar{\mathbf{\Omega}} - \bar{\mathbf{\Omega}}_0\|_{\text{F}},$$

where

$$\bar{\mathbf{\Omega}} := \mathbf{U}(\mathbf{A}_{\bar{S}})\mathbf{M}\mathbf{U}(\mathbf{A}_{\bar{S}})^{\text{T}} + \mathbf{I}_{\bar{s}+r}, \quad \bar{\mathbf{\Omega}}_0 := \mathbf{U}(\mathbf{A}_{0\bar{S}})\mathbf{M}_0\mathbf{U}(\mathbf{A}_{0\bar{S}})^{\text{T}} + \mathbf{I}_{\bar{s}+r}.$$

For each $S \subset [p]$, denote $\mathbf{y}_{iS} := [y_{ij} : j \in S]^{\text{T}}$ for $i = 1, \ldots, n$. By Lemma 5.6, for each S and $M > 4$, there exists a sequence of tests $(\phi_{nS})_{n=1}^{\infty}$, where ϕ_{nS} is a measurable function of $\{\mathbf{y}_{1\text{supp}\{\mathbf{U}(\mathbf{A}_{\bar{S}})\}}, \ldots, \mathbf{y}_{n\text{supp}\{\mathbf{U}(\mathbf{A}_{\bar{S}})\}}\}$, such that

$$\begin{aligned}
\mathbb{E}_{\bar{\mathbf{\Omega}}_0}\phi_{nS} &\leq \exp\left\{C(\kappa+1)s_0 - \frac{CM^2n\epsilon_n^2}{4\|\mathbf{\Omega}_0\|_2^2}\right\} + 2\exp\{C(\kappa+1)s_0 - C\sqrt{M}n\} \\
&\leq 3\exp\left(2C\kappa s_0 - \frac{CM^2n\epsilon_n^2}{4\|\mathbf{\Omega}_0\|_2^2}\right),
\end{aligned}$$

and

$$\begin{aligned}
\sup_{\bar{\mathbf{\Omega}} \in \bar{H}_{1S}} \mathbb{E}_{\bar{\mathbf{\Omega}}}(1 - \phi_n) &\leq \exp\left[C(\kappa+1)s_0 - \frac{CMn\epsilon_n^2}{4}\left\{1 \vee \frac{M}{(\sqrt{M}+2)^2\|\mathbf{\Omega}_0\|_2^2}\right\}\right] \\
&\leq \exp\left(2C\kappa s_0 - \frac{CMn\epsilon_n^2}{8\|\mathbf{\Omega}_0\|_2^2}\right),
\end{aligned}$$

where

$$\bar{H}_{1S} := \left\{\bar{\mathbf{\Omega}} = \mathbf{U}(\mathbf{A}_{\bar{S}})\mathbf{M}\mathbf{U}(\mathbf{A}_{\bar{S}})^{\text{T}} + \mathbf{I}_{\bar{s}+r} : \mathbf{M} \in \mathbb{M}_+(r), \|\bar{\mathbf{\Omega}} - \bar{\mathbf{\Omega}}_0\|_{\text{F}} > M\epsilon_n\right\}.$$

Hence we can combine tests by taking $\phi_n := \max_S \phi_{nS}$ and apply the union bound to obtain

$$\begin{aligned}
\mathbb{E}_0\phi_n &\leq \sum_{s=1}^{\lceil \kappa s_0 \rceil} \binom{p-r}{s} 3\exp\left(2C\kappa s_0 - \frac{CM^2n\epsilon_n^2}{4\|\mathbf{\Omega}_0\|_2^2}\right) \\
&\leq 3\kappa s_0 \exp(3\kappa s_0 \log p)\exp\left(2C\kappa s_0 - \frac{CM^2n\epsilon_n^2}{4\|\mathbf{\Omega}_0\|_2^2}\right) \\
&\leq 3\exp\left\{(2C+4)\kappa s_0 \log p - \frac{CM^2n\epsilon_n^2}{4\|\mathbf{\Omega}_0\|_2^2}\right\},
\end{aligned}$$

and

$$\sup_{\mathbf{\Omega} \in H_1} \mathbb{E}_{\mathbf{\Omega}}(1 - \phi_n) \leq \sup_{S:|S| < \kappa s_0} \sup_{\bar{\mathbf{\Omega}} \in \bar{H}_{1S}} \mathbb{E}_{\bar{\mathbf{\Omega}}}(1 - \phi_{nS}) \leq \exp\left(2C\kappa s_0 - \frac{CMn\epsilon_n^2}{8\|\mathbf{\Omega}_0\|_2^2}\right).$$

The proof is thus completed. □

5.4. Posterior contraction under Frobenius norm

Theorem 5.7. *Assume the setup and the prior specification in Section 3.1 and suppose conditions A1–A5 hold. Then there exists some large constant $M_0 > 0$ and constant $c > 0$ that possibly depend on $\|\mathbf{\Omega}_0\|_2$ but is independent of r, s, p, n, such that*

$$\mathbb{E}_0\Pi_{\boldsymbol{\theta}}\left(\|\mathbf{\Omega}(\boldsymbol{\theta}) - \mathbf{\Omega}_0\|_{\mathrm{F}} > M\sqrt{\frac{rs_0\log n + s_0\log p}{n}}\;\middle|\;\mathbf{Y}_n\right) \lesssim e^{-c(rs\log n + s\log p)}.$$

Proof of Theorem 5.7. Denote $\epsilon_n := \sqrt{(rs_0\log n + s_0\log p)/n}$. We first decompose the expected posterior probability by

$$\begin{aligned}
&\mathbb{E}_0\Pi_{\boldsymbol{\theta}}\{\|\mathbf{\Omega}(\boldsymbol{\theta}) - \mathbf{\Omega}_0\|_{\mathrm{F}} > M\epsilon_n \mid \mathbf{Y}_n\}\\
&\quad\le \mathbb{E}_0\Pi_{\boldsymbol{\theta}}\left\{\|\mathbf{\Omega}(\boldsymbol{\theta}) - \mathbf{\Omega}_0\|_{\mathrm{F}} > M\epsilon_n, |\mathrm{supp}(\mathbf{A})| \le \kappa_0 s_0\;\middle|\;\mathbf{Y}_n\right\}\\
&\qquad + \mathbb{E}_0\Pi_{\boldsymbol{\theta}}\left\{\boldsymbol{\theta} : |\mathrm{supp}(\mathbf{A})| > \kappa_0 s_0 \mid \mathbf{Y}_n\right\}
\end{aligned}$$

where κ_0 is set to be large enough such that the second term on the right-hand side is $O(e^{-cn\epsilon_n^2})$ for some constant $c > 0$ (possibly depending on $\|\mathbf{\Omega}_0\|_2$) according to Lemma 5.4. It suffices to focus on the first term consequently. Let $\Xi_n := \{D_n \ge \exp(-C_0 n\epsilon_n^2)\}$, where C_0 is a constant depending on $\|\mathbf{\Omega}_0\|_2$ such that $\mathbb{P}_0(\Xi_n^c) = o(1)$ according to Lemma 5.2 and Lemma 5.1. Take ϕ_n to be the test function given by Lemma 5.5. Then, we can decompose the first term on the right-hand side of the previous display by

$$\begin{aligned}
&\mathbb{E}_0\Pi_{\boldsymbol{\theta}}\left\{\|\mathbf{\Omega}(\boldsymbol{\theta}) - \mathbf{\Omega}_0\|_{\mathrm{F}} > M\epsilon_n, |\mathrm{supp}(\mathbf{A})| \le \kappa_0 s_0\;\middle|\;\mathbf{Y}_n\right\}\\
&\quad\le \mathbb{E}_0\Pi_{\boldsymbol{\theta}}\left\{\|\mathbf{\Omega}(\boldsymbol{\theta}) - \mathbf{\Omega}_0\|_{\mathrm{F}} > M\epsilon_n, |\mathrm{supp}(\mathbf{A})| \le \kappa_0 s_0\;\middle|\;\mathbf{Y}_n\right\}\mathbb{1}(\Xi_n)(1-\phi_n)\\
&\qquad + \mathbb{E}_0\phi_n + \mathbb{P}_0(\Xi_n^c).
\end{aligned}$$

Since the second term on the right-hand side is upper bounded by

$$3\exp\left\{(2C+4)\kappa_0 s_0\log p - \frac{CM^2 n\epsilon_n^2}{4\|\mathbf{\Omega}_0\|_2^2}\right\} \le 3\exp\left(-\frac{CM^2 n\epsilon_n^2}{8\|\mathbf{\Omega}_0\|_2^2}\right)$$

by Lemma 5.5 with a sufficiently large $M > 0$, and the third term is also $O(e^{-cn\epsilon_n^2})$ by Lemma 5.2 and Lemma 5.1, it suffices to show that the first term is also $o(1)$. Denote

$$\begin{aligned}
H_1 := \{&\mathbf{\Omega} = \mathbf{U}\mathbf{M}\mathbf{U}^{\mathrm{T}} + \mathbf{I}_p : \mathbf{U} \in \mathbb{O}_+(p,r), \mathbf{M} \in \mathbb{M}_+(r), \|\mathbf{\Omega} - \mathbf{\Omega}_0\|_{\mathrm{F}} > M\epsilon_n,\\
&|\mathrm{supp}\{\mathbf{A}(\mathbf{U})\}| \le \kappa_0 s_0\}.
\end{aligned}$$

Then by Lemma 5.5, the Fubini's theorem, and the definition of Ξ_n,

$$\mathbb{E}_0\Pi_{\boldsymbol{\theta}}\left\{\|\mathbf{\Omega}(\boldsymbol{\theta}) - \mathbf{\Omega}_0\|_{\mathrm{F}} > M\epsilon_n, |\mathrm{supp}(\mathbf{A})| \le \kappa_0 s_0\;\middle|\;\mathbf{Y}_n\right\}\mathbb{1}(\Xi_n)(1-\phi_n)$$

$$
\begin{aligned}
&\le \exp(C_0 n\epsilon_n^2)\mathbb{E}_0(1-\phi_n)\int_{\{\boldsymbol{\theta}:\boldsymbol{\Omega}(\boldsymbol{\theta})\in H_1\}}\exp\{\ell(\boldsymbol{\Omega}(\boldsymbol{\theta}))-\ell(\boldsymbol{\Omega}_0)\}\Pi_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{\theta})\\
&= \exp(C_0 n\epsilon_n^2)\int_{\boldsymbol{\theta}:\boldsymbol{\Omega}(\boldsymbol{\theta})\in H_1}\mathbb{E}_0(1-\phi_n)\exp\{\ell(\boldsymbol{\Omega}(\boldsymbol{\theta}))-\ell(\boldsymbol{\Omega}_0)\}\Pi_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{\theta})\\
&= \exp(C_0 n\epsilon_n^2)\int_{\{\boldsymbol{\theta}:\boldsymbol{\Omega}(\boldsymbol{\theta})\in H_1\}}\mathbb{E}_{\boldsymbol{\Omega}}(1-\phi_n)\Pi(\mathrm{d}\boldsymbol{\Omega})\\
&\le \exp(C_0 n\epsilon_n^2)\sup_{\boldsymbol{\Omega}\in H_1}\mathbb{E}_{\boldsymbol{\Omega}}(1-\phi_n)\\
&\le \exp\left(C_0 n\epsilon_n^2+2C\kappa_0 s_0-\frac{CMn\epsilon_n^2}{8\|\boldsymbol{\Omega}_0\|_2^2}\right)\le \exp\left(-\frac{CMn\epsilon_n^2}{16\|\boldsymbol{\Omega}_0\|_2^2}\right)
\end{aligned}
$$

by taking M to be suffciently large enough. The proof is thus completed. □

5.5. *Local asymptotic normality*

In this subsection, we establish the local asymptotic normality of the spiked covariance model under the sparsity constraint through Theorem 5.8 below. Some preliminaries are needed to prove this theorem. Define

$$
\mathcal{A}_n := \{\boldsymbol{\theta} : \|\boldsymbol{\Omega}(\boldsymbol{\theta})-\boldsymbol{\Omega}_0\|_{\mathrm{F}} < M\epsilon_n, |\mathrm{supp}(\mathbf{A})|\le\kappa_0 s_0\}\,, \tag{23}
$$

where $\epsilon_n := \sqrt{(rs\log n + s\log p)/n}$. By Theorem 5.7 and Lemma 5.4, there exists some constants $M,\kappa_0,c>0$, possibly depending on $\|\boldsymbol{\Omega}_0\|_2$, such that

$$
\mathbb{E}_0\Pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}\in\mathcal{A}_n \mid \mathbf{Y}_n)\ge 1-O(e^{-cn\epsilon_n^2}).
$$

Under the assumption that $\sup_{n\ge1}(\mathbf{A}_0)$ is bounded away from 1, by Theorem 3.2 in [55],

$$
\mathcal{A}_n \subset \{\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2 \le M'\epsilon_n, |\mathrm{supp}(\mathbf{A})|\le\kappa_0 s_0\}
$$

for some large constant $M'>0$. Note that with a slight abuse of notation, we may use M to denote a generic constant that is sufficiently large such that we can write

$$
\mathcal{A}_n \subset \{\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2 \le M\epsilon_n, |\mathrm{supp}(\mathbf{A})|\le\kappa_0 s_0\}
$$

and $\mathcal{A}_n$ still satisfies $\mathbb{E}_0\Pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}\in\mathcal{A}_n \mid \mathbf{Y}_n)\ge 1-O(e^{-cn\epsilon_n^2})$. For any $S\subset[p-r]$ with $|S|\le\kappa_0 s_0$, let

$$
\mathcal{A}_n(S) := \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1 \le M\sqrt{\frac{r^2 s_0^2\log n + rs_0^2\log p}{n}}, \mathrm{supp}(\mathbf{A}) = S\right\}
$$

for some large constant $M>0$. It follows that

$$
\mathcal{A}_n \subset \mathcal{B}_n := \bigcup_{S:|S|\le\kappa_0 s_0}\mathcal{A}_n(S)
$$

This is because for all $\boldsymbol{\theta} = [\mathrm{vec}(\mathbf{A})^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{T}}]^{\mathrm{T}} \in \mathcal{A}_n$ with $\mathrm{supp}(\mathbf{A}) = S$, $|S| \leq \kappa_0 s_0$, we have

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1 \leq \sqrt{(|S| + |S_0|)r + 2r^2 + r(r+1)}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \lesssim \sqrt{(\kappa_0 + 1)rs_0}\sqrt{\frac{rs_0 \log n + s_0 \log p}{n}} \lesssim \sqrt{\frac{r^2 s_0^2 \log n + rs_0^2 \log p}{n}}.$$

Theorem 5.8. *Let* $\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}_0)$ *with* $\boldsymbol{\Omega}_0 := \boldsymbol{\Omega}(\boldsymbol{\theta}_0)$*, and* $\ell(\boldsymbol{\Omega}(\boldsymbol{\theta}))$ *be the log-likelihood function of* $\boldsymbol{\theta}$*. Suppose conditions A1–A5 hold. Then, the* $\ell(\boldsymbol{\Omega}(\boldsymbol{\theta}))$ *yields the following local asymptotic normality expansion:*

$$\begin{aligned}\ell(\boldsymbol{\Omega}(\boldsymbol{\theta})) - \ell(\boldsymbol{\Omega}_0) =& \frac{n}{2}\mathrm{vec}\left(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\right)^{\mathrm{T}} (\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1}) D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &- \frac{n}{4}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}} (\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1}) D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &+ R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0),\end{aligned}$$

where $\widehat{\boldsymbol{\Omega}} := (1/n)\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^{\mathrm{T}}$ *denotes the sample covariance matrix, and the remainder* R_n *satisfies*

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_n} |R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| \leq C\sqrt{\frac{(r^2 s^2 \log n + rs^2 \log p)^3}{n}}$$

with probability at least $1 - 2e^{-\kappa_0 n \epsilon_n^2}$ *for some constant* $C > 0$ *depending on* $\|\boldsymbol{\Omega}_0\|_2$.

The key to the proof of the local asymptotic normality expansion in Theorem 5.8 is the following lemma that controls the stochastic remainder in the Taylor expansion of the log-likelihood function. For convenience, we denote by

$$\begin{aligned}\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0) &:= \mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}(\boldsymbol{\varphi}_0) - 2(\mathbf{I}_p - \mathbf{X}_0)^{-1}(\mathbf{X}_{\boldsymbol{\varphi}} - \mathbf{X}_0)(\mathbf{I}_p - \mathbf{X}_0)^{-1}\mathbf{I}_{p \times r}, \\ \mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) &:= \mathbf{U}(\boldsymbol{\varphi})(\mathbf{M} - \mathbf{M}_0)\{\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\}^{\mathrm{T}} \\ &\quad + \{\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\}\mathbf{M}_0\{\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\}^{\mathrm{T}} \\ &\quad + \{\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\}(\mathbf{M} - \mathbf{M}_0)\mathbf{U}_0^{\mathrm{T}}, \\ \mathbf{R}_1(\boldsymbol{\Omega}, \boldsymbol{\Omega}_0) &:= \boldsymbol{\Omega}_0^{-1}(\boldsymbol{\Omega}_0 - \boldsymbol{\Omega})\boldsymbol{\Omega}_0^{-1}\sum_{m=1}^{\infty}\{(\boldsymbol{\Omega}_0 - \boldsymbol{\Omega})\boldsymbol{\Omega}_0^{-1}\}^m, \quad \|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 < 1.\end{aligned}$$

Lemma 5.9. *Let* $\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}_0)$ *with* $\boldsymbol{\Omega}_0 := \boldsymbol{\Omega}(\boldsymbol{\theta}_0)$*. Suppose conditions A1–A5 hold. Then, with probability at least* $1 - 2e^{-\kappa_0 n \epsilon_n^2}$ *the following stochastic remainders satisfy*

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_n} \left|2n\mathrm{vec}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1})\mathrm{vec}\{\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0)\mathbf{M}_0\mathbf{U}_0^{\mathrm{T}}\}\right| \leq C\delta_n, \quad (24)$$

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}_n} \left|n\mathrm{vec}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1})\mathrm{vec}\{\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\}\right| \leq C\delta_n, \quad (25)$$

$$\sup_{\boldsymbol{\theta}\in\mathcal{B}_n}\left|n\mathrm{vec}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)^{\mathrm{T}}\mathrm{vec}\{\mathbf{R}_1(\boldsymbol{\Omega}(\boldsymbol{\theta}),\boldsymbol{\Omega}_0)\}\right|\leq C\delta_n, \tag{26}$$

where $\widehat{\boldsymbol{\Omega}} := (1/n)\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}$ *denotes the sample covariance matrix,* $C>0$ *is a constant depending on* $\|\boldsymbol{\Omega}_0\|_2$*, and* $\delta_n=\sqrt{(r^2s^2\log n+rs^2\log p)^3/n}$*.*

Proof of Lemma 5.9. The proof is based on reducing the dimension of the deterministic remainders $\mathbf{R}_{\mathbf{U}}$, $\mathbf{R}_{\boldsymbol{\Sigma}}$, and $\mathbf{R}_1$ because, for $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$, the intrinsic dimension is much smaller than the ambient dimension due to the sparsity. We first fix $S\in\mathcal{S}(\kappa_0 s_0):=\{S\subset[p-r]:|S|\leq\kappa_0 s_0\}$. Let $\bar{S}:=S\cup S_0$, where $S_0:=\mathrm{supp}(\mathbf{A}_0)$ and $\boldsymbol{\theta}_0:=[\mathrm{vec}(\mathbf{A}_0)^{\mathrm{T}},\boldsymbol{\mu}_0^{\mathrm{T}}]^{\mathrm{T}}$, and let $\bar{s}:=|\bar{S}|$. Denote by

$$\mathbf{A}:=\mathbf{P}_S\begin{bmatrix}\mathbf{A}_{\bar{S}}\\ \mathbf{0}\end{bmatrix},\quad \boldsymbol{\varphi}_{\bar{S}}:=\mathrm{vec}(\mathbf{A}_{\bar{S}}),\quad \boldsymbol{\theta}_{\bar{S}}:=\begin{bmatrix}\boldsymbol{\varphi}_{\bar{S}}\\ \boldsymbol{\mu}\end{bmatrix}$$

for a suitable permutation matrix $\mathbf{P}_S$. Similarly, denote by

$$\mathbf{A}_0:=\mathbf{P}_S\begin{bmatrix}\mathbf{A}_{0\bar{S}}\\ \mathbf{0}\end{bmatrix},\quad \boldsymbol{\varphi}_{0\bar{S}}=\mathrm{vec}(\mathbf{A}_{0\bar{S}}),\quad \boldsymbol{\theta}_{0\bar{S}}=\begin{bmatrix}\boldsymbol{\varphi}_{0\bar{S}}\\ \boldsymbol{\mu}\end{bmatrix}.$$

By definition of the Cayley parameterization $\boldsymbol{\varphi}\mapsto\mathbf{U}(\boldsymbol{\varphi})$,

$$\mathbf{U}(\boldsymbol{\varphi})=\begin{bmatrix}(\mathbf{I}_r-\mathbf{A}_{\bar{S}}^{\mathrm{T}}\mathbf{A}_{\bar{S}})(\mathbf{I}_r+\mathbf{A}_{\bar{S}}^{\mathrm{T}}\mathbf{A}_{\bar{S}})^{-1}\\ \mathbf{P}_S\begin{bmatrix}\mathbf{A}_{\bar{S}}(\mathbf{I}_r+\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\\ \mathbf{0}\end{bmatrix}\end{bmatrix}=\begin{bmatrix}\mathbf{I}_r & \\ & \mathbf{P}_S\end{bmatrix}\begin{bmatrix}\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})\\ \mathbf{0}\end{bmatrix}$$

where $\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})$ is the Cayley parameterization that maps the vector $\boldsymbol{\varphi}_{\bar{S}}$ to a Stiefel matrix in $\in\mathbb{O}(\bar{s}+r,r)$. Write $\mathbf{Q}_S=\mathrm{diag}(\mathbf{I}_r,\mathbf{P}_S)$. Similarly, we can also write $\mathbf{U}_0=\mathbf{Q}_S[\mathbf{U}_{0\bar{S}}^{\mathrm{T}},\mathbf{0}]^{\mathrm{T}}$, where $\mathbf{U}_{0\bar{S}}:=\mathbf{U}(\boldsymbol{\varphi}_{0\bar{S}})$. The permutation matrix $\mathbf{Q}_S$ will be useful in this proof. For $\mathbf{R}_{\mathbf{U}}$, write

$$\mathbf{I}_p-\mathbf{X}_0=\mathbf{Q}_S\begin{bmatrix}\mathbf{I}-\mathbf{X}_{0\bar{S}} & \\ & \mathbf{I}_{p-(\bar{s}+r)}\end{bmatrix}\mathbf{Q}_S^{\mathrm{T}},$$

where

$$\mathbf{X}_{0\bar{S}}:=\mathbf{X}_{\boldsymbol{\varphi}_{0\bar{S}}}=\begin{bmatrix}\mathbf{0} & -\mathbf{A}_{0\bar{S}}^{\mathrm{T}}\\ \mathbf{A}_{0\bar{S}} & \mathbf{0}\end{bmatrix}.$$

Similarly, we also have

$$\mathbf{X}_{\boldsymbol{\varphi}}-\mathbf{X}_0=\mathbf{Q}_S\begin{bmatrix}\mathbf{X}_{\boldsymbol{\varphi}_{\bar{S}}}-\mathbf{X}_{0\bar{S}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{Q}_S^{\mathrm{T}},\quad\text{where}\quad \mathbf{X}_{\boldsymbol{\varphi}_{\bar{S}}}:=\begin{bmatrix}\mathbf{0} & -\mathbf{A}_{\bar{S}}^{\mathrm{T}}\\ \mathbf{A}_{\bar{S}} & \mathbf{0}\end{bmatrix}.$$

■ We first consider $\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi},\boldsymbol{\varphi}_0)$. Write $\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi},\boldsymbol{\varphi}_0)$ in the following block form with a zero matrix in the lower block:

$$\begin{aligned}\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi},\boldsymbol{\varphi}_0)&=\mathbf{Q}_S\begin{bmatrix}\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})-\mathbf{U}(\boldsymbol{\varphi}_{0\bar{S}})\\ \mathbf{0}\end{bmatrix}\\ &\quad-\mathbf{Q}_S\begin{bmatrix}(\mathbf{I}-\mathbf{X}_{0\bar{S}})^{-1}(\mathbf{X}_{\boldsymbol{\varphi}_{\bar{S}}}-\mathbf{X}_{0\bar{S}})(\mathbf{I}-\mathbf{X}_{0\bar{S}})^{-1}\mathbf{I}_{(\bar{s}+r)\times r}\\ \mathbf{0}\end{bmatrix}\end{aligned}$$

$$:= \mathbf{Q}_S \begin{bmatrix} \mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}_{\bar{S}}, \boldsymbol{\varphi}_{0\bar{S}}) \\ \mathbf{0} \end{bmatrix},$$

where we have used the fact that $\mathbf{Q}_S \mathbf{I}_{p\times r} = \mathbf{I}_{p\times r}$. Therefore,

$$\begin{aligned}
&\mathrm{vec}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1})\mathrm{vec}\{\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0)\mathbf{M}_0\mathbf{U}_0^{\mathrm{T}}\} \\
&\quad = \mathrm{tr}\left\{(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0)\mathbf{M}_0\mathbf{U}_0^{\mathrm{T}}\boldsymbol{\Omega}_0^{-1}\right\} \\
&\quad = \mathrm{tr}\left\{(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0)\tilde{\mathbf{M}}_0\mathbf{U}_0^{\mathrm{T}}\right\} \\
&\quad = \mathrm{tr}\left\{\mathbf{U}_0^{\mathrm{T}}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0)\tilde{\mathbf{M}}_0\right\},
\end{aligned}$$

where $\tilde{\mathbf{M}}_0 := \mathbf{M}_0(\mathbf{M}_0^{-1} + \mathbf{I})^{-1}$. Write $\widehat{\boldsymbol{\Omega}}$ and $\boldsymbol{\Omega}_0$ in the block forms

$$\widehat{\boldsymbol{\Omega}} = \mathbf{Q}_S \begin{bmatrix} \widehat{\boldsymbol{\Omega}}_{\bar{S}} & \widehat{\boldsymbol{\Omega}}_{12} \\ \widehat{\boldsymbol{\Omega}}_{21} & \widehat{\boldsymbol{\Omega}}_{22} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}}.$$

and

$$\begin{aligned}
\boldsymbol{\Omega}_0 &= \mathbf{Q}_S \left(\begin{bmatrix} \mathbf{U}_{0\bar{S}} \\ \mathbf{0} \end{bmatrix} \mathbf{M}_0 \begin{bmatrix} \mathbf{U}_{0\bar{S}}^{\mathrm{T}} & \mathbf{0} \end{bmatrix} + \mathbf{I}_p \right) \mathbf{Q}_S^{\mathrm{T}} \\
&= \mathbf{Q}_S \begin{bmatrix} \mathbf{U}_{0\bar{S}}\mathbf{M}_0\mathbf{U}_{0\bar{S}}^{\mathrm{T}} + \mathbf{I}_{\bar{s}+r} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-(\bar{s}+r)} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}} = \mathbf{Q}_S \begin{bmatrix} \boldsymbol{\Omega}_{0\bar{S}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-(\bar{s}+r)} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}},
\end{aligned}$$

where $\boldsymbol{\Omega}_{0\bar{S}} := \mathbf{U}_{0\bar{S}}\mathbf{M}_0\mathbf{U}_{0\bar{S}}^{\mathrm{T}} + \mathbf{I}_{\bar{s}+r}$. It follows that

$$\begin{aligned}
&\mathbf{U}_0^{\mathrm{T}}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0)\tilde{\mathbf{M}}_0 \\
&\quad = \begin{bmatrix} \mathbf{U}_{0\bar{S}}^{\mathrm{T}} & \mathbf{0} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S \begin{bmatrix} \widehat{\boldsymbol{\Omega}}_{\bar{S}} - \boldsymbol{\Omega}_{0\bar{S}} & \widehat{\boldsymbol{\Omega}}_{12} \\ \widehat{\boldsymbol{\Omega}}_{21} & \widehat{\boldsymbol{\Omega}}_{22} - \mathbf{I}_{p-(\bar{s}+r)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Omega}_{0\bar{S}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-(\bar{s}+r)} \end{bmatrix} \\
&\qquad \times \begin{bmatrix} \mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}_{\bar{S}}, \boldsymbol{\varphi}_{0\bar{S}}) \\ \mathbf{0} \end{bmatrix} \tilde{\mathbf{M}}_0 \\
&\quad = \mathbf{U}_{0\bar{S}}^{\mathrm{T}}(\widehat{\boldsymbol{\Omega}}_{\bar{S}} - \boldsymbol{\Omega}_{0\bar{S}})\boldsymbol{\Omega}_{0\bar{S}}^{-1}\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}_{\bar{S}}, \boldsymbol{\varphi}_{0\bar{S}})\tilde{\mathbf{M}}_0.
\end{aligned}$$

By the random matrix theory (see, for example, Section 5.4.1. in [51]), for any $t > 0$,

$$\mathbb{P}_0\left(\|\widehat{\boldsymbol{\Omega}}_{\bar{S}} - \boldsymbol{\Omega}_{0\bar{S}}\|_2 > C_0\sqrt{\frac{\kappa_0 s_0 + t^2}{n}} \right) \leq 2\exp(-ct^2)$$

for some absolute constant $c > 0$ and some constant $C_0 > 0$ that depends on $\|\boldsymbol{\Sigma}_0\|_2$, where $\kappa_n := \kappa_0 s_0$ and $\boldsymbol{\Sigma}_{0\bar{S}} := \mathbf{U}_{0\bar{S}}\mathbf{M}_0\mathbf{U}_{0\bar{S}}^{\mathrm{T}} + \mathbf{I}_{\bar{s}+r}$. Therefore, with t^2 replaced by $(2\kappa_0/c)(rs_0\log n + s_0\log p)$, we have,

$$\begin{aligned}
&\mathbb{P}_0\left(\sup_{S \in \mathcal{S}(\kappa_0 s_0)} \|\widehat{\boldsymbol{\Omega}}_{\bar{S}} - \boldsymbol{\Omega}_{0\bar{S}}\|_2 > C_0\sqrt{\frac{rs_0\log n + s_0\log p}{n}} \right) \\
&\quad \leq 2\binom{p-r}{\kappa_0 s_0} \exp\left\{-2\kappa_0(rs_0\log n + s_0\log p)\right\} \\
&\quad \leq 2\exp\{-\kappa_0(rs_0\log n + s_0\log p)\}
\end{aligned} \tag{27}$$

Denote $\|\mathbf{A}\|_* := \sum_i \sigma_i(\mathbf{A})$ the nuclear norm of a matrix. By Hölder's inequality, the equivalence between nuclear norm and Frobenius norm, and Theorem 2.1 in [55], the left-hand side of (24) is upper bounded by

$$\begin{aligned}
&\sup_{S\in\mathcal{S}(\kappa_0 s_0)}\sup_{\boldsymbol{\theta}\in\mathcal{A}_n(S)} 2n\|\widehat{\boldsymbol{\Omega}}_{\bar{S}} - \boldsymbol{\Omega}_{0\bar{S}}\|_2\|\mathbf{U}_{0\bar{S}}\tilde{\mathbf{M}}_0\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}_{\bar{S}},\boldsymbol{\varphi}_{0\bar{S}})^{\mathrm{T}}\boldsymbol{\Omega}_{0\bar{S}}^{-1}\|_*\\
&\leq 2n\sqrt{r}\sup_{S\in\mathcal{S}(\kappa_0 s_0)}\|\widehat{\boldsymbol{\Omega}}_{\bar{S}} - \boldsymbol{\Omega}_{0\bar{S}}\|_2\\
&\quad\times\sup_{S\in\mathcal{S}(\kappa_0 s_0)}\sup_{\boldsymbol{\theta}\in\mathcal{A}_n(S)}\|\mathbf{U}_{0\bar{S}}\tilde{\mathbf{M}}_0\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}_{\bar{S}},\boldsymbol{\varphi}_{0\bar{S}})^{\mathrm{T}}\boldsymbol{\Omega}_{0\bar{S}}^{-1}\|_{\mathrm{F}}\\
&\lesssim n\sqrt{\frac{r(rs_0\log n + s_0\log p)}{n}}\sup_{S\in\mathcal{S}(\kappa_0 s_0)}\sup_{\boldsymbol{\theta}\in\mathcal{A}_n(S)}\|\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}_{\bar{S}},\boldsymbol{\varphi}_{0\bar{S}})\|_{\mathrm{F}}\\
&\lesssim n\sqrt{\frac{r(rs_0\log n + s_0\log p)}{n}}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1^2 \lesssim \sqrt{\frac{(r^2s^2\log n + rs^2\log p)^3}{n}}
\end{aligned}$$

with probability greater than $1 - 2\exp(-\kappa_0 n\epsilon_n^2)$. This shows that the left-hand side of (24) is upper bounded by $C\delta_n$ with probability at least $1-2\exp(-\kappa_0 n\epsilon_n^2)$.

■ For $\mathbf{R}_{\boldsymbol{\Omega}}$, we have, using the permutation matrix $\mathbf{Q}_S$,

$$\begin{aligned}
&\mathbf{U}(\boldsymbol{\varphi})(\mathbf{M}-\mathbf{M}_0)\{\mathbf{U}(\boldsymbol{\varphi})-\mathbf{U}_0\}^{\mathrm{T}}\\
&\quad= \mathbf{Q}_S\begin{bmatrix}\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})(\mathbf{M}-\mathbf{M}_0)\{\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})-\mathbf{U}_{0\bar{S}}\}^{\mathrm{T}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{Q}_S^{\mathrm{T}},
\end{aligned}$$

$$\begin{aligned}
&\{\mathbf{U}(\boldsymbol{\varphi})-\mathbf{U}_0\}\mathbf{M}_0\{\mathbf{U}(\boldsymbol{\varphi})-\mathbf{U}_0\}^{\mathrm{T}}\\
&\quad= \mathbf{Q}_S\begin{bmatrix}\{\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})-\mathbf{U}_{0\bar{S}}\}\mathbf{M}_0\{\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})-\mathbf{U}_{0\bar{S}}\}^{\mathrm{T}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{Q}_S^{\mathrm{T}},
\end{aligned}$$

$$\begin{aligned}
&\{\mathbf{U}(\boldsymbol{\varphi})-\mathbf{U}_0\}(\mathbf{M}-\mathbf{M}_0)\mathbf{U}_0^{\mathrm{T}}\\
&\quad= \mathbf{Q}_S\begin{bmatrix}\{\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})-\mathbf{U}_{0\bar{S}}\}(\mathbf{M}-\mathbf{M}_0)\mathbf{U}_{0\bar{S}}^{\mathrm{T}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{Q}_S^{\mathrm{T}},
\end{aligned}$$

which implies that

$$\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta},\boldsymbol{\theta}_0) = \mathbf{Q}_S\begin{bmatrix}\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta}_{\bar{S}},\boldsymbol{\theta}_{0\bar{S}}) & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{Q}_S^{\mathrm{T}},$$

where

$$\begin{aligned}
\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta}_{\bar{S}},\boldsymbol{\theta}_{0\bar{S}}) &= \mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})(\mathbf{M}-\mathbf{M}_0)\{\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})-\mathbf{U}_{0\bar{S}}\}^{\mathrm{T}}\\
&\quad+\{\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})-\mathbf{U}_{0\bar{S}}\}\mathbf{M}_0\{\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})-\mathbf{U}_{0\bar{S}}\}^{\mathrm{T}}\\
&\quad+\{\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})-\mathbf{U}_{0\bar{S}}\}(\mathbf{M}-\mathbf{M}_0)\mathbf{U}_{0\bar{S}}^{\mathrm{T}}.
\end{aligned}$$

Clearly, $\operatorname{rank}\{\mathbf{R}_{\mathbf{\Omega}}(\boldsymbol{\theta}_{\bar{S}}, \boldsymbol{\theta}_{0\bar{S}})\} \leq 3r$. Hence, the left-hand side of (25) can be upper bounded similarly using Hölder's inequality and (27) by

$$
\begin{aligned}
&\sup_{\mathcal{B}_n} \left| n\operatorname{vec}(\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0)^{\mathrm{T}}(\mathbf{\Omega}_0^{-1} \otimes \mathbf{\Omega}_0^{-1})\operatorname{vec}\{\mathbf{R}_{\mathbf{\Omega}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\} \right| \\
&\quad = \sup_{\mathcal{B}_n} \left| n\operatorname{tr}\left\{ (\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0)\mathbf{\Omega}_0^{-1}\mathbf{R}_{\mathbf{\Omega}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\mathbf{\Omega}_0^{-1} \right\} \right| \\
&\quad = \sup_{S} \sup_{\mathcal{A}_n(S)} \left| n\operatorname{tr}\left\{ \begin{bmatrix} \widehat{\mathbf{\Omega}}_{\bar{S}} - \mathbf{\Omega}_{0\bar{S}} & \widehat{\mathbf{\Omega}}_{12} \\ \widehat{\mathbf{\Omega}}_{21} & \widehat{\mathbf{\Omega}}_{22} - \mathbf{I}_{p-(\bar{s}+r)} \end{bmatrix} \right.\right. \\
&\qquad\qquad \left.\left. \times \begin{bmatrix} \mathbf{\Omega}_{0\bar{S}}^{-1}\mathbf{R}_{\mathbf{\Omega}}(\boldsymbol{\theta}_{\bar{S}}, \boldsymbol{\theta}_{0\bar{S}})\mathbf{\Omega}_{0\bar{S}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right\} \right| \\
&\quad = \sup_{S} \sup_{\mathcal{A}_n(S)} \left| n\operatorname{tr}\left[\begin{bmatrix} \widehat{\mathbf{\Omega}}_{\bar{S}} - \mathbf{\Omega}_{0\bar{S}} & \widehat{\mathbf{\Omega}}_{12} \\ \widehat{\mathbf{\Omega}}_{21} & \widehat{\mathbf{\Omega}}_{22} - \mathbf{I}_{p-(\bar{s}+r)} \end{bmatrix} \begin{bmatrix} \mathbf{\Omega}_{0\bar{S}}^{-1}\mathbf{R}_{\mathbf{\Omega}}(\boldsymbol{\theta}_{\bar{S}}, \boldsymbol{\theta}_{0\bar{S}})\mathbf{\Omega}_{0\bar{S}}^{-1} \\ \mathbf{0} \end{bmatrix} \right.\right. \\
&\qquad\qquad \left.\left. \times \begin{bmatrix} \mathbf{I}_{\bar{s}+r} & \mathbf{0} \end{bmatrix} \right] \right| \\
&\quad = \sup_{S} \sup_{\mathcal{A}_n(S)} \left| n\operatorname{tr}\left[\begin{bmatrix} \mathbf{I}_{\bar{s}+r} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{\Omega}}_{\bar{S}} - \mathbf{\Omega}_{0\bar{S}} & \widehat{\mathbf{\Omega}}_{12} \\ \widehat{\mathbf{\Omega}}_{21} & \widehat{\mathbf{\Omega}}_{22} - \mathbf{I}_{p-(\bar{s}+r)} \end{bmatrix} \right.\right. \\
&\qquad\qquad \left.\left. \times \begin{bmatrix} \mathbf{\Omega}_{0\bar{S}}^{-1}\mathbf{R}_{\mathbf{\Omega}}(\boldsymbol{\theta}_{\bar{S}}, \boldsymbol{\theta}_{0\bar{S}})\mathbf{\Omega}_{0\bar{S}}^{-1} \\ \mathbf{0} \end{bmatrix} \right] \right| \\
&\quad \leq \sup_{S} \sup_{\mathcal{A}_n(S)} \left| n\operatorname{tr}\left\{ (\widehat{\mathbf{\Omega}}_{\bar{S}} - \mathbf{\Omega}_{0\bar{S}})\mathbf{\Omega}_{0\bar{S}}^{-1}\mathbf{R}_{\mathbf{\Omega}}(\boldsymbol{\theta}_{\bar{S}}, \boldsymbol{\theta}_{0\bar{S}})\mathbf{\Omega}_{0\bar{S}}^{-1} \right\} \right| \\
&\quad = \sup_{S} \sup_{\mathcal{A}_n(S)} \left| n\operatorname{tr}\left\{ \mathbf{\Omega}_{0\bar{S}}^{-1}(\widehat{\mathbf{\Omega}}_{\bar{S}} - \mathbf{\Omega}_{0\bar{S}})\mathbf{\Omega}_{0\bar{S}}^{-1}\mathbf{R}_{\mathbf{\Omega}}(\boldsymbol{\theta}_{\bar{S}}, \boldsymbol{\theta}_{0\bar{S}}) \right\} \right| \\
&\quad \leq n \sup_{S} \|\mathbf{\Omega}_{0\bar{S}}^{-1}\|_2^2 \|\widehat{\mathbf{\Omega}}_{\bar{S}} - \mathbf{\Omega}_{0\bar{S}}\|_2 \sup_{S} \sup_{\boldsymbol{\theta} \in \mathcal{A}_n(S)} \|\mathbf{R}_{\mathbf{\Omega}}(\boldsymbol{\theta}_{\bar{S}}, \boldsymbol{\theta}_{0\bar{S}})\|_* \\
&\quad \leq n \sup_{S} \|\widehat{\mathbf{\Omega}}_{\bar{S}} - \mathbf{\Omega}_{0\bar{S}}\|_2 \sup_{S} \sup_{\boldsymbol{\theta} \in \mathcal{A}_n(S)} \sqrt{3r}\|\mathbf{R}_{\mathbf{\Omega}}(\boldsymbol{\theta}_{\bar{S}}, \boldsymbol{\theta}_{0\bar{S}})\|_{\mathrm{F}} \\
&\quad \lesssim n\sqrt{\frac{r(rs_0 \log n + s_0 \log p)}{n}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1^2 \lesssim \sqrt{\frac{(r^2 s^2 \log n + rs^2 \log p)^3}{n}}
\end{aligned}
$$

with probability greater than $1 - 2\exp(-\kappa_0 n\epsilon_n^2)$. Hence the left-hand side of (25) is also bounded by $C\delta_n$ with probability at least $1 - 2\exp(-\kappa_0 n\epsilon_n^2)$.

■ For $\mathbf{R}_1$, we follow the same spirit and let $\mathbf{\Omega}_{\bar{S}} = \mathbf{U}(\boldsymbol{\varphi}_{\bar{S}})\mathbf{M}\mathbf{U}(\boldsymbol{\varphi}_{\bar{S}}) + \mathbf{I}_{\bar{s}+r}$ and $\mathbf{\Omega}_{0\bar{S}} := \mathbf{U}_{0\bar{S}}\mathbf{M}_0\mathbf{U}_{0\bar{S}}^{\mathrm{T}} + \mathbf{I}_{\bar{s}+r}$. Denote by $\mathbf{\Omega} := \mathbf{\Omega}(\boldsymbol{\theta})$. It follows that

$$
\begin{aligned}
(\mathbf{\Omega}_0 - \mathbf{\Omega})\mathbf{\Omega}_0^{-1} &= \mathbf{Q}_S \begin{bmatrix} \mathbf{\Omega}_{0\bar{S}} - \mathbf{\Omega}_{\bar{S}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}} \mathbf{Q}_S \begin{bmatrix} \mathbf{\Omega}_{0\bar{S}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-(\bar{s}+r)} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}} \\
&= \mathbf{Q}_s \begin{bmatrix} (\mathbf{\Omega}_{0\bar{S}} - \mathbf{\Omega}_{\bar{S}})\mathbf{\Omega}_{0\bar{S}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbf{R}_1(\boldsymbol{\Omega}, \boldsymbol{\Omega}_0) &= \mathbf{Q}_S \begin{bmatrix} \boldsymbol{\Omega}_{0\bar{S}}^{-1}(\boldsymbol{\Omega}_{0\bar{S}} - \boldsymbol{\Omega}_{\bar{S}})\boldsymbol{\Omega}_{0\bar{S}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}} \sum_{m=1}^{\infty} \mathbf{Q}_s \\
&\quad \times \begin{bmatrix} \{(\boldsymbol{\Omega}_{0\bar{S}} - \boldsymbol{\Omega}_{\bar{S}})\boldsymbol{\Omega}_{0\bar{S}}^{-1}\}^m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}} \\
&= \mathbf{Q}_S \begin{bmatrix} \boldsymbol{\Omega}_{0\bar{S}}^{-1}(\boldsymbol{\Omega}_{0\bar{S}} - \boldsymbol{\Omega}_{\bar{S}})\boldsymbol{\Omega}_{0\bar{S}}^{-1} \sum_{m=1}^{\infty}\{(\boldsymbol{\Omega}_{0\bar{S}} - \boldsymbol{\Omega}_{\bar{S}})\boldsymbol{\Omega}_{0\bar{S}}^{-1}\}^m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}} \\
&:= \mathbf{Q}_S \begin{bmatrix} \mathbf{R}_1(\boldsymbol{\Omega}_{\bar{S}}, \boldsymbol{\Omega}_{0\bar{S}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}}.
\end{aligned}
$$

Let $\mathbf{R}_1(\boldsymbol{\Omega}_{\bar{S}}, \boldsymbol{\Omega}_{0\bar{S}})$ yield singular value decomposition $\mathbf{W}_1\mathbf{S}\mathbf{W}_2^{\mathrm{T}}$. Following the same reasoning, we have,

$$
\begin{aligned}
&\sup_{\mathcal{B}_n} \left| n\mathrm{vec}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\mathrm{vec}\{\mathbf{R}_1(\boldsymbol{\Omega}(\boldsymbol{\theta}), \boldsymbol{\Omega}_0)\} \right| \\
&\quad = \sup_S \sup_{\mathcal{A}_n(S)} n \left| \mathrm{tr}\left\{ \mathbf{Q}_S \begin{bmatrix} \widehat{\boldsymbol{\Omega}}_{\bar{S}} - \boldsymbol{\Omega}_{0\bar{S}} & \widehat{\boldsymbol{\Omega}}_{12} \\ \widehat{\boldsymbol{\Omega}}_{21} & \widehat{\boldsymbol{\Omega}}_{22} - \mathbf{I}_{p-(\bar{s}+r)} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S \right.\right. \\
&\qquad\qquad \left.\left. \times \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{0} \end{bmatrix} \mathbf{S} \begin{bmatrix} \mathbf{W}_2^{\mathrm{T}} & \mathbf{0} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}} \right\} \right| \\
&\quad = \sup_S \sup_{\mathcal{A}_n(S)} n \left| \mathrm{tr}\left\{ \begin{bmatrix} \mathbf{W}_2^{\mathrm{T}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\Omega}}_{\bar{S}} - \boldsymbol{\Omega}_{0\bar{S}} & \widehat{\boldsymbol{\Omega}}_{12} \\ \widehat{\boldsymbol{\Omega}}_{21} & \widehat{\boldsymbol{\Omega}}_{22} - \mathbf{I}_{p-(\bar{s}+r)} \end{bmatrix} \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{0} \end{bmatrix} \mathbf{S} \right\} \right| \\
&\quad = \sup_S \sup_{\mathcal{A}_n(S)} n \left| \mathrm{tr}\left\{ \mathbf{W}_2^{\mathrm{T}}(\widehat{\boldsymbol{\Omega}}_{\bar{S}} - \boldsymbol{\Omega}_{0\bar{S}})\mathbf{W}_1\mathbf{S} \right\} \right| \\
&\quad \lesssim n\sqrt{s_0} \sup_{S\in\mathcal{S}(\kappa_0 s_0)} \|\widehat{\boldsymbol{\Omega}}_{\bar{S}} - \boldsymbol{\Omega}_{0\bar{S}}\|_2 \sup_S \sup_{\boldsymbol{\theta}\in\mathcal{A}_n(S)} \|\mathbf{R}_1(\boldsymbol{\Omega}_{\bar{S}}, \boldsymbol{\Omega}_{0\bar{S}})\|_{\mathrm{F}} \\
&\quad \lesssim \sqrt{ns_0(rs_0\log n + s_0\log p)}\|\boldsymbol{\Omega}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_0\|_{\mathrm{F}}^2 \\
&\quad \lesssim \sqrt{\frac{(r^2s^2\log n + rs^2\log p)^3}{n}}
\end{aligned}
$$

with probability at least $1 - 2\exp(-\kappa_0 n\epsilon_n^2)$. The proof is thus completed. □

We are now in a position to complete the proof of local asymptotic normality (Theorem 5.8).

Proof of Theorem 5.8. We first consider the Taylor expansion of ℓ as a function of $\boldsymbol{\Omega}$ when $\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_{\mathrm{F}}$ is sufficiently small. By definition,

$$
\begin{aligned}
\ell(\boldsymbol{\Omega}) - \ell(\boldsymbol{\Omega}_0) &= \frac{n}{2}\mathrm{tr}\left\{\widehat{\boldsymbol{\Omega}}(\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Omega}^{-1})\right\} + \frac{n}{2}\log\det(\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}_0) \\
&= \frac{n}{2}\mathrm{tr}\left\{(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)(\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Omega}^{-1})\right\} \\
&\quad + \frac{n}{2}\mathrm{tr}(\mathbf{I}_p - \boldsymbol{\Omega}_0\boldsymbol{\Omega}^{-1}) + \frac{n}{2}\log\det\{(\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}_0^{-1})\boldsymbol{\Omega}_0 + \mathbf{I}_p\}
\end{aligned}
$$

$$= \frac{n}{2}\mathrm{tr}\left\{(\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0)(\mathbf{\Omega}_0^{-1} - \mathbf{\Omega}^{-1})\right\} + \frac{n}{2}\mathrm{tr}\{\mathbf{\Omega}_0^{1/2}(\mathbf{\Omega}_0^{-1} - \mathbf{\Omega}^{-1})\mathbf{\Omega}_0^{1/2}\}$$
$$+ \frac{n}{2}\log\det\{\mathbf{I}_p - \mathbf{\Omega}_0^{1/2}(\mathbf{\Omega}_0^{-1} - \mathbf{\Omega}^{-1})\mathbf{\Omega}_0^{1/2}\}.$$

Let $h_k := \lambda_k\{\mathbf{\Omega}_0^{1/2}(\mathbf{\Omega}_0^{-1} - \mathbf{\Omega}^{-1})\mathbf{\Omega}_0^{1/2}\}$, $k = 1, \dots, 2r$. Clearly,

$$\begin{aligned}\max_{1\le k\le 2r}|h_k| &\le \|\mathbf{\Omega}_0^{1/2}\|_2\|\mathbf{\Omega}_0^{-1} - \mathbf{\Omega}^{-1}\|_2\|\mathbf{\Omega}_0^{1/2}\|_2 \\ &\le \|\mathbf{\Omega}_0\|_2\|\mathbf{\Omega}^{-1}(\mathbf{\Omega} - \mathbf{\Omega}_0)\mathbf{\Omega}_0^{-1}\|_2 \\ &\le \|\mathbf{\Omega}_0\|_2\|\mathbf{\Omega} - \mathbf{\Omega}_0\|_2.\end{aligned}$$

Furthermore, using the Taylor expansion technique with the integral remainder (see, for example, Lemma 6.2 in [22]),

$$\begin{aligned}&\frac{n}{2}\mathrm{tr}\{\mathbf{\Omega}_0^{1/2}(\mathbf{\Omega}_0^{-1} - \mathbf{\Omega}^{-1})\mathbf{\Omega}_0^{1/2}\} + \frac{n}{2}\log\det\{\mathbf{I}_p - \mathbf{\Omega}_0^{1/2}(\mathbf{\Omega}_0^{-1} - \mathbf{\Omega}^{-1})\mathbf{\Omega}_0^{1/2}\} \\ &\quad= \frac{n}{2}\sum_{k=1}^{2r}\{h_k + \log(1-h_k)\} = -\frac{n}{4}\sum_{k=1}^{2r}h_k^2 - \frac{n}{2}\sum_{k=1}^{2r}\int_0^{h_k}\frac{(h_k - s)^2}{(1-s)^3}\mathrm{d}s \\ &\quad= -\frac{n}{4}\|\mathbf{\Omega}_0^{1/2}(\mathbf{\Omega}^{-1} - \mathbf{\Omega}_0^{-1})\mathbf{\Omega}_0^{1/2}\|_{\mathrm{F}}^2 - \frac{n}{2}\sum_{k=1}^{2r}\int_0^{h_k}\frac{(h_k - s)^2}{(1-s)^3}\mathrm{d}s.\end{aligned}$$

We now analyze the linear term $\mathrm{tr}\{(\widehat{\mathbf{\Omega}} - \mathbf{\Omega}_0)(\mathbf{\Omega}_0^{-1} - \mathbf{\Omega}^{-1})\}$ and the quadratic term $\|\mathbf{\Omega}_0^{1/2}(\mathbf{\Omega}_0^{-1} - \mathbf{\Omega}^{-1})\mathbf{\Omega}_0^{1/2}\|_{\mathrm{F}}^2$ separately. By the matrix series expansion,

$$\begin{aligned}\mathbf{\Omega}_0^{-1} - \mathbf{\Omega}^{-1} &= \mathbf{\Omega}_0^{-1}(\mathbf{\Omega} - \mathbf{\Omega}_0)(\mathbf{\Omega} - \mathbf{\Omega}_0 + \mathbf{\Omega}_0)^{-1} \\ &= \mathbf{\Omega}_0^{-1}(\mathbf{\Omega} - \mathbf{\Omega}_0)\mathbf{\Omega}_0^{-1}\{(\mathbf{\Omega} - \mathbf{\Omega}_0)\mathbf{\Omega}_0^{-1} + \mathbf{I}_p\}^{-1} \\ &= \mathbf{\Omega}_0^{-1}(\mathbf{\Omega} - \mathbf{\Omega}_0)\mathbf{\Omega}_0^{-1} + \mathbf{R}_1(\mathbf{\Omega}, \mathbf{\Omega}_0),\end{aligned}$$

where the remainder

$$\mathbf{R}_1(\mathbf{\Omega}, \mathbf{\Omega}_0) := \mathbf{\Omega}_0^{-1}(\mathbf{\Omega}_0 - \mathbf{\Omega})\mathbf{\Omega}_0^{-1}\sum_{m=1}^{\infty}\{(\mathbf{\Omega}_0 - \mathbf{\Omega})\mathbf{\Omega}_0^{-1}\}^m$$

satisfies $\|\mathbf{R}_1(\mathbf{\Omega}, \mathbf{\Omega}_0)\|_{\mathrm{F}} \lesssim \|\mathbf{\Omega} - \mathbf{\Omega}_0\|_{\mathrm{F}}^2$. The vectorization form of the previous equation can be written as

$$\mathrm{vec}(\mathbf{\Omega}_0^{-1} - \mathbf{\Omega}^{-1}) = (\mathbf{\Omega}_0^{-1} \otimes \mathbf{\Omega}_0^{-1})\mathrm{vec}(\mathbf{\Omega} - \mathbf{\Omega}_0) + \mathrm{vec}\{\mathbf{R}_1(\mathbf{\Omega}, \mathbf{\Omega}_0)\}.$$

Now we consider parameterizing $\mathbf{\Omega}$ by $\mathbf{\Omega} = \mathbf{\Omega}(\boldsymbol{\theta})$. It follows from Theorem 3.1 in [55] that

$$\|\mathbf{R}_1(\mathbf{\Omega}(\boldsymbol{\theta}), \mathbf{\Omega}_0)\|_{\mathrm{F}} \lesssim \|\mathbf{\Omega}(\boldsymbol{\theta}) - \mathbf{\Omega}_0\|_{\mathrm{F}}^2 \lesssim \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1^2.$$

Following the proof of Theorem 3.1 in [55], we obtain the following matrix decomposition

$$\mathbf{\Omega}(\boldsymbol{\theta}) - \mathbf{\Omega}_0 = \mathbf{\Sigma}(\boldsymbol{\theta}) - \mathbf{\Sigma}(\boldsymbol{\theta}_0)$$

$$= \mathbf{U}_0(\mathbf{M} - \mathbf{M}_0)\mathbf{U}_0^{\mathrm{T}} + \mathbf{U}_0\mathbf{M}_0\{\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\}^{\mathrm{T}} + \{\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\}\mathbf{M}_0\mathbf{U}_0^{\mathrm{T}} + \mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0),$$

where the remainder

$$\begin{aligned}\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) := \mathbf{U}(\boldsymbol{\varphi})(\mathbf{M} - \mathbf{M}_0)\{\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\}^{\mathrm{T}} &\\ + \{\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\}\mathbf{M}_0\{\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\}^{\mathrm{T}} &\\ + \{\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\}(\mathbf{M} - \mathbf{M}_0)\mathbf{U}_0^{\mathrm{T}} &\end{aligned}$$

satisfies $\|\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\|_{\mathrm{F}} \lesssim \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2 \lesssim \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1^2$. In the vectorization form, we can write

$$\operatorname{vec}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0) = D_{\boldsymbol{\mu}}\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0) + (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\mathbf{U}_0\mathbf{M}_0 \otimes \mathbf{I}_p)\operatorname{vec}\{\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\} + \operatorname{vec}\{\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\}.$$

In addition, by Theorem 2.1 in [55], we have,

$$\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0 = 2(\mathbf{I}_p - \mathbf{X}_0)^{-1}(\mathbf{X}_{\boldsymbol{\varphi}} - \mathbf{X}_0)(\mathbf{I}_p - \mathbf{X}_0)^{-1}\mathbf{I}_{p\times r} + \mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0),$$

where $\mathbf{U}_0 = \mathbf{U}(\boldsymbol{\varphi}_0)$, and $\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0)$ satisfies $\|\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0)\|_{\mathrm{F}} \lesssim \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1^2$. The vectorization version of the previous display can be written as

$$\operatorname{vec}\{\mathbf{U}(\boldsymbol{\varphi}) - \mathbf{U}_0\} = D\mathbf{U}(\boldsymbol{\varphi}_0)(\boldsymbol{\varphi} - \boldsymbol{\varphi}_0) + \operatorname{vec}\{\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0)\},$$

where the Fréchet derivative $D\mathbf{U}$ is defined by (4). Recall that $D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ is defined by (8). It follows from the above derivations that

$$\begin{aligned}\operatorname{vec}\{\boldsymbol{\Omega}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_0\} &= D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + (\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\mathbf{U}_0\mathbf{M}_0 \otimes \mathbf{I}_p)\operatorname{vec}\{\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0)\} \\ &\quad + \operatorname{vec}\{\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\}.\end{aligned}$$

This means that over $\mathcal{B}_n$, we can have well control of the Frobenius norm deviation $\|\boldsymbol{\Omega}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_0\|_{\mathrm{F}}$: For any $\boldsymbol{\theta} \in \mathcal{B}_n$,

$$\begin{aligned}\|\boldsymbol{\Omega}(\boldsymbol{\theta}) - \boldsymbol{\Omega}_0\|_{\mathrm{F}} &= \|\operatorname{vec}\{\boldsymbol{\Omega}(\boldsymbol{\theta})\} - \operatorname{vec}(\boldsymbol{\Omega}_0)\|_2 \\ &\leq \|D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\|_2\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 + 2\|\mathbf{M}_0\|_2\|\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0)\|_{\mathrm{F}} \\ &\quad + \|\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\|_{\mathrm{F}} \\ &\lesssim \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1 \lesssim \sqrt{\frac{r^2 s^2 \log n + rs^2 \log p}{n}}.\end{aligned}$$

Hence for the precision matrix $\boldsymbol{\Omega}^{-1}$, we have

$$\begin{aligned}\operatorname{vec}(\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Omega}^{-1}) &= (\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + (\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1})(\mathbf{I}_{p^2} + \mathbf{K}_{pp})(\mathbf{U}_0\mathbf{M}_0 \otimes \mathbf{I}_p)\operatorname{vec}\{\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi}, \boldsymbol{\varphi}_0)\} \\ &\quad + (\boldsymbol{\Omega}_0^{-1} \otimes \boldsymbol{\Omega}_0^{-1})\operatorname{vec}\{\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)\} + \operatorname{vec}\{\mathbf{R}_1(\boldsymbol{\Omega}(\boldsymbol{\theta}), \boldsymbol{\Omega}_0)\},\end{aligned}$$

with the remainder $\mathbf{R}_1$ satisfying $\|\mathbf{R}_1(\boldsymbol{\Omega}(\boldsymbol{\theta}),\boldsymbol{\Omega}_0)\|_{\mathrm{F}} \lesssim \|\boldsymbol{\Omega}(\boldsymbol{\theta})-\boldsymbol{\Omega}_0\|_{\mathrm{F}}^2 \lesssim \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1^2$. Denote by

$$\begin{aligned}\mathbf{r}_{\boldsymbol{\Omega}^{-1}}(\boldsymbol{\theta},\boldsymbol{\theta}_0) &:= (\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})(\mathbf{I}_{p^2}+\mathbf{K}_{pp})(\mathbf{U}_0\mathbf{M}_0\otimes\mathbf{I}_p)\mathrm{vec}\{\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi},\boldsymbol{\varphi}_0)\}\\
&\quad+(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})\mathrm{vec}\{\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\}+\mathrm{vec}\{\mathbf{R}_1(\boldsymbol{\Omega}(\boldsymbol{\theta}),\boldsymbol{\Omega}_0)\}\\
&=(\mathbf{I}_{p^2}+\mathbf{K}_{pp})(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})\mathrm{vec}\{\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi},\boldsymbol{\varphi}_0)\mathbf{M}_0\mathbf{U}_0^{\mathrm{T}}\}\\
&\quad+(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})\mathrm{vec}\{\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\}+\mathrm{vec}\{\mathbf{R}_1(\boldsymbol{\Omega}(\boldsymbol{\theta}),\boldsymbol{\Omega}_0)\}.\end{aligned}$$

Clearly, $\|\mathbf{r}_{\boldsymbol{\Omega}^{-1}}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\|_2 \lesssim \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1^2$ by the properties of the remainders $\mathbf{R}_1$, $\mathbf{R}_{\boldsymbol{\Omega}}$, and $\mathbf{R}_{\mathbf{U}}$. It follows that

$$\begin{aligned}\mathrm{tr}\{(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)(\boldsymbol{\Omega}_0^{-1}-\boldsymbol{\Omega}^{-1})\} &= \mathrm{vec}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\\
&\quad+\mathrm{vec}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)^{\mathrm{T}}\mathbf{r}_{\boldsymbol{\Omega}^{-1}}(\boldsymbol{\theta},\boldsymbol{\theta}_0),\end{aligned}$$

and

$$\begin{aligned}\|\boldsymbol{\Omega}_0^{1/2}(\boldsymbol{\Omega}^{-1}-\boldsymbol{\Omega}_0^{-1})\boldsymbol{\Omega}_0^{1/2}\|_{\mathrm{F}}^2 &= \mathrm{tr}\left\{(\boldsymbol{\Omega}^{-1}-\boldsymbol{\Omega}_0^{-1})\boldsymbol{\Omega}_0(\boldsymbol{\Omega}^{-1}-\boldsymbol{\Omega}_0^{-1})\boldsymbol{\Omega}_0\right\}\\
&=\mathrm{vec}(\boldsymbol{\Omega}_0^{-1}-\boldsymbol{\Omega}^{-1})(\boldsymbol{\Omega}_0\otimes\boldsymbol{\Omega}_0)\mathrm{vec}(\boldsymbol{\Omega}_0^{-1}-\boldsymbol{\Omega}^{-1})\\
&=(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\\
&\quad+2(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{r}_{\boldsymbol{\Omega}^{-1}}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\\
&\quad+\mathbf{r}_{\boldsymbol{\Omega}^{-1}}(\boldsymbol{\theta},\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0\otimes\boldsymbol{\Omega}_0)\mathbf{r}_{\boldsymbol{\Omega}^{-1}}(\boldsymbol{\theta},\boldsymbol{\theta}_0).\end{aligned}$$

Denote by

$$\begin{aligned}r_q(\boldsymbol{\theta},\boldsymbol{\theta}_0) &:= 2(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{r}_{\boldsymbol{\Omega}^{-1}}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\\
&\quad+\mathbf{r}_{\boldsymbol{\Omega}^{-1}}(\boldsymbol{\theta},\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0\otimes\boldsymbol{\Omega}_0)\mathbf{r}_{\boldsymbol{\Omega}^{-1}}(\boldsymbol{\theta},\boldsymbol{\theta}_0).\end{aligned}$$

By the property of the remainder $\mathbf{r}_{\boldsymbol{\Omega}^{-1}}$, we see that $|r_q(\boldsymbol{\theta},\boldsymbol{\theta}_0)| \lesssim \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1^3$.

Now putting all the above derivations together, we obtain the following expansion of the log-likelihood function:

$$\begin{aligned}\ell(\boldsymbol{\Omega}(\boldsymbol{\theta}))-\ell(\boldsymbol{\Omega}(\boldsymbol{\theta}_0)) &= \frac{n}{2}\mathrm{vec}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\\
&\quad-\frac{n}{4}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\\
&\quad-\frac{n}{2}\sum_{k=1}^{2r}\int_0^{h_k}\frac{(h_k-s)^2}{(1-s)^3}\mathrm{d}s-\frac{n}{4}r_q(\boldsymbol{\theta},\boldsymbol{\theta}_0)\\
&\quad+\frac{n}{2}\mathrm{vec}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)^{\mathrm{T}}\mathbf{r}_{\boldsymbol{\Omega}^{-1}}(\boldsymbol{\theta},\boldsymbol{\theta}_0).\end{aligned}$$

The third line of the previous equation is the deterministic remainder and the fourth line is the stochastic remainder. For the sum of the integrals in the third line of the above display, since $\|\boldsymbol{\Omega}(\boldsymbol{\theta})-\boldsymbol{\Omega}_0\|_{\mathrm{F}}\to 0$, we may assume that

$\max_k |h_k| \leq 1/2$, and hence,

$$\sup_{\boldsymbol{\theta}\in\mathcal{A}_n(S)}\left|\frac{n}{2}\sum_{k=1}^{2r}\int_0^{h_k}\frac{(h_k-s)^2}{(1-s)^3}\mathrm{d}s\right| \lesssim n\sup_{\|\boldsymbol{\Omega}(\boldsymbol{\theta})-\boldsymbol{\Omega}_0\|_{\mathrm{F}}<M\epsilon_n}\sum_{k=1}^{2r}\int_0^{h_k}(h_k-s)^2\mathrm{d}s$$
$$\leq n\max_{1\leq k\leq 2r}|h_k|\frac{1}{3}\sum_{k=1}^{2r}h_k^2 \lesssim n\|\boldsymbol{\Omega}(\boldsymbol{\theta})-\boldsymbol{\Omega}_0\|_{\mathrm{F}}^3$$
$$\lesssim n\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1^3 \lesssim \delta_n,$$

where $\delta_n = \sqrt{(r^2s^2\log n + rs^2\log p)^3/n}$. The stochastic remainder is given by

$$\begin{aligned}
&\frac{n}{2}\mathrm{vec}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)^{\mathrm{T}}\mathbf{r}_{\boldsymbol{\Omega}^{-1}}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\\
&\quad= n\mathrm{vec}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})\mathrm{vec}\{\mathbf{R}_{\mathbf{U}}(\boldsymbol{\varphi},\boldsymbol{\varphi}_0)\mathbf{M}_0\mathbf{U}_0^{\mathrm{T}}\}\\
&\qquad+\frac{n}{2}\mathrm{vec}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})\mathrm{vec}\{\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\}\\
&\qquad+\frac{n}{2}\mathrm{vec}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)^{\mathrm{T}}\mathrm{vec}\{\mathbf{R}_1(\boldsymbol{\Omega}(\boldsymbol{\theta}),\boldsymbol{\Omega}_0)\}.
\end{aligned}$$

By Lemma 5.9, the supremum of the stochastic remainder over $\boldsymbol{\theta}\in\mathcal{B}_n$ is also $O(\delta_n)$ with probability at least $1-2e^{-\kappa_0 n\epsilon_n^2}$, and hence completing the proof. □

5.6. Distributional approximation: Proof of Theorem 3.2

This subsection elaborates on the proof of Theorem 3.2. We remark that the proof is a generalization of the proof of Theorem 6 in [11], but it also requires a local asymptotic normality argument developed in Section 5.5. For convenience, we introduce additional notations that will be used to characterize the limit shape of the posterior distribution. Denote by

$$\mathbf{Z}_0 := \sqrt{\frac{n}{2}}(\boldsymbol{\Omega}_0^{-1/2}\otimes\boldsymbol{\Omega}_0^{-1/2})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0),\ \boldsymbol{\varepsilon}_n := \sqrt{\frac{n}{2}}(\boldsymbol{\Omega}_0^{-1/2}\otimes\boldsymbol{\Omega}_0^{-1/2})\mathrm{vec}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0).$$

Let $\mathbf{Z}_{0S} := \mathbf{Z}_0\mathbf{F}_S$. Then $\widehat{\boldsymbol{\theta}}_S$, $\mathbf{I}_S(\boldsymbol{\theta}_0)$, and $\widehat{w}_S$ can be equivalently written as

$$\widehat{\boldsymbol{\theta}}_S = (\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\mathbf{Z}_{0S}^{\mathrm{T}}(\mathbf{Z}_0\boldsymbol{\theta}_0+\boldsymbol{\varepsilon}_n),\quad n\mathbf{I}_S(\boldsymbol{\theta}_0) = \mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S},$$
$$\widehat{w}_S \propto \frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}\det\{2\pi(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\}^{1/2}\exp\left(\frac{1}{2}\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2\right),$$

where $\gamma(|S|)$ is defined by

$$\gamma(|S|) =: \int_{\|\mathbf{A}_S\|_2<1}\exp(-2\|\mathrm{vec}(\mathbf{A}_S)\|_1)\mathrm{d}\mathbf{A}_S.$$

The proof is based on the following collection of technical lemmas. Recall that the sub-Gaussian norm and the sub-exponential norm of a random variable X are defined by

$$\|X\|_{\psi_2} = \sup_{p\geq 1}p^{-1/2}(\mathbb{E}_0|X|^p)^{1/p},\quad \|X\|_{\psi_1} = \sup_{p\geq 1}p^{-1}(\mathbb{E}_0|X|^p)^{1/p}.$$

We refer to [51] for a detailed review on the concept of these (Orlicz) norms.

In order to establish Theorem 3.2 via local asymptotic normality, it is necessary to understand the behavior of $(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n$, which we discuss in Lemma 5.10 below.

Lemma 5.10. *Let* $\mathbf{y}_1,\ldots,\mathbf{y}_n\overset{\text{i.i.d.}}{\sim}\mathrm{N}(\mathbf{0},\boldsymbol{\Omega}_0)$ *with* $\boldsymbol{\Omega}_0:=\boldsymbol{\Omega}(\theta_0)$. *Suppose conditions A1–A5 hold. Then there exists some constant* $C_0>0$ *only depending on the spectra of* $\boldsymbol{\Omega}_0$, *such that*

$$\mathbb{P}_0\left\{|(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n|>C_0\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1\sqrt{n\log p}\right\}\leq\frac{2}{p}.$$

Proof of Lemma 5.10. Denote by

$$\mathbf{X}_0:=\mathbf{X}_{\boldsymbol{\varphi}_0}=\begin{bmatrix}\mathbf{0}_{r\times r} & -\mathbf{A}_0^{\mathrm{T}}\\ \mathbf{A}_0 & \mathbf{0}_{(p-r)\times(p-r)}\end{bmatrix}\quad\text{and}\quad\mathbf{C}_0:=(\mathbf{I}_p-\mathbf{X}_0)^{-1}.$$

By definition, we have

$$\begin{aligned}
&\frac{2}{n}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n\\
&\quad=(\boldsymbol{\varphi}-\boldsymbol{\varphi}_0)^{\mathrm{T}}D\mathbf{U}(\boldsymbol{\varphi}_0)^{\mathrm{T}}(\mathbf{M}_0\mathbf{U}_0^{\mathrm{T}}\otimes\mathbf{I}_p)(\mathbf{I}_{p^2}+\mathbf{K}_{pp})(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})\mathrm{vec}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)\\
&\qquad+(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^{\mathrm{T}}D_{\boldsymbol{\mu}}\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})\mathrm{vec}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)\\
&\quad=4\mathrm{vec}(\mathbf{X}_{\boldsymbol{\varphi}}-\mathbf{X}_0)^{\mathrm{T}}\{(\mathbf{I}_p-\mathbf{X}_0)^{-1}\mathbf{I}_{p\times r}\mathbf{M}_0\mathbf{U}_0^{\mathrm{T}}\otimes(\mathbf{I}_p-\mathbf{X}_0)^{-\mathrm{T}}\}\\
&\qquad\times\mathrm{vec}\{\boldsymbol{\Omega}_0^{-1}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\}\\
&\qquad+(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^{\mathrm{T}}\mathbb{D}_r^{\mathrm{T}}\mathrm{vec}\{\boldsymbol{\Omega}_0^{-1}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\}\\
&\quad=4\mathrm{vec}(\mathbf{X}_{\boldsymbol{\varphi}}-\mathbf{X}_0)^{\mathrm{T}}\mathrm{vec}\{\mathbf{C}_0^{\mathrm{T}}\boldsymbol{\Omega}_0^{-1}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\mathbf{U}_0\mathbf{M}_0\mathbf{I}_{p\times r}^{\mathrm{T}}\mathbf{C}_0^{\mathrm{T}}\}\\
&\qquad+\mathrm{vec}(\mathbf{M}-\mathbf{M}_0)^{\mathrm{T}}\mathrm{vec}\{\mathbf{U}_0^{\mathrm{T}}\boldsymbol{\Omega}_0^{-1}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)\widehat{\boldsymbol{\Omega}}_0^{-1}\mathbf{U}_0\}
\end{aligned}$$

Let $\mathbf{e}_j$ be the standard basis vector along the jth coordinate in $\mathbb{R}^p$, i.e., the jth coordinate being 1 and the rest of the coordinates being zeros, $\boldsymbol{\alpha}_i:=\boldsymbol{\Omega}_0^{-1}\mathbf{C}_0\mathbf{e}_i$, $\boldsymbol{\beta}_j:=\boldsymbol{\Omega}_0^{-1}\mathbf{U}_0\mathbf{M}_0\mathbf{I}_{p\times r}^{\mathrm{T}}\mathbf{C}_0^{\mathrm{T}}\mathbf{e}_j$, and $\boldsymbol{\gamma}_k:=\boldsymbol{\Omega}_0^{-1}\mathbf{U}_0\mathbf{e}_k$. Then by the Hölder's inequality,

$$\begin{aligned}
&|(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n|\\
&\quad\leq 2n\|\mathrm{vec}(\mathbf{X}_{\boldsymbol{\varphi}}-\mathbf{X}_0)\|_1\max_{j,h\in[p]}\left|\boldsymbol{\alpha}_j^{\mathrm{T}}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)\boldsymbol{\beta}_h\right|\\
&\qquad+n\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|_1\max_{k,l\in[r]}\left|\boldsymbol{\gamma}_k^{\mathrm{T}}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)\boldsymbol{\gamma}_l\right|\\
&\quad\leq n\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1\max_{j,h\in[p],k,l\in[r]}\left\{2\left|\boldsymbol{\alpha}_j^{\mathrm{T}}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)\boldsymbol{\beta}_h\right|,\left|\boldsymbol{\gamma}_k^{\mathrm{T}}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)\boldsymbol{\gamma}_l\right|\right\}.
\end{aligned}$$

Observe that

$$\|\mathbf{C}_0\|_2^2=\lambda_{\max}(\mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0)=\lambda_{\min}^{-1}\{(\mathbf{I}+\mathbf{X}_0)(\mathbf{I}-\mathbf{X}_0)\}=\lambda_{\min}^{-1}(\mathbf{I}+\mathbf{X}_0\mathbf{X}_0^{\mathrm{T}})\leq 1,$$

and that

$$\|\boldsymbol{\alpha}_j\|_2 \le \|\boldsymbol{\Omega}_0^{-1}\|_2\|\mathbf{C}_0\|_2 \le 1,\ \|\boldsymbol{\beta}_j\|_2 \le \|\boldsymbol{\Omega}_0^{-1}\|_2\|\boldsymbol{\Lambda}_0\|_2\|\mathbf{C}_0\|_2 \le \|\boldsymbol{\Omega}_0\|_2,$$
$$\|\boldsymbol{\gamma}_k\|_2 \le \|\boldsymbol{\Omega}_0^{-1}\|_2 \le 1,\ \|\mathbf{u}^{\mathrm{T}}\mathbf{y}_i\|_{\psi_2} \lesssim (\mathbf{u}^{\mathrm{T}}\boldsymbol{\Omega}_0\mathbf{u})^{1/2} \le \|\boldsymbol{\Omega}_0\|_2^{1/2}\|\mathbf{u}\|_2 \text{ for all } \mathbf{u}\in\mathbb{R}^p.$$

It follows from the properties of Orlicz norms that

$$\max_{j,h\in[p]}\|\boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\boldsymbol{\beta}_h\|_{\psi_1} \le \max_{j\in[p]}\|\boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{y}_i\|_{\psi_2}\max_{j\in[p]}\|\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{y}_i\|_{\psi_2} = O(1),$$
$$\max_{k,l\in[r]}\|\boldsymbol{\gamma}_k^{\mathrm{T}}\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\boldsymbol{\gamma}_l\|_{\psi_1} \le \left(\max_{k\in[r]}\|\boldsymbol{\gamma}_k^{\mathrm{T}}\mathbf{y}_i\|_{\psi_2}\right)^2 = O(1),$$

and hence, by the union bound and the Bernstein-type inequality for sub-exponential random variables (see, for example, Proposition 5.16 in [51]), we have, for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}_0\left(|(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n| > t\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1\right)\\
&\quad\le \sum_{j=1}^p\sum_{h=1}^p\mathbb{P}_0\left(\left|\boldsymbol{\alpha}_j^{\mathrm{T}}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)\boldsymbol{\beta}_h\right| > \frac{t}{2n}\right)\\
&\qquad+\sum_{k=1}^r\sum_{l=1}^r\mathbb{P}_0\left(\left|\boldsymbol{\gamma}_k^{\mathrm{T}}(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}_0)\boldsymbol{\gamma}_l\right| > \frac{t}{n}\right)\\
&\quad= \sum_{j=1}^p\sum_{h=1}^p\mathbb{P}_0\left(\left|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\boldsymbol{\beta}_h - \boldsymbol{\alpha}_j^{\mathrm{T}}\boldsymbol{\Omega}_0\boldsymbol{\beta}_h)\right| > \frac{t}{2n}\right)\\
&\qquad+\sum_{k=1}^r\sum_{l=1}^r\mathbb{P}_0\left(\left|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{\gamma}_k^{\mathrm{T}}\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\boldsymbol{\gamma}_l - \boldsymbol{\gamma}_k^{\mathrm{T}}\boldsymbol{\Omega}_0\boldsymbol{\gamma}_l)\right| > \frac{t}{n}\right)\\
&\quad\le 2(p^2+r^2)\exp\left\{-C_0\min\left(\frac{t^2}{n}, t\right)\right\}\\
&\quad\le 4\exp\left\{2\log p - C_0\min\left(\frac{t^2}{n}, t\right)\right\}
\end{aligned}$$

for some constant $C_0 > 0$ (possibly depending on the spectra of $\boldsymbol{\Omega}_0$). The proof is completed by taking $t = (4/C_0)^{1/2}\sqrt{n\log p}$ □

Lemma 5.11 establishes the ℓ_1-norm posterior contraction of $\boldsymbol{\theta}$, which serves as an intermediate step towards proving Theorem 3.2.

Lemma 5.11. *Assume the setup and prior specification in Section 3.1 and suppose conditions A1–A5 hold. Then, there exists some constant $M > 0$ independent of r, s, p, n, such that*

$$\mathbb{E}_0\Pi_{\boldsymbol{\theta}}^{\infty}\left(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1 > M\sqrt{\frac{r^2s^2\log n + rs^2\log p}{n}}\ \middle|\ \mathbf{Y}_n\right) \lesssim \frac{1}{p}.$$

Proof of Lemma 5.11. Denote by

$$\widetilde{\mathcal{A}}_n := \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1 \leq M\sqrt{\frac{r^2 s^2 \log n + r s^2 \log p}{n}}\right\}$$

and

$$Q_S(\mathrm{d}\boldsymbol{\theta}) := \{\phi(\boldsymbol{\theta}_S \mid \widehat{\boldsymbol{\theta}}_S, (\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1})\mathrm{d}\boldsymbol{\theta}_S\}\{\delta_{\mathbf{0}}(\mathrm{d}\boldsymbol{\theta}_{S^c})\}.$$

By definition, we can write

$$\begin{aligned}
&\Pi_{\boldsymbol{\theta}}^{\infty}(\boldsymbol{\theta} \in \widetilde{\mathcal{A}}_n^c \mid \mathbf{Y}_n) \\
&\quad = \left[\sum_{S\in\mathcal{S}_0} \frac{\pi_p(|S|)\exp(\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2/2)}{\binom{p-r}{|S|}\gamma(|S|)} \det\{2\pi(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\}^{1/2} Q_S(\widetilde{\mathcal{A}}_n^c)\right] \\
&\qquad \times \left[\sum_{S\in\mathcal{S}_0} \frac{\pi_p(|S|)\exp(\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2/2)}{\binom{p-r}{|S|}\gamma(|S|)} \det\{2\pi(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\}^{1/2}\right. \\
&\qquad\qquad \left.\times\, Q_S(\mathbb{R}^{(p-r)r+r(r+1)/2})\right]^{-1}.
\end{aligned}$$

Using the fact that for any fixed index set $S \in \mathcal{S}_0$, and any measurable set $\mathcal{B} \subset \mathbb{R}^{(p-r)r+r(r+1)/2}$,

$$\det\{2\pi(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\}^{1/2} Q_S(\mathcal{B}) = \int_{\mathcal{B}_S} \exp\left\{-\frac{1}{2}\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)\|_2^2\right\}\mathrm{d}\boldsymbol{\theta}_S,$$

where $\mathcal{B}_S = \{\boldsymbol{\theta}_S : [\boldsymbol{\theta}_S^{\mathrm{T}}, \mathbf{0}_{S^c}^{\mathrm{T}}]^{\mathrm{T}} \in \mathcal{B}\}$ is the intersection of $\mathcal{B}$ with the subspace $\mathbb{R}^{|S|r+r(r+1)/2}$, we write

$$\begin{aligned}
&\Pi_{\boldsymbol{\theta}}^{\infty}(\boldsymbol{\theta} \in \widetilde{\mathcal{A}}_n^c \mid \mathbf{Y}_n) \\
&\quad = \left[\sum_{S\in\mathcal{S}_0} \frac{\pi_p(|S|)e^{\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2/2}}{\binom{p-r}{|S|}\gamma(|S|)} \int_{(\widetilde{\mathcal{A}}_n^c)_S} e^{-(1/2)\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\widehat{\boldsymbol{\theta}}_S)\|_2^2}\mathrm{d}\boldsymbol{\theta}_S\right] \\
&\qquad \times \left[\sum_{S\in\mathcal{S}_0} \frac{\pi_p(|S|)e^{\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2/2}}{\binom{p-r}{|S|}\gamma(|S|)} \int e^{-(1/2)\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\widehat{\boldsymbol{\theta}}_S)\|_2^2}\mathrm{d}\boldsymbol{\theta}_S\right]^{-1} \\
&\quad = \left[\sum_{S\in\mathcal{S}_0} \frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)} \int_{(\widetilde{\mathcal{A}}_n^c)_S} e^{(1/2)\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2-(1/2)\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\widehat{\boldsymbol{\theta}}_S)\|_2^2}\mathrm{d}\boldsymbol{\theta}_S\right] \\
&\qquad \times \left[\sum_{S\in\mathcal{S}_0} \frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)} \int e^{(1/2)\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2-(1/2)\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\widehat{\boldsymbol{\theta}}_S)\|_2^2}\mathrm{d}\boldsymbol{\theta}_S\right]^{-1}.
\end{aligned}$$

Let $\mathbf{t}_n = \mathbf{Z}_0\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}_n$. Observe that $\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S$ is the projection of $\mathbf{t}_n$ onto the subspace spanned by the columns of $\mathbf{Z}_{0S}$, and by Parseval's identity, we have

$$\frac{1}{2}\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2 - \frac{1}{2}\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)\|_2^2 = \frac{1}{2}\|\mathbf{t}_n\|_2^2 - \frac{1}{2}\|\mathbf{t}_n - \mathbf{Z}_{0S}\boldsymbol{\theta}_S\|_2^2$$

and that

$$
\begin{aligned}
&-\frac{1}{2}\|\mathbf{t}_n-\mathbf{Z}_{0S}\boldsymbol{\theta}_S\|_2^2+\frac{1}{2}\|\mathbf{t}_n-\mathbf{Z}_0\boldsymbol{\theta}_0\|_2^2\\
&\quad=(\mathbf{t}_n-\mathbf{Z}_0\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0(\boldsymbol{\theta}_S-\boldsymbol{\theta}_{0S})-\frac{1}{2}\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\boldsymbol{\theta}_{0S})\|_2^2\\
&\quad=\boldsymbol{\varepsilon}_n^{\mathrm{T}}\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\boldsymbol{\theta}_{0S})-\frac{1}{2}\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\boldsymbol{\theta}_{0S})\|_2^2.
\end{aligned}
$$

Note that $\|\mathbf{t}_n\|_2^2$ and $\|\mathbf{t}_n-\mathbf{Z}_0\boldsymbol{\theta}_0\|_2^2$ do not depend on $\boldsymbol{\theta}$ or the indexing set S. It follows that

$$
\begin{aligned}
&\Pi_{\boldsymbol{\theta}}^{\infty}(\boldsymbol{\theta}\in\widetilde{\mathcal{A}}_n^c\mid\mathbf{Y}_n)\\
&\quad=\left[\sum_{S\in\mathcal{S}_0}\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}\int_{(\widetilde{\mathcal{A}}_n^c)_S}\exp\left\{\frac{1}{2}\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2-\frac{1}{2}\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\widehat{\boldsymbol{\theta}}_S)\|_2^2\right\}\mathrm{d}\boldsymbol{\theta}_S\right]\\
&\qquad\times\left[\sum_{S\in\mathcal{S}_0}\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}\int\exp\left\{\frac{1}{2}\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2-\frac{1}{2}\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\widehat{\boldsymbol{\theta}}_S)\|_2^2\right\}\mathrm{d}\boldsymbol{\theta}_S\right]^{-1}\\
&\quad=\left[\sum_{S\in\mathcal{S}_0}\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}\int_{(\widetilde{\mathcal{A}}_n^c)_S}e^{\boldsymbol{\varepsilon}_n^{\mathrm{T}}\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\boldsymbol{\theta}_{0S})-\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\boldsymbol{\theta}_{0S})\|_2^2/2}\mathrm{d}\boldsymbol{\theta}_S\right]\\
&\qquad\times\left[\sum_{S\in\mathcal{S}_0}\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}\int e^{\boldsymbol{\varepsilon}_n^{\mathrm{T}}\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\boldsymbol{\theta}_{0S})-\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\boldsymbol{\theta}_{0S})\|_2^2/2}\mathrm{d}\boldsymbol{\theta}_S\right]^{-1}\\
&\quad:=\frac{N_n^{\infty}}{D_n^{\infty}}.
\end{aligned}
$$

We now analyze the numerator N_n^{∞} and the denominator D_n^{∞} separately.

■ We first analyze the denominator D_n^{∞}. Denote by $U_S(\mathrm{d}\boldsymbol{\theta}):=(\mathrm{d}\boldsymbol{\theta}_S)\{\delta_{\mathbf{0}}(\mathrm{d}\boldsymbol{\theta}_{S^c})\}$ for any $S\in\mathcal{S}_0$. It follows that

$$
\begin{aligned}
D_n^{\infty}&=\sum_{S\in\mathcal{S}_0}\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}\int\exp\left\{\boldsymbol{\varepsilon}_n^{\mathrm{T}}\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)-\frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2\right\}U_S(\mathrm{d}\boldsymbol{\theta})\\
&\geq\frac{\pi_p(|S_0|)}{\binom{p-r}{|S_0|}\gamma(|S_0|)}\int\exp\left\{\boldsymbol{\varepsilon}_n^{\mathrm{T}}\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)-\frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2\right\}U_{S_0}(\mathrm{d}\boldsymbol{\theta})
\end{aligned}
$$

By definition of the multivariate normal distribution, we have,

$$
\begin{aligned}
&Q_{S_0}(\mathrm{d}\boldsymbol{\theta})\\
&\quad=\frac{\exp\{-(1/2)\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2+(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n\}U_{S_0}(\mathrm{d}\boldsymbol{\theta})}{\int_{\mathbb{R}^{(p-r)r+r(r+1)/2}}\exp\{-(1/2)\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2+(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n\}U_{S_0}(\mathrm{d}\boldsymbol{\theta})}.
\end{aligned}
$$

Define the measures

$$
\eta_{S_0}(\mathrm{d}\boldsymbol{\theta}):=\exp\left\{-\frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2+(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n\right\}U_{S_0}(\mathrm{d}\boldsymbol{\theta}),
$$

$$\zeta_{S_0}(\mathrm{d}\boldsymbol{\beta}_{S_0}) := \exp\left\{-\frac{1}{2}\|\mathbf{Z}_{0S_0}\boldsymbol{\beta}_{S_0}\|_2^2\right\}\mathrm{d}\boldsymbol{\beta}_{S_0}, \quad \boldsymbol{\beta}_{S_0} := \boldsymbol{\theta}_{S_0} - \boldsymbol{\theta}_{0S_0}$$

and the probability distribution $\bar{\sigma}_{S_0}(\mathrm{d}\boldsymbol{\beta}_{S_0}) := \zeta_{S_0}(\mathrm{d}\boldsymbol{\beta}_{S_0})/\zeta_{S_0}(\mathbb{R}^{|S_0|r+r(r+1)/2})$. Then the denominator $\eta(\mathbb{R}^{(p-r)r+r(r+1)/2})$ can be lower bounded as follows:

$$\begin{aligned}
&\eta_{S_0}(\mathbb{R}^{(p-r)r+r(r+1)/2})\\
&\quad= \int_{\mathbb{R}^{(p-r)r+r(r+1)/2}} \exp\left\{-\frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2 + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n\right\} U_{S_0}(\mathrm{d}\boldsymbol{\theta})\\
&\quad= \int_{\mathbb{R}^{|S_0|r+r(r+1)/2}} \exp\left\{-\frac{1}{2}\|\mathbf{Z}_{0S_0}\boldsymbol{\beta}_{S_0}\|_2^2 + \boldsymbol{\beta}_{S_0}^{\mathrm{T}}\mathbf{Z}_{0S_0}^{\mathrm{T}}\boldsymbol{\varepsilon}_n\right\}\mathrm{d}\boldsymbol{\beta}_{S_0}\\
&\quad= \zeta_{S_0}(\mathbb{R}^{|S_0|r+r(r+1)/2})\int_{\mathbb{R}^{|S_0|r+r(r+1)/2}} \exp(\boldsymbol{\beta}_{S_0}^{\mathrm{T}}\mathbf{Z}_{0S_0}^{\mathrm{T}}\boldsymbol{\varepsilon}_n)\bar{\sigma}_{S_0}(\mathrm{d}\boldsymbol{\beta}_{S_0})\\
&\quad\geq \zeta_{S_0}(\mathbb{R}^{|S_0|r+r(r+1)/2})\exp\left\{\int_{\mathbb{R}^{|S_0|r+r(r+1)/2}} (\boldsymbol{\beta}_{S_0}^{\mathrm{T}}\mathbf{Z}_{0S_0}^{\mathrm{T}}\boldsymbol{\varepsilon}_n)\mathrm{d}\bar{\sigma}_{S_0}(\mathrm{d}\boldsymbol{\beta}_{S_0})\right\}\\
&\quad= \zeta_{S_0}(\mathbb{R}^{|S_0|r+r(r+1)/2}) = \det\{2\pi(\mathbf{Z}_{0S_0}^{\mathrm{T}}\mathbf{Z}_{0S_0})^{-1}\}^{1/2},
\end{aligned}$$

where we have used the change of variable $\boldsymbol{\beta}_{S_0} = \boldsymbol{\theta}_{S_0} - \boldsymbol{\theta}_{0S_0}$, Jensen's inequality applied to the distribution $\bar{\sigma}$, and the fact that $\bar{\sigma}_{S_0}$ is symmetric about zero so that the expected value of $\boldsymbol{\beta}_{S_0}^{\mathrm{T}}\mathbf{Z}_{0S_0}^{\mathrm{T}}\boldsymbol{\varepsilon}_n$ with regard to $\bar{\sigma}(\mathrm{d}\boldsymbol{\beta}_{S_0})$ is 0. Hence, we obtain the following lower bound for the denominator D_n^∞:

$$\begin{aligned}
D_n^\infty &\geq \frac{\pi_p(|S_0|)}{\binom{p-r}{|S_0|}\gamma(|S_0|)}\int \exp\left\{\boldsymbol{\varepsilon}_n^{\mathrm{T}}\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2\right\} U_{S_0}(\mathrm{d}\boldsymbol{\theta})\\
&= \frac{\pi_p(|S_0|)}{\binom{p-r}{|S_0|}\gamma_{S_0}(|S_0|)}\eta_{S_0}(\mathbb{R}^{(p-r)r+r(r+1)/2}) \geq \frac{\pi_p(s_0)(2\pi)^{s_0r/2+r(r+1)/4}}{\exp(s_0\log p)\det(\mathbf{Z}_{0S_0}^{\mathrm{T}}\mathbf{Z}_{0S_0})^{1/2}}.
\end{aligned}$$

By the geometric-algorithmic mean inequality,

$$\begin{aligned}
&\det(\mathbf{Z}_{0S_0}^{\mathrm{T}}\mathbf{Z}_{0S_0})^{1/2}\\
&\quad= \left\{\prod_{i=1}^{|S_0|r+r(r+1)/2} \lambda_i(\mathbf{Z}_{0S_0}^{\mathrm{T}}\mathbf{Z}_{0S_0})\right\}^{1/2}\\
&\quad\leq \left\{\frac{1}{|S|r+r(r+1)/2}\sum_{i=1}^{|S_0|r+r(r+1)/2} \lambda_i(\mathbf{Z}_{0S_0}^{\mathrm{T}}\mathbf{Z}_{0S_0})\right\}^{|S|r/2+r(r+1)/4}\\
&\quad= \left\{\frac{\mathrm{tr}(\mathbf{Z}_{0S_0}^{\mathrm{T}}\mathbf{Z}_{0S_0})}{|S|r+r(r+1)/2}\right\}^{|S_0|r/2+r(r+1)/4} \leq (\|\mathbf{Z}_{0S_0}\|_2^2)^{s_0r/2+r(r+1)/4}\\
&\quad\leq \left(n\|D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\|_2^2/2\right)^{s_0r/2+r(r+1)/4} \leq \exp(Crs_0\log n)
\end{aligned}$$

for some constant $C > 0$. Therefore,

$$D_n^\infty \geq \pi_p(s_0)\exp\left(-s_0\log p - Crs_0\log n\right).$$

■ We next analyze the numerator N_n^∞. Write

$$\begin{aligned}
N_n^\infty &= \sum_{S\in\mathcal{S}_0}\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}\int_{(\widetilde{\mathcal{A}}_n^c)_S} e^{\boldsymbol{\varepsilon}_n^{\mathrm{T}}\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\boldsymbol{\theta}_{0S})-(1/2)\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\boldsymbol{\theta}_{0S})\|_2^2}\mathrm{d}\boldsymbol{\theta}_S\\
&= \sum_{S\in\mathcal{S}_0}\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}\int_{\widetilde{\mathcal{A}}_n^c} e^{\boldsymbol{\varepsilon}_n^{\mathrm{T}}\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)-(1/2)\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2}U_S(\mathrm{d}\boldsymbol{\theta})
\end{aligned}$$

Denote by $\Xi_n := \{|(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n| \leq \bar{\lambda}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1,\}$, where $\bar{\lambda} := C_0\sqrt{n\log p}$ is such that $\mathbb{P}_0(\Xi_n^c) \leq 2/p$ by Lemma 5.10. By definition of $\mathcal{A}_n$, for any $\boldsymbol{\theta}\in\widetilde{\mathcal{A}}_n^c$, we have, $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1 > M\sqrt{(r^2s^2\log n + rs^2\log p)/n}$. By Theorem 2.1, we have

$$\begin{aligned}
\sigma_{\min}^2(\mathbf{Z}_0) &= \frac{n}{2}\lambda_{\min}\left\{D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\Sigma}_0^{-1}\otimes\boldsymbol{\Sigma}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\right\}\\
&= \frac{n}{2}\min_{\|\boldsymbol{\theta}\|_2=1}\|(\boldsymbol{\Omega}_0^{-1/2}\otimes\boldsymbol{\Omega}_0^{-1/2})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\boldsymbol{\theta}\|_2^2\\
&\geq \frac{n}{2}\lambda_{\min}(\boldsymbol{\Omega}_0^{-1})^2\min_{\|\boldsymbol{\theta}\|_2=1}\|D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\boldsymbol{\theta}\|_2^2 \geq \frac{n\sigma_{\min}^2\{D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\}}{2\|\boldsymbol{\Omega}_0\|_2^2} \gtrsim n.
\end{aligned}$$

Then over the event Ξ_n, with $\mathrm{supp}(\mathbf{A})\in\mathcal{S}_0$ and $\boldsymbol{\theta} := [\mathrm{vec}(\mathbf{A})^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{T}}]^{\mathrm{T}}$, we have,

$$\begin{aligned}
(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n &\leq 2\bar{\lambda}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1 - \bar{\lambda}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1\\
&= 2\left\{\frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2\right\}^{1/2}\left\{\frac{\sqrt{2}\bar{\lambda}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1}{\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2}\right\} - \bar{\lambda}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1\\
&\leq \frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2 + \frac{2\bar{\lambda}^2\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1^2}{\sigma_{\min}(\mathbf{Z}_0)^2\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2} - \bar{\lambda}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1\\
&\leq \frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2 + \frac{\bar{C}_0 rs_0\log p\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2} - \bar{\lambda}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1\\
&= \frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2 + \bar{C}_0 rs_0\log p - \bar{\lambda}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1.
\end{aligned}$$

where $\bar{C}_0$ is a constant depending on the spectra of $\boldsymbol{\Omega}_0$. For any $\boldsymbol{\theta}\in\widetilde{\mathcal{A}}_n^c$, we have,

$$\begin{aligned}
\bar{\lambda}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1 &= C_0\sqrt{n\log p}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1 \geq C_0M\sqrt{r^2s^2(\log n)(\log p) + rs^2(\log p)^2}\\
&\geq C_1M(rs_0\log n + s_0\sqrt{r}\log p)
\end{aligned}$$

for some constant $C_1 > 0$. Therefore, by choosing a sufficiently large $M > 0$, we have, for any $\boldsymbol{\theta}\in\widetilde{\mathcal{A}}_n^c$,

$$\exp\left\{-\frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2 + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n\right\}$$

$$\leq \exp\left\{-\frac{C_1 M(rs_0\log n + s_0\sqrt{r}\log p)}{4} - \frac{\bar{\lambda}}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1\right\}$$

for some constant $C_1 > 0$ (possibly depending on the spectra of $\boldsymbol{\Omega}_0$), which further implies that over the event Ξ_n,

$$\begin{aligned}
N_n^\infty &= \sum_{S\in\mathcal{S}_0}\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}\int_{(\widetilde{\mathcal{A}}_n^c)_S}\exp\left\{\boldsymbol{\varepsilon}_n^{\mathrm{T}}\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_2^2\right\}U_S(\mathrm{d}\boldsymbol{\theta}) \\
&\leq \exp\left\{-\frac{C_1 M(rs_0\log n + s_0\sqrt{r}\log p)}{4}\right\} \\
&\quad\times \sum_{S\in\mathcal{S}_0}\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}\int_{\widetilde{\mathcal{A}}_n^c}\exp\left(-\frac{\bar{\lambda}}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_1\right)U_S(\mathrm{d}\boldsymbol{\theta}) \\
&= \exp\left\{-\frac{C_1 M(rs_0\log n + s_0\sqrt{r}\log p)}{4}\right\}\sum_{S\in\mathcal{S}_0}\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}\frac{4^{|S|r+r(r+1)/2}}{\bar{\lambda}^{|S|r+r(r+1)/2}} \\
&\leq \exp\left\{-\frac{C_1 M(rs_0\log n + s_0\sqrt{r}\log p)}{4}\right\}\sum_{t=0}^{\kappa_0 s_0}\frac{\pi_p(t)}{\gamma(t)} \\
&= \exp\left\{-\frac{C_1 M(rs_0\log n + s_0\sqrt{r}\log p)}{4}\right\}e^{(1/2)\kappa_0 s_0 r\log(\kappa_0 s_0 r) + 2\kappa_0 s_0 r} \\
&\leq \exp\left\{-\frac{C_1 M(rs_0\log n + s_0\sqrt{r}\log p)}{8}\right\}
\end{aligned}$$

for a sufficiently large $M > 0$, where we have used the fact that $\bar{\lambda}\infty 0$ as $n\to\infty$ and (21).

■ We are now finally able to analyze the ratio N_n^∞/D_n^∞. Write

$$\begin{aligned}
&\mathbb{E}_0\Pi_{\boldsymbol{\theta}}^\infty(\boldsymbol{\theta}\in\widetilde{\mathcal{A}}_n^c \mid \mathbf{Y}_n) \\
&\quad\leq \mathbb{P}_0(\Xi_n^c) + \mathbb{E}_0\mathbb{1}(\Xi_n)\frac{N_n^\infty}{D_n^\infty} \\
&\quad\leq \frac{2}{p} + \frac{1}{\pi_p(s_0)}\exp\left\{s_0\log p - \frac{C_1 M(rs_0\log n + s_0\sqrt{r}\log p)}{8} + Cs_0 r\log n\right\} \\
&\quad\leq 2\exp\left\{-\frac{C_0 M(rs_0\log n + s_0\sqrt{r}\log p)}{16}\right\} + \frac{2}{p}\lesssim\frac{1}{p}
\end{aligned}$$

by taking a sufficiently large $M > 0$ again. The proof is thus completed. □

We are now in a position to present the proof of Theorem 3.2.

Proof of Theorem 3.2. We first claim that condition A5 implies that

$$\mathbb{E}_0\Pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} : S_0\subset\mathrm{supp}(\mathbf{A}) \mid \mathbf{Y}_n)\to 1.$$

In fact, if $S_0 \cap \mathrm{supp}(\mathbf{A})^c \neq \varnothing$, then there exists some $j \in S_0$ such that $j \in \mathrm{supp}(\mathbf{A})^c$. Therefore,

$$\begin{aligned}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \geq \|\mathrm{vec}(\mathbf{A}) - \mathrm{vec}(\mathbf{A}_0)\|_{\mathrm{F}} &= \left\{\sum_{l=1}^p \|\mathbf{e}_l^{\mathrm{T}}(\mathbf{A} - \mathbf{A}_0)\|_2^2\right\}^{1/2} \\ &\geq \|[\mathbf{A}]_{j*} - [\mathbf{A}_0]_{j*}\|_2 = \|[\mathbf{A}_0]_{j*}\|_2 \geq \min_{j \in S_0} \|[\mathbf{A}_0]_{j*}\|_2.\end{aligned}$$

Denote by $\epsilon_n = \sqrt{(rs\log n + s\log p)/n}$. Using the result from Theorem 5.7, we have,

$$\mathbb{E}_0 \Pi_{\boldsymbol{\theta}} \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 < M\sqrt{\frac{rs\log n + s\log p}{n}} \;\middle|\; \mathbf{Y}_n\right\} \geq 1 - O(e^{-cn\epsilon_n^2})$$

for some constants $M, c > 0$. Hence,

$$\begin{aligned}&\mathbb{E}_0 \Pi_{\boldsymbol{\theta}}\{\boldsymbol{\theta} : S_0 \subset \mathrm{supp}(\mathbf{A}) \mid \mathbf{Y}_n\} \\ &\quad = 1 - \mathbb{E}_0 \Pi_{\boldsymbol{\theta}}\{\boldsymbol{\theta} : S_0 \cap \mathrm{supp}(\mathbf{A})^c \neq \varnothing \mid \mathbf{Y}_n\} \\ &\quad \geq 1 - \mathbb{E}_0 \Pi_{\boldsymbol{\theta}} \left(\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \geq \min_{j \in S_0} \|[\mathbf{A}_0]_{j*}\|_2 \mid \mathbf{Y}_n\right) \\ &\quad \geq 1 - \mathbb{E}_0 \Pi_{\boldsymbol{\theta}} \left(\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \geq M\sqrt{\frac{rs\log n + s\log p}{n}} \;\middle|\; \mathbf{Y}_n\right) \geq 1 - O(e^{-cn\epsilon_n^2}).\end{aligned}$$

Let

$$\begin{aligned}\mathcal{E}_n &= \bigcup_{S \in \mathcal{S}_0} \mathcal{A}_n(S) \\ &= \bigcup_{S \in \mathcal{S}_0} \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1 \leq M\sqrt{\frac{r^2s^2\log n + rs^2\log p}{n}}, \mathrm{supp}(\mathbf{A}) = S\right\}\end{aligned}$$

By Theorem 5.7, Lemma 5.4, and condition A5, we immediately see that

$$\mathbb{E}_0 \Pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in \mathcal{E}_n \mid \mathbf{Y}_n) \geq 1 - O(e^{-cn\epsilon_n^2}).$$

For notational convenience, we denote by $\mathbf{t}_n := \mathbf{Z}_0\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}_n$,

$$\begin{aligned}v_S &\propto \frac{\pi_p(|S|)}{\binom{p-r}{|S|}} \int \exp\{\ell(\boldsymbol{\Omega}(\boldsymbol{\theta})) - \ell(\boldsymbol{\Omega}_0)\}\Pi_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{\theta})\mathbb{1}\{S \in \mathcal{S}_0\}, \\ G_S(\mathrm{d}\boldsymbol{\theta}) &:= \{\exp(-2\|\boldsymbol{\theta}_S\|_1)\mathrm{d}\boldsymbol{\theta}_S\}\{\delta_{\mathbf{0}}(\mathrm{d}\boldsymbol{\theta}_{S^c})\}, \\ \mu_S(\mathrm{d}\boldsymbol{\theta}) &:= \frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)} \mathbb{1}\{\boldsymbol{\theta} \in \mathcal{A}_n(S)\} \exp\{\ell(\boldsymbol{\Omega}(\boldsymbol{\theta})) - \ell(\boldsymbol{\Omega}_0)\} G_S(\mathrm{d}\boldsymbol{\theta}), \\ \nu_S(\mathrm{d}\boldsymbol{\theta}) &:= \frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)} \exp\left(\frac{1}{2}\boldsymbol{\varepsilon}_n^{\mathrm{T}}\boldsymbol{\varepsilon}_n - \frac{1}{2}\|\mathbf{t}_n - \mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2 - 2\|\boldsymbol{\theta}_0\|_1\right)\end{aligned}$$

$$
\begin{aligned}
&\times \mathbb{1}\{\boldsymbol{\theta} \in \mathcal{A}_n(S)\} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)^{\mathrm{T}}\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S}(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)\right\} \mathrm{d}\boldsymbol{\theta}_S \\
&\times \delta_{\mathbf{0}}(\mathrm{d}\boldsymbol{\theta}_{S^c}), \\
\widetilde{\varpi}_S &:= \frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)} \exp\left(\frac{1}{2}\boldsymbol{\varepsilon}_n^{\mathrm{T}}\boldsymbol{\varepsilon}_n - \frac{1}{2}\|\mathbf{t}_n - \mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2 - 2\|\boldsymbol{\theta}_0\|_1\right) \mathbb{1}\{S \in \mathcal{S}_0\} \\
&\times \int_{(\mathcal{A}_n(S))_S} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)^{\mathrm{T}}\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S}(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)\right\} \mathrm{d}\boldsymbol{\theta}_S, \\
\widehat{\varpi}_S &:= \frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)} \exp\left(\frac{1}{2}\boldsymbol{\varepsilon}_n^{\mathrm{T}}\boldsymbol{\varepsilon}_n - \frac{1}{2}\|\mathbf{t}_n - \mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2 - 2\|\boldsymbol{\theta}_0\|_1\right) \mathbb{1}\{S \in \mathcal{S}_0\} \\
&\times \int_{\mathbb{R}^{|S|r + r(r+1)/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)^{\mathrm{T}}\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S}(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)\right\} \mathrm{d}\boldsymbol{\theta}_S, \\
Q_S(\mathrm{d}\boldsymbol{\theta}) &:= \{\phi(\boldsymbol{\theta}_S \mid \widehat{\boldsymbol{\theta}}_S, (\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1})\mathrm{d}\boldsymbol{\theta}_S\}\{\delta_{\mathbf{0}}(\mathrm{d}\boldsymbol{\theta}_{S^c})\}, \\
\widetilde{Q}_S(\mathrm{d}\boldsymbol{\theta}) &:= \frac{\mathbb{1}\{\boldsymbol{\theta} \in \mathcal{A}_n(S)\}Q_S(\mathrm{d}\boldsymbol{\theta})}{Q_S(\boldsymbol{\theta} \in \mathcal{A}_n(S))}.
\end{aligned}
$$

By Parseval's identity, we have

$$
\begin{aligned}
&\exp\left(-2\|\boldsymbol{\theta}_0\|_1 - \frac{1}{2}\|\mathbf{t}_n - \mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2\right) \\
&\quad = \exp\left\{-2\|\boldsymbol{\theta}_0\|_1 - \frac{1}{2}\|\mathbf{t}_n - \mathbf{Z}_{0S}(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\mathbf{Z}_{0S}\mathbf{t}_n\|_2^2\right\} \\
&\quad = \exp\left\{-2\|\boldsymbol{\theta}_0\|_1 - \frac{1}{2}\|\mathbf{t}_n\|_2^2 + \frac{1}{2}\|\mathbf{Z}_{0S}(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\mathbf{Z}_{0S}\mathbf{t}_n\|_2^2\right\} \\
&\quad = \exp\left(-2\|\boldsymbol{\theta}_0\|_1 - \frac{1}{2}\|\mathbf{t}_n\|_2^2\right) \exp\left(\frac{1}{2}\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2\right).
\end{aligned}
$$

Note that $\exp\{-2\|\boldsymbol{\theta}_0\|_1 - (1/2)\|\mathbf{t}_n\|_2^2\}$ does not depend on the supporting set S. Therefore, by definiton of $\widehat{w}_S$, we have

$$
\begin{aligned}
\widehat{w}_S &\propto \frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)} \det\{2\pi(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\}^{1/2} \\
&\quad \times \exp\left(\frac{1}{2}\|\boldsymbol{\varepsilon}_n\|_2^2 - 2\|\boldsymbol{\theta}_0\|_1 - \frac{1}{2}\|\mathbf{t}_n - \mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2\right) \\
&= \widehat{\varpi}_S.
\end{aligned}
$$

Namely, $\widehat{w}_S = \widehat{\varpi}_S / \sum_{T \in \mathcal{S}_0} \widehat{\varpi}_T$, $S \in \mathcal{S}_0$. For any probability distribution $\mathbb{P}(\cdot)$ and any event $\mathcal{A}$, we have, by the law of total probability,

$$
\begin{aligned}
\left\|\mathbb{P}(\cdot) - \frac{\mathbb{P}(\cdot \cap \mathcal{A})}{\mathbb{P}(\mathcal{A})}\right\|_{\mathrm{TV}} &= \sup_{\mathcal{B}} \left|\frac{\mathbb{P}(\mathcal{B} \cap \mathcal{A}^c)\mathbb{P}(\mathcal{A}) - \mathbb{P}(\mathcal{B} \cap \mathcal{A})\mathbb{P}(\mathcal{A}^c)}{\mathbb{P}(\mathcal{A})}\right| \\
&\le \sup_{\mathcal{B}} \frac{\mathbb{P}(\mathcal{B} \cap \mathcal{A}^c)\mathbb{P}(\mathcal{A}) + \sup_{\mathcal{B}} \mathbb{P}(\mathcal{B} \cap \mathcal{A})\mathbb{P}(\mathcal{A}^c)}{\mathbb{P}(\mathcal{A})} \le 2\mathbb{P}(\mathcal{A}^c).
\end{aligned}
$$

For any measurable set $\mathcal{B}\subset\mathscr{D}(p,r)$, the (exact) posterior probability of $\mathcal{B}$ given $\mathcal{E}_n$ and $\mathbf{Y}_n$ can be written as

$$\begin{aligned}
&\frac{\Pi_{\boldsymbol\theta}(\mathcal{B}\cap\mathcal{E}_n\mid\mathbf{Y}_n)}{\Pi_{\boldsymbol\theta}(\mathcal{E}_n\mid\mathbf{Y}_n)}\\
&\quad=\frac{\sum_{S\in\mathcal{S}_0}\pi_p(|S|)\binom{p-r}{|S|}^{-1}\gamma(|S|)^{-1}\int_{\mathcal{B}\cap\mathcal{E}_n}\exp\{\ell(\boldsymbol\Omega(\boldsymbol\theta))-\ell(\boldsymbol\Omega_0)\}G_S(\mathrm{d}\boldsymbol\theta)}{\sum_{S\in\mathcal{S}_0}\pi_p(|S|)\binom{p-r}{|S|}^{-1}\gamma(|S|)^{-1}\int_{\mathcal{E}_n}\exp\{\ell(\boldsymbol\Omega(\boldsymbol\theta))-\ell(\boldsymbol\Omega_0)\}G_S(\mathrm{d}\boldsymbol\theta)}\\
&\quad=\frac{\sum_{S\in\mathcal{S}_0}\pi_p(|S|)\binom{p-r}{|S|}^{-1}\gamma(|S|)^{-1}\int_{\mathcal{B}\cap\mathcal{A}_n(S)}\exp\{\ell(\boldsymbol\Omega(\boldsymbol\theta))-\ell(\boldsymbol\Omega_0)\}G_S(\mathrm{d}\boldsymbol\theta)}{\sum_{S\in\mathcal{S}_0}\pi_p(|S|)\binom{p-r}{|S|}^{-1}\gamma(|S|)^{-1}\int_{\mathcal{A}_n(S)}\exp\{\ell(\boldsymbol\Omega(\boldsymbol\theta))-\ell(\boldsymbol\Omega_0)\}G_S(\mathrm{d}\boldsymbol\theta)}\\
&\quad=\sum_{S\in\mathcal{S}_0}\frac{\mu_S(\mathcal{B})}{\|\sum_{S\in\mathcal{S}_0}\mu_S(\cdot)\|_{\mathrm{TV}}}.
\end{aligned}$$

By the triangle inequality, the total variation distance between $\Pi(\mathrm{d}\boldsymbol\theta\mid\mathbf{Y}_n)$ and $\Pi^\infty(\mathrm{d}\boldsymbol\theta\mid\mathbf{Y}_n)$ can be decomposed as follows:

$$\begin{aligned}
&\mathbb{E}_0\|\Pi_{\boldsymbol\theta}(\boldsymbol\theta\in\cdot\mid\mathbf{Y}_n)-\Pi^\infty_{\boldsymbol\theta}(\boldsymbol\theta\in\cdot\mid\mathbf{Y}_n)\|_{\mathrm{TV}}\\
&\quad\le\mathbb{E}_0\left\|\Pi_{\boldsymbol\theta}(\boldsymbol\theta\in\cdot\mid\mathbf{Y}_n)-\frac{\Pi_{\boldsymbol\theta}(\boldsymbol\theta\in\cdot\cap\mathcal{E}_n)}{\Pi_{\boldsymbol\theta}(\boldsymbol\theta\in\mathcal{E}_n)}\right\|_{\mathrm{TV}} &(28)\\
&\qquad+\mathbb{E}_0\left\|\sum_{S\in\mathcal{S}_0}\frac{\mu_S(\cdot)}{\|\sum_{S\in\mathcal{S}_0}\mu_S(\cdot)\|_{\mathrm{TV}}}-\sum_{S\in\mathcal{S}(\kappa_n)}\frac{\nu_S(\cdot)}{\|\sum_{S\in\mathcal{S}_0}\nu_S(\cdot)\|_{\mathrm{TV}}}\right\|_{\mathrm{TV}} &(29)\\
&\qquad+\mathbb{E}_0\left\|\sum_{S\in\mathcal{S}_0}\frac{\nu_S(\cdot)}{\|\sum_{S\in\mathcal{S}_0}\nu_S(\cdot)\|_{\mathrm{TV}}}-\Pi^\infty_{\boldsymbol\theta}(\boldsymbol\theta\in\cdot\mid\mathbf{Y}_n)\|\right\|_{\mathrm{TV}}. &(30)
\end{aligned}$$

The first term on the right-hand side is upper bounded by $2\mathbb{E}_0\Pi_{\boldsymbol\theta}(\boldsymbol\theta\in\mathcal{E}_n^c\mid\mathbf{Y}_n)$, which is $O(e^{-cn\epsilon_n^2})$ by Theorem 5.7, Lemma 5.4, and condition A5. It suffices to focus on the second and the third term. For the second term on line (29), write

$$\begin{aligned}
&\left\|\sum_S\frac{\mu_S}{\|\sum_S\mu_S\|_{\mathrm{TV}}}-\sum_S\frac{\nu_S}{\|\sum_S\nu_S\|_{\mathrm{TV}}}\right\|_{\mathrm{TV}}\\
&\quad=\left\|\frac{\|\sum_S\nu_S\|_{\mathrm{TV}}\sum_S(\mu_S-\nu_S)+(\|\sum_S\nu_S\|_{\mathrm{TV}}-\|\sum_S\mu_S\|_{\mathrm{TV}})\sum_S\nu_S}{\|\sum_S\mu_S\|_{\mathrm{TV}}\|\sum_S\nu_S\|_{\mathrm{TV}}}\right\|_{\mathrm{TV}}\\
&\quad\le\frac{2\sum_S\|\mu_S-\nu_S\|_{\mathrm{TV}}}{\|\sum_S\mu_S\|_{\mathrm{TV}}}=\frac{2}{\|\sum_S\mu_S\|_{\mathrm{TV}}}\sum_S\sup_{\mathcal{B}}\left|\int_{\mathcal{B}}\mathrm{d}\mu_S-\int_{\mathcal{B}}\left(\frac{\mathrm{d}\nu_S}{\mathrm{d}\mu_S}\right)\mathrm{d}\mu_S\right|\\
&\quad\le\frac{2}{\|\sum_S\mu_S\|_{\mathrm{TV}}}\sum_S\sup_{\mathcal{B}}\int_{\mathcal{B}}\left|1-\left(\frac{\mathrm{d}\nu_S}{\mathrm{d}\mu_S}\right)\right|\mathrm{d}\mu_S\\
&\quad\le\frac{2}{\|\sum_S\mu_S\|_{\mathrm{TV}}}\sum_S\sup_{\mathcal{B}}\mu_S(\mathcal{B})\left\|1-\left(\frac{\mathrm{d}\nu_S}{\mathrm{d}\mu_S}\right)\right\|_{L_\infty(\mathcal{A}_n(S))}\\
&\quad=2\sup_{S\in\mathcal{S}_0}\left\|1-\left(\frac{\mathrm{d}\nu_S}{\mathrm{d}\mu_S}\right)\right\|_{L_\infty(\mathcal{A}_n(S))}.
\end{aligned}$$

By definition of ν_S and μ_S, for any $\boldsymbol{\theta} \in \mathcal{A}_n(S)$, we have

$$-\log \frac{\mathrm{d}\nu_S}{\mathrm{d}\mu_S}(\boldsymbol{\theta}) = \ell(\boldsymbol{\Omega}(\boldsymbol{\theta})) - \ell(\boldsymbol{\Omega}_0) - \frac{1}{2}\|\boldsymbol{\varepsilon}_n\|_2^2 + \frac{1}{2}\|\mathbf{t}_n - \mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2 + \frac{1}{2}\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)\|_2^2 + 2\|\boldsymbol{\theta}_{0S}\|_1 - 2\|\boldsymbol{\theta}_S\|_1.$$

Observe that $\mathbf{Z}_{0S}(\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S)$ is inside the column space of $\mathbf{Z}_{0S}$, and $\mathbf{t}_n - \mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S$ lies in the orthogonal complement of the column space of $\mathbf{Z}_{0S}$, it follows from the Parseval's identity that for $\boldsymbol{\theta} \in \mathcal{A}_n(S)$,

$$\begin{aligned}\|\mathbf{t}_n - \mathbf{Z}_0\boldsymbol{\theta}\|_2^2 &= \|\mathbf{t}_n - \mathbf{Z}_{0S}\boldsymbol{\theta}_S\|_2^2 = \|\mathbf{t}_n - \mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S + \mathbf{Z}_{0S}(\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S)\|_2^2 \\ &= \|\mathbf{t}_n - \mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2 + \|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)\|_2^2.\end{aligned}$$

Denote by $\delta_n = \sqrt{(r^2 s^2 \log n + r s^2 \log p)^3/n}$. Using the fact that

$$(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n - \frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_2^2 = -\frac{1}{2}\|\mathbf{t}_n - \mathbf{Z}_0\boldsymbol{\theta}\|_2^2 + \frac{1}{2}\|\mathbf{t}_n - \mathbf{Z}_0\boldsymbol{\theta}_0\|_2^2,$$

we can further obtain

$$\begin{aligned}-\log \frac{\mathrm{d}\nu_S}{\mathrm{d}\mu_S}(\boldsymbol{\theta}) &= \ell(\boldsymbol{\Omega}(\boldsymbol{\theta})) - \ell(\boldsymbol{\Omega}_0) - \frac{1}{2}\|\mathbf{t}_n - \mathbf{Z}_0\boldsymbol{\theta}_0\|_2^2 + \frac{1}{2}\|\mathbf{t}_n - \mathbf{Z}_0\boldsymbol{\theta}\|_2^2 \\ &\quad + 2(\|\boldsymbol{\theta}_{0S}\|_1 - \|\boldsymbol{\theta}_S\|_1) \\ &= \ell(\boldsymbol{\Omega}(\boldsymbol{\theta})) - \ell(\boldsymbol{\Omega}_0) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\boldsymbol{\varepsilon}_n + \frac{1}{2}\|\mathbf{Z}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|_2^2 \\ &\quad + 2(\|\boldsymbol{\theta}_{0S}\|_1 - \|\boldsymbol{\theta}_S\|_1) \\ &= R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + 2(\|\boldsymbol{\theta}_{0S}\|_1 - \|\boldsymbol{\theta}_S\|_1),\end{aligned}$$

where the remainder R_n satisfies

$$\sup_{S \in \mathcal{S}_0} \sup_{\boldsymbol{\theta} \in \mathcal{A}_n(S)} |R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| \le \sup_{\boldsymbol{\theta} \in \mathcal{B}_n} |R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)| \le C\delta_n$$

with probability at least $1 - 2e^{-\kappa_0 n \epsilon_n^2}$ by Theorem 5.8. In addition, we also have

$$\begin{aligned}\sup_{S \in \mathcal{S}_0} \sup_{\boldsymbol{\theta} \in \mathcal{A}_n(S)} 2|\|\boldsymbol{\theta}_{0S}\|_1 - \|\boldsymbol{\theta}_S\|_1| &\le \sup_{S \in \mathcal{S}_0} \sup_{\boldsymbol{\theta} \in \mathcal{A}_n(S)} 2\|\boldsymbol{\theta}_S - \boldsymbol{\theta}_{0S}\|_1 \\ &\lesssim \sqrt{\frac{r^2 s^2 \log n + r s^2 \log p}{n}} = o(1)\end{aligned}$$

by definition. Therefore, the term on line (29) is upper bounded by

$$\begin{aligned}&2 \sup_{S \in \mathcal{S}_0} \left\|1 - \frac{\mathrm{d}\nu_S}{\mathrm{d}\mu_S}\right\|_{L_\infty(\mathcal{A}_n(S))} \\ &\quad = 2 \sup_{S \in \mathcal{S}_0} \|1 - \exp\{-R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - 2(\|\boldsymbol{\theta}_{0S}\|_1 - \|\boldsymbol{\theta}_S\|_1)\}\|_{L_\infty(\mathcal{A}_n(S))}\end{aligned}$$

$$\leq 2\left[\exp\left\{\sup_{S\in\mathcal{S}_0}\sup_{\boldsymbol{\theta}\in\mathcal{A}_n(S)}(|R_n(\boldsymbol{\theta},\boldsymbol{\theta}_0)|+2(\|\boldsymbol{\theta}_{0S}\|_1-\|\boldsymbol{\theta}_S\|_1))\right\}-1\right]\leq C\delta_n$$

with probability at least $1-2e^{-\kappa_0 n\epsilon_n^2}$, where we have used the fact that $e^x-1\leq 2x$ for all $x\in(0,1]$. This shows that the term on line (29) is $O(\delta_n+e^{-\kappa_0 n\epsilon_n^2})$. We now finally focus on the term on (30). Using the fact that $\widehat{w}_S=\widehat{\varpi}_S/\sum_{T\in\mathcal{S}_0}\widehat{\varpi}_T$, we have

$$\Pi^\infty_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{\theta}\mid\mathbf{Y}_n)=\sum_{S\in\mathcal{S}_0}\left(\frac{\widehat{\varpi}_S}{\sum_{S\in\mathcal{S}_0}\widehat{\varpi}_S}\right)Q_S(\mathrm{d}\boldsymbol{\theta})$$

and

$$\sum_{S\in\mathcal{S}_0}\frac{\nu_S(\mathrm{d}\boldsymbol{\theta})}{\|\sum_{S\in\mathcal{S}_0}\nu_S(\cdot)\|_{\mathrm{TV}}}=\sum_{S\in\mathcal{S}_0}\left(\frac{\widetilde{\varpi}_S}{\sum_{S\in\mathcal{S}_0}\widetilde{\varpi}_S}\right)\widetilde{Q}_S(\mathrm{d}\boldsymbol{\theta})$$

because by construction, $\nu_S(\mathrm{d}\boldsymbol{\theta})=\widetilde{\varpi}_S\widetilde{Q}_S(\mathrm{d}\boldsymbol{\theta})$. Note that for any measurable set $\mathcal{B}$, $Q_S(\mathcal{B}\cap\mathcal{E}_n)=Q_S(\mathcal{B}\cap\mathcal{A}_n(S))$, and by definition, $\widetilde{Q}_S(\mathcal{B})=Q_S(\mathcal{B}\cap\mathcal{A}_n(S))/Q_S(\mathcal{A}_n(S))$. Therefore,

$$\begin{aligned}\frac{\Pi^\infty_{\boldsymbol{\theta}}(\boldsymbol{\theta}\in\mathcal{B}\cap\mathcal{E}_n\mid\mathbf{Y}_n)}{\Pi^\infty_{\boldsymbol{\theta}}(\boldsymbol{\theta}\in\mathcal{E}_n\mid\mathbf{Y}_n)}&=\frac{\sum_{S\in\mathcal{S}_0}\widehat{w}_SQ_S(\mathcal{B}\cap\mathcal{E}_n)}{\sum_{S\in\mathcal{S}_0}\widehat{w}_SQ_S(\mathcal{E}_n)}=\frac{\sum_{S\in\mathcal{S}_0}\widehat{\varpi}_SQ_S(\mathcal{B}\cap\mathcal{A}_n(S))}{\sum_{S\in\mathcal{S}_0}\widehat{\varpi}_SQ_S(\mathcal{A}_n(S))}\\&=\frac{\sum_{S\in\mathcal{S}_0}\widehat{\varpi}_SQ_S(\mathcal{A}_n(S))\widetilde{Q}_S(\mathcal{B})}{\sum_{S\in\mathcal{S}_0}\widehat{\varpi}_SQ_S(\mathcal{A}_n(S))}.\end{aligned}$$

Furthermore, by Parseval's identity, we have

$$\begin{aligned}&\widehat{\varpi}_SQ_S(\mathcal{A}_n(S))\\&\quad=\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}e^{\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2/2}\det\{2\pi(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\}^{1/2}Q_S(\mathcal{A}_n(S))\\&\qquad\times\exp\left(\frac{1}{2}\|\boldsymbol{\varepsilon}_n\|_2^2-2\|\boldsymbol{\theta}_0\|_1-\frac{1}{2}\|\mathbf{t}_n\|_2^2\right)\\&\quad=\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}e^{\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2/2}\int_{\mathcal{A}_n}\exp\left\{-\frac{1}{2}\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\widehat{\boldsymbol{\theta}}_S)\|_2^2\right\}\mathrm{d}\boldsymbol{\theta}_S\times\delta_{\mathbf{0}}(\mathrm{d}\boldsymbol{\theta}_{S^c})\\&\qquad\times\exp\left(\frac{1}{2}\|\boldsymbol{\varepsilon}_n\|_2^2-2\|\boldsymbol{\theta}_0\|_1-\frac{1}{2}\|\mathbf{t}_n\|_2^2\right)\\&\quad=\frac{\pi_p(|S|)}{\binom{p-r}{|S|}\gamma(|S|)}e^{\|\mathbf{Z}_{0S}\widehat{\boldsymbol{\theta}}_S\|_2^2/2}\int_{(\mathcal{A}_n(S))_S}\exp\left\{-\frac{1}{2}\|\mathbf{Z}_{0S}(\boldsymbol{\theta}_S-\widehat{\boldsymbol{\theta}}_S)\|_2^2\right\}\mathrm{d}\boldsymbol{\theta}_S\\&\qquad\times\exp\left(\frac{1}{2}\|\boldsymbol{\varepsilon}_n\|_2^2-2\|\boldsymbol{\theta}_0\|_1-\frac{1}{2}\|\mathbf{t}_n\|_2^2\right)\\&\quad=\widetilde{\varpi}_S,\end{aligned}$$

implying that

$$\frac{\Pi_{\boldsymbol\theta}^{\infty}(\boldsymbol\theta\in\mathcal{B}\cap\mathcal{E}_n\mid\mathbf{Y}_n)}{\Pi_{\boldsymbol\theta}^{\infty}(\mathcal{E}_n\mid\mathbf{Y}_n)}=\frac{\sum_{S\in\mathcal{S}_0}\widehat{\varpi}_S Q_S(\mathcal{A}_n(S))\widetilde{Q}_S(\mathcal{B})}{\sum_{S\in\mathcal{S}_0}\widehat{\varpi}_S Q_S(\mathcal{A}_n(S))}$$
$$=\sum_{S\in\mathcal{S}_0}\left(\frac{\widetilde{\varpi}_S}{\sum_{S\in\mathcal{S}_0}\widetilde{\varpi}_S}\right)\widetilde{Q}_S(\mathcal{B})=\frac{\sum_{S\in\mathcal{S}_0}\nu_S(\mathcal{B})}{\|\sum_{S\in\mathcal{S}_0}\nu_S(\cdot)\|_{\mathrm{TV}}}.$$

Hence, by Lemma 5.11, we know that the term on line (30) is upper bounded by

$$\mathbb{E}_0\left\|\frac{\sum_{S\in\mathcal{S}_0}\nu_S(\cdot)}{\|\sum_{S\in\mathcal{S}_0}\nu_S(\cdot)\|_{\mathrm{TV}}}-\Pi_{\boldsymbol\theta}^{\infty}(\boldsymbol\theta\in\cdot\mid\mathbf{Y}_n)\right\|_{\mathrm{TV}}$$
$$=\mathbb{E}_0\left\|\frac{\Pi_{\boldsymbol\theta}^{\infty}(\boldsymbol\theta\in\cdot\cap\mathcal{E}_n\mid\mathbf{Y}_n)}{\Pi_{\boldsymbol\theta}^{\infty}(\boldsymbol\theta\in\mathcal{E}_n\mid\mathbf{Y}_n)}-\Pi_{\boldsymbol\theta}^{\infty}(\boldsymbol\theta\in\cdot\mid\mathbf{Y}_n)\right\|_{\mathrm{TV}}$$
$$\leq 2\mathbb{E}_0\Pi_{\boldsymbol\theta}^{\infty}(\boldsymbol\theta\in\mathcal{E}_n^c\mid\mathbf{Y}_n)\lesssim\frac{1}{p}.$$

The proof is thus completed. □

5.7. *Posterior contraction under spectral norm*

Proof of Theorem 3.1. The proof of Theorem 3.1 is based on Theorem 3.2 and a discretization trick for the spectral norm loss. By Davis-Kahan theorem [13], $\|\sin\Theta(\mathbf{U}(\boldsymbol\varphi),\mathbf{U}_0)\|_2\leq 2\|\boldsymbol\Sigma(\boldsymbol\theta)-\boldsymbol\Sigma_0\|_2/\lambda_r(\boldsymbol\Sigma_0)$, so it suffices to consider the posterior contraction under $\|\boldsymbol\Omega(\boldsymbol\theta)-\boldsymbol\Omega_0\|_2=\|\boldsymbol\Sigma(\boldsymbol\theta)-\boldsymbol\Sigma_0\|_2$. Because

$$\mathbb{E}_0\Pi_{\boldsymbol\theta}\left\{\|\boldsymbol\Omega(\boldsymbol\theta)-\boldsymbol\Omega_0\|_2>M\sqrt{\frac{s\log p}{n}}\;\middle|\;\mathbf{Y}_n\right\}$$
$$\leq\mathbb{E}_0\Pi_{\boldsymbol\theta}^{\infty}\left\{\|\boldsymbol\Sigma(\boldsymbol\theta)-\boldsymbol\Sigma_0\|_2>M\sqrt{\frac{s\log p}{n}}\;\middle|\;\mathbf{Y}_n\right\}$$
$$+\mathbb{E}_0\|\Pi_{\boldsymbol\theta}(\mathrm{d}\boldsymbol\theta\mid\mathbf{Y}_n)-\Pi_{\boldsymbol\theta}^{\infty}(\mathrm{d}\boldsymbol\theta\mid\mathbf{Y}_n)\|_{\mathrm{TV}},$$

and $\mathbb{E}_0\|\Pi_{\boldsymbol\theta}(\mathrm{d}\boldsymbol\theta\mid\mathbf{Y}_n)-\Pi_{\boldsymbol\theta}^{\infty}(\mathrm{d}\boldsymbol\theta\mid\mathbf{Y}_n)\|_{\mathrm{TV}}=O(\delta_n)$ by Theorem 3.2, where $\delta_n=\sqrt{(r^2s^2\log n+rs^2\log p)^3/n}$. Therefore, it suffices to focus on

$$\mathbb{E}_0\Pi_{\boldsymbol\theta}^{\infty}\left\{\|\boldsymbol\Sigma(\boldsymbol\theta)-\boldsymbol\Sigma_0\|_2>M\sqrt{\frac{s\log p}{n}}\;\middle|\;\mathbf{Y}_n\right\}.$$

For any $S\in\mathcal{S}_0$, denote by $Q_S(\mathrm{d}\boldsymbol\theta):=\{\phi(\boldsymbol\theta_S\mid\widehat{\boldsymbol\theta}_S,(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1})\mathrm{d}\boldsymbol\theta_S\}\{\delta_{\mathbf{0}}(\mathrm{d}\boldsymbol\theta_{S^c})\}$, $\boldsymbol\Sigma_S(\boldsymbol\theta_S):=\mathbf{U}(\boldsymbol\varphi_S)\mathbf{M}\mathbf{U}(\boldsymbol\varphi_S)$, where $\boldsymbol\theta_S:=[\boldsymbol\varphi_S^{\mathrm{T}},\boldsymbol\mu^{\mathrm{T}}]^{\mathrm{T}}=[\mathrm{vec}(\mathbf{A}_S)^{\mathrm{T}},\boldsymbol\mu^{\mathrm{T}}]^{\mathrm{T}}$, and $\mathbf{U}(\boldsymbol\varphi_S)$ denotes the Cayley parameterization of $\boldsymbol\varphi_S=\mathrm{vec}(\mathbf{A}_S)$ from $\mathbb{R}^{|S|r}$ to $\mathbb{O}(|S|+r,r)$. Then from the proof of Lemma 5.9, we see that for any $\boldsymbol\theta=$

$[\text{vec}(\mathbf{A})^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{T}}]^{\mathrm{T}} = [\boldsymbol{\varphi}^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{T}}]^{\mathrm{T}}$ with $\text{supp}(\mathbf{A}) = S \supset S_0$, there exists a permutation matrix $\mathbf{P}_S$ such that

$$\mathbf{A} = \mathbf{P}_S \begin{bmatrix} \mathbf{A}_S \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{A}_0 = \mathbf{P}_S \begin{bmatrix} \mathbf{A}_{0S} \\ \mathbf{0} \end{bmatrix},$$

which further implies that

$$\mathbf{U}(\boldsymbol{\varphi}) = \mathbf{Q}_S \begin{bmatrix} \mathbf{U}(\boldsymbol{\varphi}_S) \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{U}_0 = \mathbf{Q}_S \begin{bmatrix} \mathbf{U}(\boldsymbol{\varphi}_{0S}) \\ \mathbf{0} \end{bmatrix},$$

where $\mathbf{Q}_S = \text{diag}(\mathbf{I}_r, \mathbf{P}_S)$, and $\boldsymbol{\varphi}_{0S} = \text{vec}(\mathbf{A}_{0S})$ for an appropriate $\mathbf{A}_{0S} \in \mathbb{R}^{|S| \times r}$. Therefore, for any $\boldsymbol{\theta} = [\text{vec}(\mathbf{A})^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{T}}]^{\mathrm{T}}$ with $\text{supp}(\mathbf{A}) = S \supset S_0$, we have

$$\begin{aligned} \|\boldsymbol{\Sigma}(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_0\|_2 &= \|\mathbf{U}(\boldsymbol{\varphi})\mathbf{M}\mathbf{U}(\boldsymbol{\varphi})^{\mathrm{T}} - \mathbf{U}_0\mathbf{M}_0\mathbf{U}_0^{\mathrm{T}}\|_2 \\ &= \left\| \mathbf{Q}_S \begin{bmatrix} \mathbf{U}(\boldsymbol{\varphi}_S)\mathbf{M}\mathbf{U}(\boldsymbol{\varphi}_S) - \mathbf{U}(\boldsymbol{\varphi}_{0S})\mathbf{M}_0\mathbf{U}(\boldsymbol{\varphi}_{0S})^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Q}_S^{\mathrm{T}} \right\|_2 \\ &= \|\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\|_2, \end{aligned}$$

where $\boldsymbol{\theta}_{0S} = [\boldsymbol{\varphi}_{0S}^{\mathrm{T}}, \boldsymbol{\mu}_0^{\mathrm{T}}]^{\mathrm{T}}$.

Now we proceed to analyze the probability of the event

$$\left\{ \|\boldsymbol{\Sigma}(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_0\|_2 > M\sqrt{\frac{s_0 \log p}{n}} \right\}$$

under the $\Pi_{\boldsymbol{\theta}}^{\infty}(\mathrm{d}\boldsymbol{\theta} \mid \mathbf{Y}_0)$ distribution. By Lemma 5.11, there exists some constants $M_1 > 0$, such that $\Pi_{\boldsymbol{\theta}}^{\infty}(\boldsymbol{\theta} \in \widetilde{\mathcal{A}}_n^c \mid \mathbf{Y}_n) = o_{\mathbb{P}_0}(1)$, where

$$\widetilde{\mathcal{A}}_n := \left\{ \boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1 \le M_1 \sqrt{\frac{r^2 s^2 \log n + r s^2 \log p}{n}} \right\}.$$

Therefore,

$$\begin{aligned} &\mathbb{E}_0 \Pi_{\boldsymbol{\theta}}^{\infty} \left\{ \|\boldsymbol{\Sigma}(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_0\|_2 > M\sqrt{\frac{s \log p}{n}} \,\middle|\, \mathbf{Y}_n \right\} \\ &\quad \le \mathbb{E}_0 \Pi_{\boldsymbol{\theta}}^{\infty} \left\{ \|\boldsymbol{\Sigma}(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_0\|_2 > M\sqrt{\frac{s \log p}{n}}, \boldsymbol{\theta} \in \widetilde{\mathcal{A}}_n \,\middle|\, \mathbf{Y}_n \right\} \\ &\qquad + \mathbb{E}_0 \Pi_{\boldsymbol{\theta}}^{\infty} \left(\boldsymbol{\theta} \in \widetilde{\mathcal{A}}_n^c \mid \mathbf{Y}_n \right) \\ &\quad = \mathbb{E}_0 \sum_{S \in \mathcal{S}_0} \widehat{w}_S Q_S \left\{ \boldsymbol{\theta} \in \widetilde{\mathcal{A}}_n : \|\boldsymbol{\Sigma}(\boldsymbol{\theta}) - \boldsymbol{\Sigma}_0\|_2 > M\sqrt{\frac{s \log p}{n}} \right\} + O(1/p) \\ &\quad = \mathbb{E}_0 \sum_{S \in \mathcal{S}_0} \widehat{w}_S Q_S \left\{ \boldsymbol{\theta} \in \widetilde{\mathcal{A}}_n : \|\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\|_2 > M\sqrt{\frac{s \log p}{n}} \right\} + O(1/p) \end{aligned}$$

$$= \mathbb{E}_0 \sum_{S\in\mathcal{S}_0} \widehat{w}_S Q_S(\mathcal{B}_n(S)) + O(1/p),$$

where

$$\mathcal{B}_n(S) := \left\{\boldsymbol{\theta} : \|\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\|_2 > M\sqrt{\frac{s_0\log p}{n}}, \boldsymbol{\theta}\in\widetilde{\mathcal{A}}_n, \mathrm{supp}(\mathbf{A}) = S\right\}.$$

Now let $p(S) := |S| + r$, $\mathcal{S}^{p(S)-1} := \{\mathbf{v}\in\mathbb{R}^{p(S)} : \|\mathbf{v}\|_2 = 1\}$ be the unit sphere in $\mathbb{R}^{p(S)}$, and let $\mathcal{S}^{p(S)-1}(1/5)$ be a 1/5-net of $\mathcal{S}^{p(S)-1}$ with smallest cardinality, namely, for any $\mathbf{v}\in\mathcal{S}^{p(S)-1}$, there exists some $\mathbf{u}(\mathbf{v})\in\mathcal{S}^{p(S)-1}(1/5)$, such that $\|\mathbf{u}(\mathbf{v}) - \mathbf{v}\|_2 < 1/5$. It follows that

$$\begin{aligned}
&\|\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\|_2\\
&\quad= \max_{\mathbf{v}\in\mathcal{S}^{p(S)-1}} |\mathbf{v}^{\mathrm{T}}\{\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\}\mathbf{v}|\\
&\quad= \max_{\mathbf{v}\in\mathcal{S}^{p(S)-1}} |\{\mathbf{v} - \mathbf{u}(\mathbf{v}) + \mathbf{u}(\mathbf{v})\}^{\mathrm{T}}\{\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\}\{\mathbf{v} - \mathbf{u}(\mathbf{v}) + \mathbf{u}(\mathbf{v})\}|\\
&\quad\leq \max_{\mathbf{v}\in\mathcal{S}^{p(S)-1}} \{2\|\mathbf{v} - \mathbf{u}(\mathbf{v})\|_2 + \|\mathbf{v} - \mathbf{u}(\mathbf{v})\|_2^2\}\|\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\|_2\\
&\qquad+ \max_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)} |\mathbf{u}^{\mathrm{T}}\{\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\}\mathbf{u}|\\
&\quad\leq \frac{1}{2}\|\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\|_2 + \max_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)} |\mathbf{u}^{\mathrm{T}}\{\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\}\mathbf{u}|,
\end{aligned}$$

implying that

$$\|\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\|_2 \leq 2\max_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)} |\mathbf{u}^{\mathrm{T}}\{\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\}\mathbf{u}|.$$

In addition, we also observe that there exists some constant $c > 0$ such that $\log|\mathcal{S}^{p(S)-1}(1/5)| \leq cp(S) = |S| + r \leq c\kappa_0 s_0$. Clearly, for any $\boldsymbol{\theta}\in\mathcal{B}_n(S)$, $\|\boldsymbol{\theta}_S - \boldsymbol{\theta}_{0S}\|_2 \leq \|\boldsymbol{\theta}_S - \boldsymbol{\theta}_{0S}\|_1 \leq M_1(\boldsymbol{\theta})\epsilon_n(\boldsymbol{\theta}) \to 0$. Then by Theorem 3.1 in [55], we have

$$\mathrm{vec}\{\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\} = D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\boldsymbol{\theta}_S - \boldsymbol{\theta}_{0S}) + \mathrm{vec}\{\mathbf{R}_S(\boldsymbol{\theta}_S, \boldsymbol{\theta}_{0S})\},$$

where

$$\|\mathbf{R}_S(\boldsymbol{\theta}_S, \boldsymbol{\theta}_{0S})\|_{\mathrm{F}} \lesssim \|\boldsymbol{\theta}_S - \boldsymbol{\theta}_{0S}\|_2^2 \leq \|\boldsymbol{\theta}_S - \boldsymbol{\theta}_{0S}\|_1^2 \lesssim \frac{r^2s^2\log n + rs^2\log p}{n}$$

whenever $\boldsymbol{\theta}\in\mathcal{B}_n(S)$. Hence, for all $\boldsymbol{\theta}\in\mathcal{B}_n(S)$ and all $\mathbf{u}\in\mathcal{S}^{p(S)-1}$,

$$\begin{aligned}
&|\mathbf{u}^{\mathrm{T}}\{\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\}\mathbf{u}|\\
&\quad\leq |(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}} D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\boldsymbol{\theta}_S - \boldsymbol{\theta}_{0S})| + \|\mathbf{R}_S(\boldsymbol{\theta}_S, \boldsymbol{\theta}_{0S})\|_{\mathrm{F}}.
\end{aligned}$$

Note that

$$\frac{r^2s^2\log n + rs^2\log p}{n} \lesssim \sqrt{\frac{s\log p}{n}}\sqrt{\max\left\{\frac{r^4s^3(\log n)^2}{n\log p}, \frac{r^2s^3\log p}{n}\right\}}$$

$$
\begin{aligned}
&\lesssim \sqrt{\frac{s\log p}{n}\max\left\{\frac{r^4s^3(\log n)}{n}, \frac{r^2s^3\log p}{n}\right\}}\\
&\lesssim \sqrt{\frac{s\log p}{n}}\sqrt{\max\left\{\frac{(r^2s^2\log n)^3}{n}, \frac{(rs^2\log p)^3}{n}\right\}}\\
&= o(1)\sqrt{\frac{s_0\log p}{n}}
\end{aligned}
$$

because of condition A4. This implies that $\|\mathbf{R}_S(\boldsymbol{\theta}_S, \boldsymbol{\theta}_{0S})\|_{\mathrm{F}} = o(\sqrt{(s\log p)/n})$ whenever $\boldsymbol{\theta} \in \widetilde{\mathcal{A}}_n$. Hence, by the union bound, we further write

$$
\begin{aligned}
&Q_S(\mathcal{B}_n(S))\\
&\quad\le Q_S\left\{\max_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)}|\mathbf{u}^{\mathrm{T}}\{\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\}\mathbf{u}| > \frac{M}{2}\sqrt{\frac{s\log p}{n}}, \boldsymbol{\theta}\in\widetilde{\mathcal{A}}_n\right\}\\
&\quad\le \sum_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)} Q_S\left\{|\mathbf{u}^{\mathrm{T}}\{\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_S) - \boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\}\mathbf{u}| > \frac{M}{2}\sqrt{\frac{s\log p}{n}}, \boldsymbol{\theta}\in\widetilde{\mathcal{A}}_n\right\}\\
&\quad\le \sum_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)} Q_S\left\{|(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)| > \frac{M}{8}\sqrt{\frac{s\log p}{n}}\right\}\\
&\qquad+ \sum_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)} Q_S\left\{|(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S})| > \frac{M}{8}\sqrt{\frac{s\log p}{n}}\right\}\\
&\quad= \sum_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)} Q_S\left\{|(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)| > \frac{M}{8}\sqrt{\frac{s\log p}{n}}\right\}\\
&\qquad+ \sum_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)} \mathbb{1}\left\{|(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S})| > \frac{M}{8}\sqrt{\frac{s\log p}{n}}\right\}.
\end{aligned}
$$

Therefore, we obtain

$$
\begin{aligned}
\sum_{S\in\mathcal{S}_0}\mathbb{E}_0\widehat{w}_S Q_S(\mathcal{B}_n(S)) &\le \sum_{S\in\mathcal{S}_0}\sum_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)}\mathbb{E}_0 Q_S\left\{\mathcal{C}_n(\mathbf{u}, S)\right\} &(31)\\
&\quad+\sum_{S\in\mathcal{S}_0}\sum_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)}\mathbb{P}_0\{\mathcal{D}_n(\mathbf{u}, S)\}, &(32)
\end{aligned}
$$

where for any $\mathbf{S}\in\mathcal{S}_0$ and $\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)$, we define

$$
\begin{aligned}
\mathcal{C}_n(\mathbf{u}, \mathbf{S}) &:= \left\{(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) : |(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\boldsymbol{\theta}_S - \widehat{\boldsymbol{\theta}}_S)| > \frac{M}{8}\sqrt{\frac{s\log p}{n}}\right\}\\
\mathcal{D}_n(\mathbf{u}, \mathbf{S}) &:= \left\{\widehat{\boldsymbol{\theta}} : |(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S})| > \frac{M}{8}\sqrt{\frac{s\log p}{n}}\right\}.
\end{aligned}
$$

We analyze the two terms on line (31) and line (32) separately.

▲ For the term on line (31), we use the fact that Q_S is a (degenerate) multivariate normal distribution and write

$$Q_S\left\{|(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\boldsymbol{\theta}_S-\widehat{\boldsymbol{\theta}}_S)|>\frac{M}{8}\sqrt{\frac{s_0\log p}{n}}\right\}$$
$$=\mathbb{P}_{\omega_S}\left(|\omega_S|>\frac{M}{8}\sqrt{\frac{s_0\log p}{n}}\right),$$

where condition on the data $\mathbf{Y}_n$, and hence, $\widehat{\boldsymbol{\theta}}_S$,

$$\omega_S:=(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\boldsymbol{\theta}_S-\widehat{\boldsymbol{\theta}}_S)\overset{Q_S}{\sim}\mathrm{N}\left(0,V_S\right),$$
$$V_S:=(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})^{\mathrm{T}}(\mathbf{u}\otimes\mathbf{u}).$$

Note that by Theorem 2.1 and condition A2,

$$\begin{aligned}\|(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\|_2&=\frac{2}{n}\|\{\mathbf{F}_S^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{F}_S\}^{-1}\|_2\\&=\frac{2}{n}\sigma_{\min}^{-1}\{\mathbf{F}_S^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{F}_S\}\\&=\frac{2}{n}\left[\min_{\|\boldsymbol{\theta}_S\|_2=1}\boldsymbol{\theta}_S^{\mathrm{T}}\mathbf{F}_S^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{F}_S\boldsymbol{\theta}_S\right]^{-1}\\&\le\frac{2}{n}\left[\lambda_{\min}(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})\sigma_{\min}^2\{D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\}\sigma_{\min}^2(\mathbf{F}_S)\right]^{-1}=O(1/n),\end{aligned}$$

implying that $V_S\le\|D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\|_2^2\|(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\|_2\le C/n$ for some constant $C>0$. By Chernoff bound and the fact that $\omega_S\overset{\mathcal{L}}{=}-\omega_S$ under Q_S, we further have

$$\begin{aligned}&\mathbb{P}_{\omega_S}\left(|\omega_S|>\frac{M}{8}\sqrt{\frac{s\log p}{n}}\right)\\&\quad=\mathbb{P}_{\omega_S}\left(\omega_S>\frac{M}{8}\sqrt{\frac{s\log p}{n}}\right)+\mathbb{P}_{\omega_S}\left(\omega_S<-\frac{M}{8}\sqrt{\frac{s\log p}{n}}\right)\\&\quad=2\mathbb{P}_{\omega_S}\left(\omega_S\sqrt{ns\log p}>\frac{Ms\log p}{8}\right)\\&\quad\le\frac{2\mathbb{E}_{\omega_S}\exp(\omega_S\sqrt{ns\log p})\}}{\exp\{(Ms\log p)/8}=\frac{2\exp\{(V_Sns\log p)/2\}}{\exp\{(Ms\log p)/8\}}\\&\quad\le\frac{2\exp\{(Cs\log p)/2\}}{\exp\{(Ms\log p)/8\}}=2\exp\left\{-\left(\frac{M}{8}-\frac{C}{2}\right)s\log p\right\}.\end{aligned}$$

Therefore, the term on line (31) is upper bounded by

$$2\sum_{S\in\mathcal{S}_0}\sum_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)}\exp\left\{-\left(\frac{M}{8}-\frac{C}{2}\right)s\log p\right\}$$

$$\leq 2\sum_{t=s_0}^{\kappa_0 s_0}\binom{p-r}{t}\exp(c\kappa_0 s_0)\exp\left\{-\left(\frac{M}{8}-\frac{C}{2}\right)s\log p\right\}$$

$$\leq 2\kappa_0 s_0 \exp\left\{\kappa_0 s_0\log p + c\kappa_0 s_0 - \left(\frac{M}{8}-\frac{C}{2}\right)s\log p\right\} \leq e^{-(Ms\log p)/16}$$

by taking a sufficiently large $M > 0$ because $s = s_0 + r$.

▲ We are now left with the concentration of $\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S}$ on line (32). Since $S_0 \subset S$, it follows that

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S} &= (\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\mathbf{Z}_{0S}^{\mathrm{T}}(\mathbf{Z}_{0S}\boldsymbol{\theta}_{0S} + \boldsymbol{\varepsilon}_n) - \boldsymbol{\theta}_{0S} = (\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\mathbf{Z}_{0S}^{\mathrm{T}}\boldsymbol{\varepsilon}_n\\
&= \sqrt{\frac{n}{2}}(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\mathbf{F}_S^{\mathrm{T}}\mathbf{Z}_0^{\mathrm{T}}\mathrm{vec}\{\boldsymbol{\Omega}_0^{-1/2}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1/2}\}\\
&= \frac{n}{2}(\mathbf{Z}_{0S}^{\mathrm{T}}\mathbf{Z}_{0S})^{-1}\mathbf{F}_S^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}\mathrm{vec}\{\boldsymbol{\Omega}_0^{-1}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\}\\
&= \left\{\mathbf{F}_S^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{F}_S\right\}^{-1}\mathbf{F}_S^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}\\
&\quad\times\mathrm{vec}\{\boldsymbol{\Omega}_0^{-1}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\}.\end{aligned}$$

Therefore,

$$(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S}) = \boldsymbol{\beta}_S^{\mathrm{T}}\mathrm{vec}\{\boldsymbol{\Omega}_0^{-1}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\},$$

where

$$\begin{aligned}\boldsymbol{\beta}_S^{\mathrm{T}} := (\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})&\left\{\mathbf{F}_S^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{F}_S\right\}^{-1}\\
&\times\mathbf{F}_S^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}.\end{aligned}$$

Consider a $p\times p$ matrix $\mathbf{B}_S$ such that $\mathrm{vec}(\mathbf{B}_S) = \boldsymbol{\beta}_S$. It follows that

$$\begin{aligned}\rho_S &:= (\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S}) = \mathrm{vec}(\mathbf{B}_S)^{\mathrm{T}}\mathrm{vec}\{\boldsymbol{\Omega}_0^{-1}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\}\\
&= \mathrm{tr}\left\{\mathbf{B}_S^{\mathrm{T}}\boldsymbol{\Omega}_0^{-1}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\right\} = \mathrm{tr}\left\{\widetilde{\mathbf{B}}_S\boldsymbol{\Omega}_0^{-1}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1}\right\}\\
&= \mathrm{tr}\left\{\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S\boldsymbol{\Omega}_0^{-1}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\right\}\\
&= -\mathrm{tr}(\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S) + \frac{1}{n}\mathrm{tr}\left(\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S\boldsymbol{\Omega}_0^{-1}\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\right),\end{aligned}$$

where $\widetilde{\mathbf{B}}_S$ is the symmetrization of $\mathbf{B}_S$ defined by $\widetilde{\mathbf{B}}_S := (\mathbf{B}_S + \mathbf{B}_S^{\mathrm{T}})/2$. Also note that

$$\begin{aligned}&\|\widetilde{\mathbf{B}}_S\|_{\mathrm{F}}\\
&\quad\leq \|\mathbf{B}_S\|_{\mathrm{F}} = \|\boldsymbol{\beta}_S\|_2\\
&\quad\leq \|D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\|_2\|D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})\|_2\|\{\mathbf{F}_S^{\mathrm{T}}D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\Omega}_0^{-1}\otimes\boldsymbol{\Omega}_0^{-1})D\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{F}_S\}^{-1}\|_2\\
&\quad= O(1).\end{aligned}$$

Since $\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^{\mathrm{T}} \sim \text{Wishart}(n, \boldsymbol{\Omega}_0)$, it follows from the moment-generating function of the Wishart distribution that (see, e.g., Chapter 8 of [40]) for any $u \in \mathbb{R}$ with $u/n \to 0$ and sufficiently large n,

$$\mathbb{E}_0 \exp\left\{\operatorname{tr}\left(\frac{u}{n}\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S\boldsymbol{\Omega}_0^{-1}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\right)\right\} = \exp\left\{-\frac{n}{2}\log\det\left(\mathbf{I} - \frac{2u}{n}\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S\right)\right\}.$$

Observe that $\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S$ and $\boldsymbol{\Omega}_0^{-1/2}\widetilde{\mathbf{B}}_S\boldsymbol{\Omega}_0^{-1/2}$ are similar matrices having the same set of eigenvalues, that $(2u/n)\lambda_j(\boldsymbol{\Omega}_0^{-1/2}\widetilde{\mathbf{B}}_S\boldsymbol{\Omega}_0^{-1/2}) = o(1)$, and that $\log(1+x) \geq x - x^2$ for sufficiently small $|x|$, we further write

$$\begin{aligned}
&\log\det\left(\mathbf{I} - \frac{2u}{n}\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S\right)\\
&\quad = \sum_{j=1}^p \log\lambda_j\left(\mathbf{I} - \frac{2u}{n}\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S\right) = \sum_{j=1}^p \log\left\{1 - \frac{2u}{n}\lambda_j(\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S)\right\}\\
&\quad = \sum_{j=1}^p \log\left\{1 - \frac{2u}{n}\lambda_j(\boldsymbol{\Omega}_0^{-1/2}\widetilde{\mathbf{B}}_S\boldsymbol{\Omega}_0^{-1/2})\right\}\\
&\quad \geq -\sum_{j=1}^p \frac{2u}{n}\lambda_j(\boldsymbol{\Omega}_0^{-1/2}\widetilde{\mathbf{B}}_S\boldsymbol{\Omega}_0^{-1/2}) - \sum_{j=1}^p \left\{\frac{2u}{n}\lambda_j(\boldsymbol{\Omega}_0^{-1/2}\widetilde{\mathbf{B}}_S\boldsymbol{\Omega}_0^{-1/2})\right\}^2\\
&\quad \geq -\frac{2u}{n}\operatorname{tr}(\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S) - \frac{4u^2}{n^2}\|\boldsymbol{\Omega}_0^{-1}\|_2^2\|\widetilde{\mathbf{B}}_S\|_{\mathrm{F}}^2\\
&\quad \geq -\frac{2u}{n}\operatorname{tr}(\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S) - \frac{Cu^2}{n^2}.
\end{aligned}$$

for some constant $C > 0$. Therefore, with $u/n = o(1)$, for sufficiently large n, we have

$$\begin{aligned}
\mathbb{E}_0\exp(u\rho_S) &= \exp\{-u\operatorname{tr}(\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S)\}\mathbb{E}_0\exp\left\{\frac{u}{n}\operatorname{tr}\left(\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S\boldsymbol{\Omega}_0^{-1}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\right)\right\}\\
&= \exp\{-u\operatorname{tr}(\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S)\}\exp\left\{-\frac{n}{2}\log\det\left(\mathbf{I} - \frac{2u}{n}\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S\right)\right\}\\
&\leq \exp\left\{-u\operatorname{tr}(\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S) + \frac{n}{2}\frac{2u}{n}\operatorname{tr}(\boldsymbol{\Omega}_0^{-1}\widetilde{\mathbf{B}}_S) + \frac{Cu^2}{2n}\right\}\\
&\leq \exp(Cu^2/n).
\end{aligned}$$

Hence, by the Chernoff bound for normal, we obtain

$$\begin{aligned}
&\mathbb{P}_0\left\{|(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}}D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S})| > \frac{M}{8}\sqrt{\frac{s\log p}{n}}\right\}\\
&\quad = \mathbb{P}_0\left(\rho_S > \frac{M}{8}\sqrt{\frac{s\log p}{n}}\right) + \mathbb{P}_0\left(\rho_S < -\frac{M}{8}\sqrt{\frac{s\log p}{n}}\right)
\end{aligned}$$

$$\leq \mathbb{P}_0\left(\rho_S\sqrt{ns\log p} > \frac{Ms\log p}{8}\right) + \mathbb{P}_0\left(-\rho_S\sqrt{ns\log p} > \frac{Ms\log p}{8}\right)$$
$$\leq \frac{\mathbb{E}_0\exp(\rho_S\sqrt{ns\log p})}{\exp\{(Ms\log p)/8\}} + \frac{\mathbb{E}_0\exp(-\rho_S\sqrt{ns\log p})}{\exp\{(Ms\log p)/8\}}$$
$$\leq 2\exp\left\{-\left(\frac{1}{8}M - C\right)s\log p\right\}.$$

Finally, the above bound leads to the following upper bound for the term on line (32):

$$\sum_{S\in\mathcal{S}_0}\sum_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)} \mathbb{P}_0\left\{|(\mathbf{u}\otimes\mathbf{u})^{\mathrm{T}} D\boldsymbol{\Sigma}_S(\boldsymbol{\theta}_{0S})(\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S})| > \frac{M}{8}\sqrt{\frac{s\log p}{n}}\right\}$$
$$\leq 2\sum_{S\in\mathcal{S}_0}\sum_{\mathbf{u}\in\mathcal{S}^{p(S)-1}(1/5)} \exp\left\{-\left(\frac{1}{8}M - C\right)s\log p\right\}$$
$$\leq 2\sum_{t=s_0}^{\kappa_0 s_0}\binom{p-r}{t}|S^{(t+r)-1}(1/5)|\exp\left\{-\left(\frac{1}{8}M - C\right)s\log p\right\}$$
$$\leq 2\sum_{t=s_0}^{\kappa_0 s_0}(\kappa_0 s_0)^{p-r}\exp\left\{c\kappa_0 s_0 - \left(\frac{1}{8}M - C\right)s\log p\right\}$$
$$\leq 2\kappa_0 s_0\exp\left\{(c+1)\kappa_0 s_0\log p - \frac{1}{8}Ms\log p + Cs\log p\right\} \leq 2e^{-(Ms\log p)/16}$$

by taking a sufficiently large $M > 0$ as $n \to \infty$. The proof is thus completed. □

References

[1] Absil, P.-A., Mahony, R. and Sepulchre, R. (2004). Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Appl. Math.* **80** 199–220. MR2035507

[2] Absil, P. A., Mahony, R. and Sepulchre, R. (2009). *Optimization Algorithms on Matrix Manifolds*. Princeton University Press. https://doi.org/doi:10.1515/9781400830244. MR2364186

[3] Amini, A. A. and Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. https://doi.org/10.1214/08-AOS664. MR2541450

[4] Bendokat, T., Zimmermann, R. and Absil, P.-A. (2020). A Grassmann manifold handbook: Basic geometry and computational aspects. *arXiv preprint arXiv:2011.13699*. MR4685952

[5] Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. https://doi.org/10.1214/13-AOS1127. MR3127849

[6] Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media. MR2807761

[7] Cai, T., Ma, Z. and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815. https://doi.org/10.1007/s00440-014-0562-z. MR3334281

[8] Cai, T. T., Ma, Z. and Wu, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. https://doi.org/10.1214/13-AOS1178. MR3161458

[9] Cai, T. T., Ren, Z. and Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Statist.* **10** 1–59. https://doi.org/10.1214/15-EJS1081. MR3466172

[10] Cai, T. T. and Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. https://doi.org/10.1214/12-AOS998. MR3097607

[11] Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. https://doi.org/10.1214/15-AOS1334. MR3375874

[12] Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *The Journal of Abnormal and Social Psychology* **38** 476–506. https://doi.org/10.1037/h0054116.

[13] Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. https://doi.org/10.1137/0707001. MR0264450

[14] Donoho, D., Gavish, M. and Johnstone, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.* **46** 1742–1778. https://doi.org/10.1214/17-AOS1601. MR3819116

[15] Driver, H. E. and Kroeber, A. L. (1932). *Quantitative Expression of Cultural Relationships. University of California Publications: American Archaeology and Ethnology.* University of California Press.

[16] Drmac, Z. (2000). On principal angles between subspaces of Euclidean space. *SIAM J. Matrix Anal. Appl.* **22** 173-194. https://doi.org/10.1137/S0895479897320824. MR1779723

[17] Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. MR3091653

[18] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7** 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[19] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostat.* **9** 432–441. https://doi.org/10.1093/biostatistics/kxm045.

[20] Fruchter, B. (1954). *Introduction to Factor Analysis. Van Nostrand Series in Psychology.* D. Van Nostrand Company.

[21] Gao, C. and Zhou, H. H. (2015). Rate-optimal posterior contraction for sparse PCA. *Ann. Statist.* **43** 785–818. https://doi.org/10.1214/14-AOS1268. MR3325710

[22] Gao, C. and Zhou, H. H. (2016). Bernstein-von Mises theorems for functionals of the covariance matrix. *Electron. J. Statist.* **10** 1751–1806. https://doi.org/10.1214/15-EJS1048. MR3522660

[23] Georghiades, A. S., Belhumeur, P. N. and Kriegman, D. J. (2001). From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.* **23** 643–660. https://doi.org/10.1109/34.927464.

[24] Ghosal, S., Ghosh, J. K. and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. https://doi.org/10.1214/aos/1016218228. MR1790007

[25] Ghosal, S., Van Der Vaart, A. et al. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.* **35** 192–223. https://doi.org/10.1214/009053606000001172. MR2332274

[26] Harman, H. H. (1976). *Modern Factor Analysis*. University of Chicago press. MR0400546

[27] Hoffmann, M., Rousseau, J. and Schmidt-Hieber, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. https://doi.org/10.1214/15-AOS1341. MR3396985

[28] Jauch, M., Hoff, P. D. and Dunson, D. B. (2020). Random orthogonal matrices and the Cayley transform. *Bernoulli* **26** 1560–1586. https://doi.org/10.3150/19-BEJ1176. MR4058378

[29] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. https://doi.org/10.1214/aos/1009210544. MR1863961

[30] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. PMID: 20617121. https://doi.org/10.1198/jasa.2009.0121. MR2751448

[31] Koltchinskii, V. and Lounici, K. (2017). New asymptotic results in principal component analysis. *Sankhya A* **79** 254–297. https://doi.org/10.1007/s13171-017-0106-6. MR3707422

[32] Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. https://doi.org/10.3150/15-BEJ730. MR3556768

[33] Lee, K.-C., Ho, J. and Kriegman, D. J. (2005). Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intell.* **27** 684–698. https://doi.org/10.1109/TPAMI.2005.92.

[34] Lei, J. and Vu, V. Q. (2015). Sparsistency and agnostic inference in sparse PCA. *Ann. Statist.* **43** 299–322. https://doi.org/10.1214/14-AOS1273. MR3311861

[35] Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. https://doi.org/10.1214/13-AOS1097. MR3099121

[36] Magnus, J. R. (1988). Linear structures. *Griffin's Statistical Monographs and Courses* **42**. MR0947343

[37] The Cancer Genome Atlas Network et al. (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489** 519–525.

[38] Ning, B. (2020). Bayesian principal component analysis with sparse priors Technical Report.

[39] Ning, Y. B. and Ning, N. (2024). Spike and slab Bayesian sparse principal component analysis. *Statist. Comput.* **34**. https://doi.org/10.1007/s11222-024-10430-8. MR4745806

[40] Page Jr, T. J. (1984). Multivariate statistics: A vector space approach. *J. Mark. Res.* **21** 236.

[41] Pati, D., Bhattacharya, A., Pillai, N. S. and Dunson, D. (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Ann. Statist.* **42** 1102–1130. https://doi.org/10.1214/14-AOS1215. MR3210997

[42] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Stat. Sin.* **17** 1617–1642. MR2399865

[43] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. MR0965765

[44] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2** 559–572. https://doi.org/10.1080/14786440109462720.

[45] Qiu, L., Zhang, Y. and Li, C.-K. (2005). Unitarily invariant metrics on the Grassmann space. *SIAM J. Matrix Anal. Appl.* **27** 507-531. https://doi.org/10.1137/040607605. MR2179686

[46] Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability* **18** 1–9. https://doi.org/10.1214/ECP.v18-2865. MR3125258

[47] Shepard, R., Brozell, S. R. and Gidofalvi, G. (2015). The representation and parametrization of orthogonal matrices. *The Journal of Physical Chemistry A* **119** 7924–7939. https://doi.org/10.1021/acs.jpca.5b02015.

[48] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267–288. MR1379242

[49] Tryon, R. C. (1939). *Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*. Edwards brother, Incorporated, lithoprinters and publishers.

[50] Van der Vaart, A. W. (2000). *Asymptotic Statistics* **3**. Cambridge University Press. MR1652247

[51] Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint:arXiv:1011.3027*. MR2963170

[52] Vu, V. and Lei, J. (2012). Minimax Rates of Estimation for Sparse PCA in High Dimensions. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* (N. D. Lawrence and M. Girolami, eds.). *Proceedings of Machine Learning Research* **22** 1278–1286.

PMLR, La Palma, Canary Islands.

[53] Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947. https://doi.org/10.1214/13-AOS1151. MR3161452

[54] Wedin, P. Å. (1983). On angles between subspaces of a finite dimensional inner product space. In *Matrix Pencils* (B. Kågström and A. Ruhe, eds.) 263–285. Springer Berlin Heidelberg, Berlin, Heidelberg.

[55] Xie, F. (2023). Euclidean representation of low-rank matrices and its geometric properties. *SIAM J. Matrix Anal. Appl.* **44** 822-866. https://doi.org/10.1137/22M1489125. MR4597265

[56] Xie, F., Cape, J., Priebe, C. E. and Xu, Y. (2022). Bayesian sparse spiked covariance model with a continuous matrix shrinkage prior. *Bayesian Anal.* **17** 1193–1217. https://doi.org/10.1214/21-BA1292. MR4506026

[57] Xu, M. and Duan, L. L. (2023). Bayesian inference with the ℓ_1-ball prior: solving combinatorial problems with exact zeros. *J. R. Stat. Soc. Ser. B Methodol.* accepted for publication. https://doi.org/10.1093/jrsssb/qkad076. MR4718547

[58] Xu, M., Zhou, H., Hu, Y. and Duan, L. L. (2023). Bayesian inference using the proximal mapping: Uncertainty quantification under varying dimensionality. *J. Amer. Statist. Assoc.* in press. https://doi.org/10.1080/01621459.2023.2220170. MR4797906

[59] Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265-286. https://doi.org/10.1198/106186006X113430. MR2252527

[60] Zubin, J. (1938). A technique for measuring like-mindedness. *J. Abnorm. Psychol.* **33** 508–516.