A unified matrix model including both CCA and F matrices in multivariate analysis: the largest eigenvalue and its applications

Xiao Han[*], Guangming Pan[†], and Qing Yang[‡]

Abstract

Let $\mathbf{Z}_{M_1\times N} = \mathbf{T}^{\frac{1}{2}}\mathbf{X}$ where $(\mathbf{T}^{\frac{1}{2}})^2 = \mathbf{T}$ is a positive definite matrix and $\mathbf{X}$ consists of independent random variables with mean zero and variance one. This paper proposes a unified matrix model

$$\mathbf{\Omega} = (\mathbf{Z}\mathbf{U}_2\mathbf{U}_2^T\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{U}_1\mathbf{U}_1^T\mathbf{Z}^T,$$

where $\mathbf{U}_1$ and $\mathbf{U}_2$ are isometric with dimensions $N\times N_1$ and $N\times(N-N_2)$ respectively such that $\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}_{N_1}$, $\mathbf{U}_2^T\mathbf{U}_2 = \mathbf{I}_{N-N_2}$ and $\mathbf{U}_1^T\mathbf{U}_2 = 0$. Moreover, $\mathbf{U}_1$ and $\mathbf{U}_2$ (random or non-random) are independent of $\mathbf{Z}_{M_1\times N}$ and with probability tending to one, $rank(\mathbf{U}_1) = N_1$ and $rank(\mathbf{U}_2) = N - N_2$. We establish the asymptotic Tracy-Widom distribution for its largest eigenvalue under moment assumptions on $\mathbf{X}$ when N_1, N_2 and M_1 are comparable.

The asymptotic distributions of the maximum eigenvalues of the matrices used in Canonical Correlation Analysis (CCA) and of F matrices (including centered and non-centered versions) can be both obtained from that of $\mathbf{\Omega}$ by selecting appropriate matrices $\mathbf{U}_1$ and $\mathbf{U}_2$. Moreover, via appropriate matrices $\mathbf{U}_1$ and $\mathbf{U}_2$, this matrix $\mathbf{\Omega}$ can be applied to some multivariate testing problems that cannot be done by the traditional CCA matrix. To see this, we explore two more applications. One is in the MANOVA approach for testing the equivalence of several high-dimensional mean vectors, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are chosen to be two nonrandom matrices. The other one is in the multivariate linear model for testing the unknown parameter matrix, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are random. For each application, theoretical results are developed and various numerical studies are conducted to confirm the satisfactory empirical performance.

KEY WORDS: Canonical correlation analysis, F matrix, Largest eigenvalue, MANOVA, multivariate linear model, Tracy-Widom distribution, random matrix theory.

[*] Xiao Han, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 637371(Email: xhan011@e.ntu.edu.sg).

[†] Guangming Pan, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 637371(Email: gmpan@ntu.edu.sg). This work was partially supported by a MOE Tier 2 grant 2014-T2-2-060 and by a MOE Tier 1 Grant RG25/14 at the Nanyang Technological University, Singapore.

[‡] Qing Yang, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, 637371(Email: qyang1@e.ntu.edu.sg).

1 Introduction

Rapid development of modern technology nowadays necessitates statistical inference on high-dimensional data in many scientific fields such as image processing, genetic engineering, machine learning and so on. This raises a boom in pursuing methodologies to remedy classical theories which are designed for the fixed dimensions. For such a purpose one popular tool is the spectral analysis of high-dimensional matrices in random matrix theory. The readers may refer to the monograph [2] and the references therein for a comprehensive reading.

This paper focuses on the largest eigenvalues. Ever since the pioneer work discovering the limiting distribution of the largest eigenvalue for the large Gaussian Wigner ensemble by Tracy and Widom in [25, 26], the largest eigenvalues of large random matrices have been widely studied. To name a few we mention [4], [6], [7] and [16]. The largest eigenvalues prove to be fruitful objects of study, playing an important role in multivariate statistical analysis such as principle component analysis (PCA), multivariate analysis of variance (MANOVA), canonical correlation analysis (CCA) and discriminant analysis. Among the vast literature, we refer the readers to a seminal work [14], as well as a recent work [12]. Johnstone in [14] considered a double Wishart setting and developed the Tracy-Widom law of its largest root when the dimension of the data matrix $\mathbf{X}$ and the sample size are comparable with the dimension being even. This limiting distribution can be applied to conduct various statistical inferences in his companion paper [15]. Considering that the results in [14] work for the Gaussian distribution only, the authors in [12] investigated an F type matrix for the general distributions without even dimension restriction. However, one may notice that the Tracy-Widom result in [12] is only verified for zero mean data.

We now set a stage to present our matrix model. The most initial motivation is the matrix frequently used in CCA. Suppose that we are given two sets of random variables, organized into two random vectors $\mathbf{x}$ and $\mathbf{y}$ with dimensions M_1 and M_2, respectively. Without loss of generality, we may assume that $M_1 \le M_2$. In multivariate analysis, CCA is the favorite method to investigate the correlation structure between two random vectors, which was introduced by Hotelling [11] first. The aim of CCA is to seek two vectors $\mathbf{a}$ and $\mathbf{b}$ such that the linear combination of $\mathbf{a}^T\mathbf{x}$ and $\mathbf{b}^T\mathbf{y}$ can get the highest correlation coefficient. i.e.

$$\rho(\mathbf{a},\mathbf{b}) \coloneqq \frac{Cov(\mathbf{a}^T\mathbf{x},\mathbf{b}^T\mathbf{y})}{\sqrt{Var(\mathbf{a}^T\mathbf{x})}\sqrt{Var(\mathbf{b}^T\mathbf{y})}}. \tag{1.1}$$

If $\rho_1 = \rho_1(\mathbf{a}_1,\mathbf{b}_1) \coloneqq \max_{\mathbf{a},\mathbf{b}} \rho(\mathbf{a},\mathbf{b})$, then ρ_1 is called the first canonical correlation coefficient. Given the first canonical correlation coefficient, one can continue to seek the second canonical correlation coefficient which is the maximum correlation coefficient of $\mathbf{a}_2^T\mathbf{x}$ and $\mathbf{b}_2^T\mathbf{y}$, uncorrelated to $\mathbf{a}_1^T\mathbf{x}$ and $\mathbf{b}_1^T\mathbf{y}$. Iterating this procedure to the end, we can get the canonical correlation coefficients ρ_1, ρ_2,...,ρ_{M_1}. Denote the population covariance matrix of any two random vectors $\mathbf{u}$ and $\mathbf{v}$ by $\Sigma_{\mathbf{uv}}$. By (1.1), it is not hard to conclude that in order to find the population canonical correlation coefficients ρ_1, ρ_2,...,ρ_{M_1}, one only need to solve the determinant equation

$$det(\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{xy}}^T - \rho^2\Sigma_{\mathbf{xx}}) = 0. \tag{1.2}$$

If $\mathbf{x}$ and $\mathbf{y}$ are independent, then $\rho_1^2 = \cdots = \rho_{M_1}^2 = 0$ or equivalently the largest eigenvalue of $\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{xy}}^T$, $\rho_1^2 = 0$. For the moment, we assume that $\mathbb{E}\mathbf{x} = \mathbb{E}\mathbf{y} = 0$ for ease of illustration, but bearing in mind that such conditions are not needed in this work. Then under the classical low-dimensional setting, i.e., both M_1 and M_2 are fixed but N is large, one can safely use γ_1, the largest eigenvalue of $\mathbf{A}_{\mathbf{xx}}^{-1}\mathbf{A}_{\mathbf{xy}}\mathbf{A}_{\mathbf{yy}}^{-1}\mathbf{A}_{\mathbf{xy}}^T$, to estimate ρ_1^2 since the sample covariance matrices converge to their population counterparts as N tends to infinity, where

$$\mathbf{A}_{\mathbf{xx}} = \mathbf{X}\mathbf{X}^T, \quad \mathbf{A}_{\mathbf{yy}} = \mathbf{Y}\mathbf{Y}^T, \quad \mathbf{A}_{\mathbf{xy}} = \mathbf{X}\mathbf{Y}^T.$$

However, when M_1 and M_2 are comparable with the sample size N, the consistency will no longer hold for the sample covariance matrices and accordingly the largest sample canonical correlation coefficient γ_1. Putting forward a theory on high-dimensional CCA is then much needed.

If $\mathbf{x}$ or $\mathbf{y}$ is Gaussian distributed, it is not difficult to derive that the largest eigenvalue of $\mathbf{S}_{xy} = \mathbf{A}_{\mathbf{xx}}^{-1}\mathbf{A}_{\mathbf{xy}}\mathbf{A}_{\mathbf{yy}}^{-1}\mathbf{A}_{\mathbf{xy}}^T$ reduces to that of the double Wishart matrices in [14], see the equation (1.3) below. Thus after centralizing and re-scaling, it converges to the Type-1 Tracy-Widom distribution as proved in [14] and [12]. However, to our best knowledge, corresponding results are not yet available for non-gaussian distributions, which is the starting point of this paper. Here we would also remark some other existing work about CCA in the high dimensional case. Central limit theorems of linear spectral statistics of CCA have been established in [28], which is for zero mean data, while spiked eigenvalues are investigated for CCA in [3]. There are also a lot of existing work about sparse CCA and we mention [10] among others.

Denote the largest eigenvalue of $\mathbf{S}_{xy}$ by γ_1. Then γ_1 is also the largest eigenvalue of $\mathbf{T}_{xy} := \mathbf{P}_y\mathbf{P}_x\mathbf{P}_y$, where

$$\mathbf{P}_x = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}, \quad \mathbf{P}_y = \mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}.$$

Equivalently, it is the largest solution to $\det(\mathbf{X}\mathbf{P}_y\mathbf{X}^T - \gamma_1\mathbf{X}\mathbf{X}^T) = 0$. Define $\lambda_1 = \frac{\gamma_1}{1-\gamma_1}$. Then under the condition that $\liminf_{N\to\infty}\frac{N}{M_1+M_2} > 1$, λ_1 is also the largest solution of

$$\det(\mathbf{X}\mathbf{P}_y\mathbf{X}^T - \lambda_1\mathbf{X}(\mathbf{I}-\mathbf{P}_y)\mathbf{X}^T) = 0.$$

The matrix of interest now becomes

$$(\mathbf{X}(I-\mathbf{P}_y)\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{P}_y\mathbf{X}^T. \tag{1.3}$$

Inspired by (1.3), we propose a unified matrix model

$$\mathbf{\Omega} = (\mathbf{Z}\mathbf{U}_2\mathbf{U}_2^T\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{U}_1\mathbf{U}_1^T\mathbf{Z}^T \tag{1.4}$$

where $\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}_{N_1}$, $\mathbf{U}_2^T\mathbf{U}_2 = \mathbf{I}_{N-N_2}$ and $\mathbf{U}_1^T\mathbf{U}_2 = 0$ (see (2.1) below for more details). We establish the asymptotic Tracy-Widom law for its largest eigenvalue in this work. An intriguing observation is that although our Tracy-Widom approximation is theoretically established for diverging dimensions, it keeps accurate for small ones (the dimension M_1 can be as small as 5 in Table 1).

The motivations behind the construction of such a matrix model $\mathbf{\Omega}$ are illustrated as follows. First, the matrix (1.3) used in CCA is a special case of $\mathbf{\Omega}$ by noticing that $\mathbf{P}_y$ and $\mathbf{I} - \mathbf{P}_y$ are orthogonal projection matrices. In addition, the non-zero mean data can be accommodated by writing $\mathbf{U}_2\mathbf{U}_2^T = \mathbf{P}_N(I - \mathbf{P}_{Ny})\mathbf{P}_N, \mathbf{U}_1\mathbf{U}_1^T = \mathbf{P}_N\mathbf{P}_{Ny}\mathbf{P}_N$ and observing that the mean vectors can be absorbed into the matrix $\mathbf{P}_N = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$, see Remark 6 below (the definition of $\mathbf{P}_{Ny}$ is given there). Further illustrations are given in Section 3, where we deal with the independence testing via CCA in detail.

Secondly, by selecting appropriate matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ (**random** or **nonrandom**) the Tracy-Widom distribution for the largest eigenvalue of this unified matrix $\mathbf{\Omega}$ can be applied to the other multivariate testing problems, which cannot be done by the traditional CCA matrix (1.3). To see this, we explore two more applications. One is the MANOVA approach in testing the equivalence of g groups' mean vectors. It is well known that classical MANOVA relies on the eigenvalues of the matrix $\mathbf{V} = \mathbf{W}^{-1}\mathbf{B}$, where $\mathbf{W}$ is the within sum of squares and cross-product matrix (SSCP) and $\mathbf{B}$ is the between SSCP, see [1]. The matrix $\mathbf{V}$ can be written in terms of $\mathbf{\Omega}$ by choosing ***nonrandom*** matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ as in equations (4.2)-(4.3) below, with the derivation details postponed to Section 4. The other one is in the multivariate linear regression model $\mathbf{Y} = \mathbf{XB} + \mathbf{Z}$ for testing the unknown parameter matrix $\mathbf{B}$. We consider both the linear hypothesis testing $H_0 : \mathbf{C}_1\mathbf{B} = \mathbf{\Gamma}_1$ and the general intra-subject hypothesis testing $H_0 : \mathbf{CBD} = \mathbf{\Gamma}$. Taking the linear one as an example, we can rewrite its testing matrix $\mathbf{M}_1 = \mathbf{E}_1^{-1}\mathbf{H}_1$ in the form of $\mathbf{\Omega}$ by selecting ***random*** matrices $\mathbf{U}_1\mathbf{U}_1^T = \mathbf{P}_{\widetilde{\mathbf{X}}}$ and $\mathbf{U}_2\mathbf{U}_2^T = \mathbf{I} - \mathbf{P}_{\mathbf{X}}$ in (5.3), where $\mathbf{E}_1$ is the error SSCP and $\mathbf{H}_1$ the hypothesis SSCP described in Section 5. Simulation results in Sections 6.3-6.4 show that the largest eigenvalue performs well in these two applications for both dense but weak alternative (DWA) and sparse but strong alternative (SSA).

Thirdly, the matrix $\mathbf{\Omega}$ generalizes the models in [14] and [12]. We would like to point out that if the matrix $\mathbf{Z}$ is generated from Gaussian distribution, then the two terms $(\mathbf{Z}\mathbf{U}_2\mathbf{U}_2^T\mathbf{Z}^T)$ and $(\mathbf{Z}\mathbf{U}_1\mathbf{U}_1^T\mathbf{Z}^T)$ in $\mathbf{\Omega}$ are independent with normal entries, which reduces to the one studied in [14]. Without this Gaussian assumption, we indeed investigate a more general case–the two terms can only be considered as uncorrelated with each other. We would also like to highlight that $\mathbf{\Omega}$ not only covers the F-matrix in [12], but also generalize it to any non-zero mean vectors by choosing some special $\mathbf{U}_2$ and $\mathbf{U}_1$. Detailed explanations will be given in Section 2. We remark that all three applications in Sections 3-5 can not be done by either [14] or [12] because we neither assume Gaussian distribution for $\mathbf{Z}$ nor impose independent structure on $(\mathbf{Z}\mathbf{U}_2\mathbf{U}_2^T\mathbf{Z}^T)$ and $(\mathbf{Z}\mathbf{U}_1\mathbf{U}_1^T\mathbf{Z}^T)$.

This paper is organized as follows. In Section 2, the main theorem about the Tracy-Widom distribution for the largest eigenvalue λ_1 of the unified matrix $\mathbf{\Omega}$ is presented. Three applications are introduced in Sections 3, 4 and 5, regarding the high-dimensional independence testing via CCA, MANOVA and multivariate linear regression, respectively. Except for theoretical results developed in previous sections, we also conduct a series of simulations in Section 6 to investigate the accuracy of the proposed asymptotic Tracy-Widom distribution (Section 6.1) as well as its numerical performance in our three applications (Sections 6.2-6.4). We give an outline and some key steps for the proof of Theorem 2.1 in the appendix of Section 7, while all detailed proofs are relegated to the supplementary material.

2 Main result on $\mathbf{\Omega}$

We investigate the largest eigenvalue of the unified matrix

$$\mathbf{\Omega} = (\mathbf{Z}\mathbf{U}_2\mathbf{U}_2^T\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{U}_1\mathbf{U}_1^T\mathbf{Z}^T \tag{2.1}$$

in this section and develop its Tracy-Widom distribution without any specific distribution assumption. Here $\mathbf{Z}_{M_1\times N} = \mathbf{T}^{\frac{1}{2}}\mathbf{X}$, $\mathbf{T}_{M_1\times M_1}$ can be any positive definite matrix and $\mathbf{X} = (X_{ij})_{M_1\times N}$ satisfies the following Condition 1. Assume that $\mathbf{U}_1$ and $\mathbf{U}_2$ are two isometries with dimensions $N\times N_1$ and $N\times(N-N_2)$, respectively such that $N_1 \leq N_2$, $\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}_{N_1}$, $\mathbf{U}_2^T\mathbf{U}_2 = \mathbf{I}_{N-N_2}$ and $\mathbf{U}_1^T\mathbf{U}_2 = 0$. Moreover, $\mathbf{U}_1$ and $\mathbf{U}_2$ (random or non-random) are independent of $\mathbf{X}$ and with probability tending to one, $rank(\mathbf{U}_1) = N_1$ and $rank(\mathbf{U}_2) = N - N_2$. The notation "0" may indicate a zero value, a zero vector or a zero matrix in this paper, changing from line to line.

Condition 1. *A matrix $\mathbf{X} = (X_{ij})_{M_1\times N}$ satisfies Condition 1 if its entries X_{ij} are independent (but not necessarily identically distributed) with all moments being finite and*

$$\mathbb{E}X_{ij} = 0, \quad \mathbb{E}X_{ij}^2 = \mathbb{E}X_{it}^2, \quad 1\leq i\leq M_1, \quad 1\leq j,t\leq N. \tag{2.2}$$

Remark 1. *Note that the matrix $\mathbf{T}$ does not influence the largest eigenvalue of $\mathbf{\Omega}$ and it can be any positive definite matrix. Indeed, let $\mathbf{\Omega}_x = (\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1\mathbf{U}_1^T\mathbf{X}^T$. One can easily observe that $\mathbf{\Omega}$ and $\mathbf{\Omega}_x$ share the same largest eigenvalue by the fact that AB and BA share the same nonzero eigenvalues.*

Before stating the main result we now make some comments about the relation between the matrix model $\mathbf{\Omega}$ and the existing models in the literature. First, as stated in the introduction, if the matrix $\mathbf{Z}$ is generated from Gaussian distribution, then the two terms $(\mathbf{Z}\mathbf{U}_2\mathbf{U}_2^T\mathbf{Z}^T)$ and $(\mathbf{Z}\mathbf{U}_1\mathbf{U}_1^T\mathbf{Z}^T)$ in $\mathbf{\Omega}$ can be considered as independent terms with normal entries, which reduces to the matrix introduced in the seminar work [14]. Secondly, we would like to point out that the matrix $\mathbf{\Omega}$ not only covers the F-matrix model studied in [12], but also generalizes it to the nonzero mean value case. To see this, choose

$$\mathbf{Z} = (\mathbf{Y}_{M_1\times n_1}, \mathbf{W}_{M_1\times n_2}), \quad \mathbf{U}_2 = \begin{pmatrix} 0 \\ \mathcal{P}_2 \end{pmatrix}, \quad \mathbf{U}_1 = \begin{pmatrix} \mathcal{P}_1 \\ 0 \end{pmatrix}$$

with appropriate dimensions, respectively. Let

$$\mathcal{P}_2\mathcal{P}_2^T = \mathbf{I}_{n_2} - \frac{1}{n_2}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^T, \quad \mathcal{P}_1\mathcal{P}_1^T = \mathbf{I}_{n_1} - \frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T,$$

$$\mathcal{P}_2^T\mathcal{P}_2 = \mathbf{U}_2^T\mathbf{U}_2 = \mathbf{I}_{N-N_2}, \quad \mathcal{P}_1^T\mathcal{P}_1 = \mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}_{N_1},$$

where "$\mathbf{1}_{n_i}$" indicates an n_i-dimensional column vector with all entries being one $(i = 1, 2)$. Then

$$\begin{aligned} \mathbf{\Omega} &= (\mathbf{Z}\mathbf{U}_2\mathbf{U}_2^T\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{U}_1\mathbf{U}_1^T\mathbf{Z}^T = (\mathbf{W}\mathcal{P}_2\mathcal{P}_2^T\mathbf{W}^T)^{-1}\mathbf{Y}\mathcal{P}_1\mathcal{P}_1^T\mathbf{Y}^T \\ &= \left[\mathbf{W}(\mathbf{I}_{n_2} - \frac{1}{n_2}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^T)\mathbf{W}^T\right]^{-1}\mathbf{Y}(\mathbf{I}_{n_1} - \frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T)\mathbf{Y}^T. \end{aligned}$$

Noticing that the data matrices $\mathbf{W}$ and $\mathbf{Y}$ are centralized in $\mathbf{\Omega}$, we thus extend the results of F-matrix under the assumption of zero mean values in [12] to the nonzero mean vectors. Finally, by assigning other forms to $\mathbf{U}_1$ and $\mathbf{U}_2$ (either random or non-random), the matrix $\mathbf{\Omega}$ can be used in various applications including centered and non-centered CCA, see Sections 3-5.

We now state the limiting distribution for the largest eigenvalue of the unified matrix $\mathbf{\Omega}$.

Theorem 2.1. *Consider the matrix $\mathbf{\Omega}$ defined in (2.1). Suppose that $\mathbf{T}$ is any positive definite matrix and $\mathbf{X}$ satisfies Condition 1. Suppose that $\liminf_{N\to\infty} \frac{N}{M_1+N_2} > 1$, $N_1 \leq N_2$, $\frac{N_1}{N_2}$ and $\frac{M_1}{N-N_2}$ are both bounded away from 0, and $\frac{N_1}{M_1}$ is bounded away from 0 and ∞. Denote the largest eigenvalue of $\mathbf{\Omega}$ by λ_1. Then there exist μ_N and σ_N such that*

$$\lim_{N\to\infty} P(\sigma_N N_1^{2/3}(\lambda_1 - \mu_N) \leq s) = F_1(s), \tag{2.3}$$

where $F_1(s)$ is the Type-1 Tracy-Widom distribution. Moreover, the mean μ_N and variance σ_N can be decided as follows. Suppose that $c_N \in [0, (1 - \sqrt{\frac{M_1}{N-N_2}})^2]$ satisfies the equation

$$\int_{-\infty}^{+\infty} (\frac{c_N}{\lambda - c_N})^2 dF(\lambda) = \frac{N_1}{M_1}, \tag{2.4}$$

where $F(\lambda)$ is the limit spectral density (LSD) of $(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}$. Then

$$\mu_N = \frac{1}{c_N}(1 + \frac{M_1}{N_1}\int_{-\infty}^{+\infty}(\frac{c_N}{\lambda - c_N})dF(\lambda)) \tag{2.5}$$

and

$$\frac{1}{\sigma_N^3} = \frac{1}{c_N^3}(1 + \frac{M_1}{N_1}\int_{-\infty}^{+\infty}(\frac{c_N}{\lambda - c_N})^3 dF(\lambda)). \tag{2.6}$$

Remark 2. *The LSD of the empirical spectral distribution of* $(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)$ *(equivalent to the sample covariance matrix in the Gaussian case) is the famous Marcenko Pastur distribution. From there one can easily find* $F(\lambda)$*.*

Remark 3. *When* $\mathbf{X}$ *is a complex random matrix, Theorem 2.1 still holds but the Tracy-Widom distribution* $F_1(s)$ *should be replaced by* $F_2(s)$*. One may refer to [26] for the definitions of* $F_i(s), i = 1, 2$*.*

Remark 4. *The condition* $\mathbf{U}_1^T\mathbf{U}_2 = 0$ *imposed on the matrices* $\mathbf{U}_1$ *and* $\mathbf{U}_2$ *can be relaxed to* $\mathbf{U}_1^T\mathbf{U}_2 = (\mathbf{I}_{N_1}, 0)$*. In fact, if* $\mathbf{U}_1^T\mathbf{U}_2 = (\mathbf{I}_{N_1}, 0)$*, then we can write* $\mathbf{U}_2$ *as* $\mathbf{U}_2 = (\mathbf{U}_1, \mathbf{U}_4)$ *such that* $\mathbf{U}_1^T\mathbf{U}_4 = 0$*. This is because if we denote* $\mathbf{U}_2 = (\mathbf{U}_3, \mathbf{U}_4)$*, then the relation* $\mathbf{U}_1^T\mathbf{U}_2 = (\mathbf{I}_{N_1}, 0)$ *suggests that* $\mathbf{U}_1^T\mathbf{U}_3 = \mathbf{I}_{N_1}$*,* $\mathbf{U}_1^T\mathbf{U}_4 = 0$ *. Denoting the i-th columns of* $\mathbf{U}_1$ *and* $\mathbf{U}_3$ *by* $\mathbf{u}_{1i}$ *and* $\mathbf{u}_{3i}$ *respectively, we have* $\mathbf{u}_{1i}^T\mathbf{u}_{3i} = 1$*. By the Cauchy-Schwarz inequality, we see that*

$$1 = \mathbf{u}_{1i}^T\mathbf{u}_{3i} \leq \|\mathbf{u}_{1i}\|\|\mathbf{u}_{3i}\| = 1,$$

which forces $\mathbf{u}_{1i} = \mathbf{u}_{3i}$ *and consequently* $\mathbf{U}_1 = \mathbf{U}_3$*,* $\mathbf{U}_2 = (\mathbf{U}_1, \mathbf{U}_4)$ *with* $\mathbf{U}_1^T\mathbf{U}_4 = 0$*. By the arguments above (1.3), the largest eigenvalue of*

$$(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1\mathbf{U}_1^T\mathbf{X}^T = (\mathbf{X}\mathbf{U}_1\mathbf{U}_1^T\mathbf{X}^T + \mathbf{X}\mathbf{U}_4\mathbf{U}_4^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1\mathbf{U}_1^T\mathbf{X}^T$$

can be transferred to a function of the largest eigenvalue of $(\mathbf{X}\mathbf{U}_4\mathbf{U}_4^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1\mathbf{U}_1^T\mathbf{X}^T$ *so that Theorem 2.1 is applicable. Therefore, one can also work out the asymptotic distribution for the largest eigenvalue of the matrix* $(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1\mathbf{U}_1^T\mathbf{X}^T$ *under the condition* $\mathbf{U}_1^T\mathbf{U}_2 = (\mathbf{I}_{N_1}, 0)$*.*

Remark 5. *Theorem 2.1 can be extended to the joint distribution of the first k largest eigenvalues, i.e.*

$$\begin{aligned} &\lim_{N\to\infty} P(\sigma_N N_1^{2/3}(\lambda_1 - \mu_N) \leq s_1, ..., \sigma_N N_1^{2/3}(\lambda_k - \mu_N) \leq s_k) \\ &= \lim_{N\to\infty} P(N_1^{2/3}(\lambda_1^{GOE} - 2) \leq s_1, ..., N_1^{2/3}(\lambda_k^{GOE} - 2) \leq s_k), \end{aligned} \tag{2.7}$$

where $\lambda_1^{GOE} \geq ...\lambda_k^{GOE}$ *are the first k largest eigenvalues of* $N_1 \times N_1$ *GOE matrix and k is a finite number independent of N. In fact, such an extension can be accomplished by a discussion parallel to Corollary 3.19 of [16] since we show the local behavior of the steitljes transform near the edge (such as Theorem 7.1). Here we omit the proof.*

A pleasant surprise from the simulated results in Section 6.1 is that although our Tracy-Widom approximation is theoretically developed for large dimensions, it keeps accurate for small dimensions regardless of the data distribution, see Table 1 where even for $M_1 = 5$, the estimated quantiles are well matched with theoretical ones.

In the next three sections, we propose three applications of this limiting Tracy-Widom distribution for λ_1. The first one is our motivation of studying $\mathbf{\Omega}$ as stated in the introduction, the high-dimensional independence testing by using canonical correlation analysis. The second one is the MANOVA approach in testing the equivalence of g groups' mean vectors. And the last one is the unknown parameter matrix testing in the multivariate linear model.

3 Unified matrix in CCA

Suppose that we have two sets of random variables, organized into two random vectors $\mathbf{z} = (z_1, \cdots, z_{M_1})^T$ and $\mathbf{y} = (y_1, \cdots, y_{M_2})^T$, with mean vectors and covariance matrices $(\boldsymbol{\mu}_z, \Sigma_{\mathbf{zz}})$ and $(\boldsymbol{\mu}_y, \Sigma_{\mathbf{yy}})$, respectively. For each of them, N observations are measured and the data matrices are denoted as $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_N)_{M_1 \times N}$ and $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)_{M_2 \times N}$. We want to test

$$H_0 : \quad \mathbf{z} \text{ and } \mathbf{y} \text{ are independent.} \tag{3.1}$$

As illustrated in the introduction, if $\mathbf{z}$ and $\mathbf{y}$ are independent, the largest eigenvalue ρ_1^2 of the matrix $\Sigma_{\mathbf{zz}}^{-1}\Sigma_{\mathbf{zy}}\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{zy}}^T$ should be zero. The corresponding sample version is

$$\begin{aligned}
\mathbf{S}_{zy} &= \left(\sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T\right)^{-1} \left(\sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{y}_i - \bar{\mathbf{y}})^T\right) && (3.2)\\
&\times \left(\sum_{i=1}^N (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T\right)^{-1} \left(\sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{y}_i - \bar{\mathbf{y}})^T\right)^T \\
&= (\mathbf{Z}\mathbf{P}_N\mathbf{Z}^T)^{-1}(\mathbf{Z}\mathbf{P}_N\mathbf{Y}^T)(\mathbf{Y}\mathbf{P}_N\mathbf{Y}^T)^{-1}(\mathbf{Z}\mathbf{P}_N\mathbf{Y}^T)^T, && (3.3)
\end{aligned}$$

where $\mathbf{P}_N = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$ and $\mathbf{1}_N$ indicates an N-dimensional column vector with all entries being one. Denote the largest eigenvalue of $\mathbf{S}_{zy}$ by $\gamma_1^{\mathbf{S}}$ and let $\lambda_1^{\mathbf{S}} = \frac{\gamma_1^{\mathbf{S}}}{1-\gamma_1^{\mathbf{S}}}$. Note that $\mathbf{P}_N$ is a projection matrix. Then the property of $\lambda_1^{\mathbf{S}}$ is a special case of λ_1 in Theorem 2.1 by observing that we can equivalently consider $\lambda_1^{\mathbf{S}}$ as the largest eigenvalue of the matrix

$$(\mathbf{Z}\mathbf{P}_N(I - \mathbf{P}_{Ny})\mathbf{P}_N\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{P}_N\mathbf{P}_{Ny}\mathbf{P}_N\mathbf{Z}^T,$$

where $\mathbf{P}_{Ny} = (\mathbf{Y}\mathbf{P}_N)^T(\mathbf{Y}\mathbf{P}_N\mathbf{Y}^T)^{-1}(\mathbf{Y}\mathbf{P}_N)$. This equivalence has been specified in the introduction, see the derivation of (1.3). It is easy to check that $(\mathbf{P}_N(I - \mathbf{P}_{Ny})\mathbf{P}_N)(\mathbf{P}_N\mathbf{P}_{Ny}\mathbf{P}_N) = 0$. Since both $\mathbf{P}_N(I - \mathbf{P}_{Ny})\mathbf{P}_N$ and $\mathbf{P}_N\mathbf{P}_{Ny}\mathbf{P}_N$ are projection matrices such that $rank(\mathbf{P}_N(I - \mathbf{P}_{Ny})\mathbf{P}_N) = N - M_2$ and $rank(\mathbf{P}_N\mathbf{P}_{Ny}\mathbf{P}_N) = M_2$ with high probability by Lemma 2 in the supplement, we can take $N_1 = N_2 = M_2$ in Theorem 2.1 to obtain the following Corollary 1.

Corollary 1. *Suppose that the data matrix $\mathbf{Z}$ can be written as $\mathbf{Z} = \mathbf{T}^{\frac{1}{2}}\mathbf{X} + \boldsymbol{\mu}_z\mathbf{1}_N^T$ for some positive definite matrix $\mathbf{T}$ and the matrix $\mathbf{X}_{M_1 \times N}$ satisfies Condition 1. We do not impose any condition on the random vector $\mathbf{y}$. Here $\boldsymbol{\mu}_z$ is the mean vector of $\mathbf{z}$ and can be any possible value. Assume that $\liminf\limits_{N\to\infty} \frac{N}{M_1+M_2} > 1$, $\frac{M_1}{N-M_2}$ is bounded away from 0 and $\frac{M_2}{M_1}$ is bounded away from 0 and ∞. Denote the largest eigenvalue of $\mathbf{S}_{zy}$ by $\gamma_1^{\mathbf{S}}$ and let $\lambda_1^{\mathbf{S}} = \frac{\gamma_1^{\mathbf{S}}}{1-\gamma_1^{\mathbf{S}}}$. Then under the null hypothesis (3.1), there* *exist μ_N and σ_N such that*

$$\lim_{N\to\infty} P(\sigma_N M_2^{2/3}(\lambda_1^{\mathbf{S}} - \mu_N) \le s) = F_1(s),$$

where $F_1(s)$ is the Type-1 Tracy-Widom distribution. Denote the LSD of $(\mathbf{X}\mathbf{P}_N(I - \mathbf{P}_{Ny})\mathbf{P}_N\mathbf{X}^T)^{-1}$ by $F(\lambda)$ and suppose that $c_N \in [0, (1-\sqrt{\frac{M_1}{N-N_1}})^2)$. Then the mean μ_N and the variance σ_N can be decided in the same way as in Theorem 1 by replacing N_1 and N_2 with M_2.

According to Corollary 1, we suggest to use $\lambda_1^{\mathbf{S}}$ for the hypothesis testing (3.1) by comparing the rescaled $\lambda_1^{\mathbf{S}}$ value with the theoretical critical point obtained from the Type-1 Tracy-Widom distribution. One can also refer to the numerical studies in Section 6.2.

Remark 6. *One may notice that there is an additional term $\boldsymbol{\mu}_z\mathbf{1}_N^T$ in the expression of $\mathbf{Z}$ in Corollary 1 compared with the one in Theorem 2.1. This allows the mean vectors to be any possible values. We would like to point out that this mean vector does not influence the analysis of $\lambda_1^{\mathbf{S}}$ due to the observation that $\boldsymbol{\mu}_z\mathbf{1}_N^T\mathbf{P}_N = 0$.*

Remark 7. *For the Tracy-Widom distribution in Corollary 1, a similar result can be concluded if we assume that the data matrix $\mathbf{Y} = \mathbf{T}^{\frac{1}{2}}\mathbf{X} + \boldsymbol{\mu}_y\mathbf{1}_N^T$ for some positive definite matrix $\mathbf{T}$ instead and $\boldsymbol{\mu}_y$ is the mean vector of $\mathbf{y}$. In this case, no condition is imposed on the random vector $\mathbf{z}$. And we only need to exchange the roles of M_1 and M_2 in the conclusions of Corollary 1. This is easy to see according to the fact that the largest eigenvalue of $\mathbf{S}_{zy}$ does not change if the roles of $\mathbf{Z}$ and $\mathbf{Y}$ are exchanged in (3.2).*

Remark 8. *For the case $N < M_1 + M_2$, it is trivial that $\gamma_1^{\mathbf{S}} \equiv 1$ and $\lambda_1^{\mathbf{S}} = +\infty$.*

4 Unified matrix in multivariate analysis of variance (MANOVA)

Suppose that we have g populations. Let n_i samples $(\mathbf{x}_{i1}, \cdots, \mathbf{x}_{in_i})$ be available from the ith population with mean vector $\boldsymbol{\mu}_i$ (p-dimensional) and common covariance matrix Σ $(i = 1, \cdots, g)$. The total sample size is denoted by $n = \sum_{i=1}^g n_i$. One frequently discussed problem in multivariate analysis is to investigate whether the g groups have the same mean vector. i.e.

$$H_0 : \quad \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_g. \tag{4.1}$$

The MANOVA approach is well-known for this testing problem. Two main SSCPs, the between SSCP $\mathbf{B}$ and the within SSCP $\mathbf{W}$ are constructed as

$$\mathbf{B} = \sum_{i=1}^g n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T, \quad \mathbf{W} = \sum_{i=1}^g \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T,$$

where $\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum\limits_{j=1}^{n_i} \mathbf{x}_{ij}$ is the i-th group sample mean and $\bar{\mathbf{x}} = \frac{1}{n}\sum\limits_{i=1}^{g}\sum\limits_{j=1}^{n_i} \mathbf{x}_{ij} = \sum\limits_{i=1}^{g} \frac{n_i}{n}\bar{\mathbf{x}}_i$ is the overall sample mean. The classical testing methods for (4.1) are based on the eigenvalues of the matrix

$\mathbf{V} = \mathbf{W}^{-1}\mathbf{B}$. We can show that under the null hypothesis (4.1), the matrix $\mathbf{V}$ can be written as a special form of $\mathbf{\Omega}$ in Section 2 and thus the limiting distribution of its largest eigenvalue $\lambda_1^{\mathbf{V}}$ follows from Theorem 1.

To see this, denote $\mathbf{X}_i = (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{in_i})^T$ (of size $n_i \times p$). Note that under the null hypothesis (4.1), the common mean vector does not influence the matrix $\mathbf{V}$. Then without loss of generality, we can simply assume that $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_g = 0$ under H_0. In this section, we use i to denote the ith group ($i = 1, \cdots, g$) and use j to denote the jth observation from the ith group ($j = 1, \cdots, n_i$). For each $\mathbf{X}_i$, let $\mathbf{H}_i$ be an $n_i \times n_i$ orthogonal matrix with the first column being $\frac{1}{\sqrt{n_i}}\mathbf{1}_{n_i}$. Here $\mathbf{1}_{n_i}$ indicates an n_i-dimensional column vector with all entries being one. The matrix $\mathbf{I}_{n_i}$ indicates an $n_i \times n_i$ identity matrix, $\mathbf{U}_{i1}$ indicates the first column of $\mathbf{I}_{n_i}$ and $\mathbf{U}_{i2}$ indicates the remaining $n_i \times (n_i - 1)$ block of $\mathbf{I}_{n_i}$. An intuitive example for easy understanding when $n_1 = 3$ is

$$\mathbf{I}_{n_1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{U}_{11} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad \mathbf{U}_{12} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Arrange these $\mathbf{U}_{i1}$'s as blocks placed on the diagonal of a block matrix $\mathbf{U}_1$ and $\mathbf{U}_{i2}$'s as blocks placed on the diagonal of another block matrix $\mathbf{U}_2$, i.e.

$$\mathbf{U}_1 = \begin{pmatrix} \underset{(n_1\times 1)}{\mathbf{U}_{11}} & & & \\ & \underset{(n_2\times 1)}{\mathbf{U}_{21}} & & \\ & & \ddots & \\ & & & \underset{(n_g\times 1)}{\mathbf{U}_{g1}} \end{pmatrix}_{n\times g}, \quad \mathbf{U}_2 = \begin{pmatrix} \underset{(n_1\times(n_1-1))}{\mathbf{U}_{12}} & & & \\ & \underset{(n_2\times(n_2-1))}{\mathbf{U}_{22}} & & \\ & & \ddots & \\ & & & \underset{(n_g\times(n_g-1))}{\mathbf{U}_{g2}} \end{pmatrix}_{n\times(n-g)}. \tag{4.2}$$

Consider the orthogonal transformations $\mathbf{Z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2}, \cdots, \mathbf{z}_{in_i})^T = \mathbf{H}_i^T\mathbf{X}_i$ (of size $n_i \times p$). It is easy to find that $\mathbf{z}_{i1} = \sqrt{n_i}\bar{\mathbf{x}}_i$. Furthermore, denote $\mathbf{a}_g = (\sqrt{\frac{n_1}{n}}, \cdots, \sqrt{\frac{n_g}{n}})^T$, $\mathbf{P}_g = \mathbf{I}_g - \mathbf{a}_g\mathbf{a}_g^T$ and $\mathbf{Z} = (\mathbf{Z}_1^T, \mathbf{Z}_2^T, \cdots, \mathbf{Z}_g^T)_{p\times n}$. Considering the relationship $\sqrt{n}\bar{\mathbf{x}} = (\mathbf{z}_{11}, \cdots, \mathbf{z}_{g1})\mathbf{a}_g$, we can obtain

$$\begin{aligned} \mathbf{B} &= \sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T = \sum_{i=1}^{g} n_i\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T - \sqrt{n}\bar{\mathbf{x}} \cdot \sqrt{n}\bar{\mathbf{x}} \\ &= (\mathbf{z}_{11}, \cdots, \mathbf{z}_{g1})(\mathbf{I}_g - \mathbf{a}_g\mathbf{a}_g^T)(\mathbf{z}_{11}, \cdots, \mathbf{z}_{g1})^T = \mathbf{Z}\mathbf{U}_1\mathbf{P}_g\mathbf{U}_1^T\mathbf{Z}^T = \mathbf{Z}\widetilde{\mathbf{U}}_1\widetilde{\mathbf{U}}_1^T\mathbf{Z}^T, \\ \mathbf{W} &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T = \sum_{i=1}^{g}(\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{x}_{ij}^T - n_i\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T) = \sum_{i=1}^{g}(\mathbf{X}_i^T\mathbf{X}_i - \mathbf{z}_{i1}\mathbf{z}_{i1}^T) \\ &= \sum_{j=2}^{n_1}\mathbf{z}_{1j}\mathbf{z}_{1j}^T + \sum_{j=2}^{n_2}\mathbf{z}_{2j}\mathbf{z}_{2j}^T + \cdots + \sum_{j=2}^{n_g}\mathbf{z}_{gj}\mathbf{z}_{gj}^T = \mathbf{Z}\mathbf{U}_2\mathbf{U}_2^T\mathbf{Z}^T, \end{aligned} \tag{4.3}$$

where $\widetilde{\mathbf{U}}_1 = \mathbf{U}_1\mathbf{P}_g$ and $\mathbb{E}(\mathbf{Z}) = 0$ under H_0. According to the construction of $\mathbf{U}_1$ and $\mathbf{U}_2$ in (4.2), we can easily conclude that $\widetilde{\mathbf{U}}_1^T\mathbf{U}_2 = 0$. Then the limiting distribution of the largest eigenvalue $\lambda_1^{\mathbf{V}}$ of

$$\mathbf{V} = \mathbf{W}^{-1}\mathbf{B} = (\mathbf{Z}\mathbf{U}_2\mathbf{U}_2^T\mathbf{Z}^T)^{-1}\mathbf{Z}\widetilde{\mathbf{U}}_1\widetilde{\mathbf{U}}_1^T\mathbf{Z}^T$$

can follow from Theorem 2.1 by assigning $M_1 = p$, $N_1 = g-1$ and $N_2 = g$ since $rank(\widetilde{\mathbf{U}}_1) = g-1, rank(\mathbf{U}_2) = n-g$. See the following Corollary 2.

Corollary 2. *Consider the multivariate mean vectors' hypothesis testing problem in (4.1). We use the largest eigenvalue $\lambda_1^{\mathbf{V}}$ of the matrix $\mathbf{V} = \mathbf{W}^{-1}\mathbf{B}$ as the test criterion. Under the null hypothesis, suppose that $\mathbf{Z}$ can be written as $\mathbf{Z} = \mathbf{T}^{\frac{1}{2}}\mathbf{X}$ for some positive definite matrix $\mathbf{T}_{p\times p}$ and the matrix $\mathbf{X}_{p\times n}$ satisfies Condition 1. Assume that $\liminf_{n\to\infty} \frac{n}{p+g} > 1$ and $\frac{g-1}{p}$ is bounded away from 0 and ∞. Then there exist μ_n and σ_n such that*

$$\lim_{n\to\infty} P(\sigma_n(g-1)^{2/3}(\lambda_1^{\mathbf{V}} - \mu_n) \leq s) = F_1(s),$$

where $F_1(s)$ is the Type-1 Tracy-Widom distribution. The mean μ_n and the variance σ_n can be decided in the same way as in Theorem 2.1 by replacing M_1 with p and N_1 with $(g-1)$.

According to Corollary 2, if the rescaled $\lambda_1^{\mathbf{V}}$ value is smaller than the theoretical critical point obtained from Type-1 Tracy-Widom distribution, we fail to reject the null hypothesis (4.1), i.e. we do not reject that the g groups share the same mean vector. Otherwise, reject H_0. In the simulation studies of Section 6.3, regarding the pattern of different mean vectors under the alternative, we consider two cases. One is the dense but weak alternative (DWA), which means that there are many different entries among the mean vectors, but these differences are faint, see the setting $H_1^{(1)}$ and $H_1^{(1)'}$ in Section 6.3 **(1)**. The other one is the sparse but strong alternative (SSA), which means that the differences are rare, but significant where they appear, see the alternative $H_1^{(2)}$, where the differences only appear in one out of p components. The numerical results in Table 3 indicate that this $\lambda_1^{\mathbf{V}}$ shows satisfactory performance for both alternatives.

Remark 9. *If we assume that all the observations come from multivariate normal distribution as in the classical setting, then the positive definite matrix $\mathbf{T}$ in Corollary 2 obviously exists by choosing $\mathbf{T} = \Sigma$. This is due to the fact that we can write each $\mathbf{X}_i$ as $\mathbf{X}_i^T = \Sigma^{\frac{1}{2}}\widetilde{\mathbf{X}}_i = \mathbf{T}^{\frac{1}{2}}\widetilde{\mathbf{X}}_i$ and the entries of $\widetilde{\mathbf{X}}_i$ are i.i.d $N(0,1)$. Then*

$$\mathbf{Z} = (\mathbf{Z}_1^T, \mathbf{Z}_2^T, \cdots, \mathbf{Z}_g^T) = (\mathbf{X}_1^T\mathbf{H}_1, \mathbf{X}_2^T\mathbf{H}_2, \cdots, \mathbf{X}_g^T\mathbf{H}_g) = \mathbf{T}^{\frac{1}{2}}(\widetilde{\mathbf{X}}_1\mathbf{H}_1, \widetilde{\mathbf{X}}_2\mathbf{H}_2, \cdots, \widetilde{\mathbf{X}}_g\mathbf{H}_g) := \mathbf{T}^{\frac{1}{2}}\mathbf{X},$$

where $\mathbf{X} = (\widetilde{\mathbf{X}}_1\mathbf{H}_1, \widetilde{\mathbf{X}}_2\mathbf{H}_2, \cdots, \widetilde{\mathbf{X}}_g\mathbf{H}_g)_{p\times n}$ satisfies Condition 1, taking into account the orthogonality of each $\mathbf{H}_i$ and the independence among each $\widetilde{\mathbf{X}}_i$.

5 Unified matrix in high-dimensional multivariate linear model

In this section, we investigate one more application of the unified matrix $\mathbf{\Omega}$ in the multivariate linear model. Let us consider a linear relationship between p_2 response variables $y_1, \cdots, y_{p_2}$ and p_1 explanatory variables $x_1, \cdots, x_{p_1}$. Suppose that there are N observations, organized into two data matrices:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1^T \\ \vdots \\ \mathbf{Y}_N^T \end{pmatrix}_{N\times p_2}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_N^T \end{pmatrix}_{N\times p_1}.$$

Then the multivariate linear model assumes that

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{Z}, \tag{5.1}$$

where $\mathbf{B}$ is a $p_1 \times p_2$ unknown parameter matrix and $\mathbf{Z}$ is a $N \times p_2$ error matrix with the assumption that the rows of $\mathbf{Z}$ are independent having mean zero and common covariance matrix Σ. We first consider the linear hypothesis testing of the form

$$H_0: \quad \mathbf{C}_1\mathbf{B} = \mathbf{\Gamma}_1, \tag{5.2}$$

where $\mathbf{C}_1$ is a $g_1 \times p_1$ known matrix of rank g_1 and $\mathbf{\Gamma}_1$ is a $g_1 \times p_2$ known matrix of rank $\min\{g_1, p_2\}$. As an example, in the simulation studies of Section 6.4, if we select $\mathbf{C}_1 = \mathbf{C}_1^{(b)} = [\mathbf{I}_{g_1}, 0]$ and $\mathbf{\Gamma}_1 = \mathbf{\Gamma}_1^{(a)} = 0$, then the testing problem (5.2) reduces to analyzing whether the first g_1 rows of $\mathbf{B}$ equal to zeros.

The initial step in conducting the linear hypothesis testing (5.2) is to estimate the unknown parameter matrix $\mathbf{B}$. As stated in Section 2, our proposed Tracy-Widom distribution performs well when the dimensions are small so that we can simply apply the classic least square estimator for $\mathbf{B}$, which is well-known to be $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. The hypothesis SSCP for testing (5.2) is given by $\mathbf{H}_1 = (\mathbf{C}_1\hat{\mathbf{B}} - \mathbf{\Gamma}_1)^T[\mathbf{C}_1(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}_1^T]^{-1}(\mathbf{C}_1\hat{\mathbf{B}} - \mathbf{\Gamma}_1)$ and the error SSCP is $\mathbf{E}_1 = \mathbf{Y}^T[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{Y}$. One can refer to chapter 7 of [9] for detailed derivations. Under the null hypothesis (5.2), $\mathbf{H}_1$ and $\mathbf{E}_1$ can be further rewritten as

$$\begin{aligned} \mathbf{H}_1 &= [\mathbf{C}_1(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}]^T[\mathbf{C}_1(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}_1^T]^{-1}[\mathbf{C}_1(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}] = \mathbf{Z}^T\mathbf{P}_{\widetilde{\mathbf{X}}}\mathbf{Z}, \\ \mathbf{E}_1 &= (\mathbf{X}\mathbf{B} + \mathbf{Z})^T[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T](\mathbf{X}\mathbf{B} + \mathbf{Z}) = \mathbf{Z}^T[\mathbf{I} - \mathbf{P}_{\mathbf{X}}]\mathbf{Z}, \end{aligned} \tag{5.3}$$

where $\widetilde{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}_1^T$, $\mathbf{P}_{\widetilde{\mathbf{X}}} = \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T$ and $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. It is easy to check that $\mathbf{P}_{\widetilde{\mathbf{X}}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = 0$. Denote the largest eigenvalue of

$$\mathbf{M}_1 = \mathbf{E}_1^{-1}\mathbf{H}_1 = (\mathbf{Z}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{P}_{\widetilde{\mathbf{X}}}\mathbf{Z} \tag{5.4}$$

by $\lambda_1^{\mathbf{M}_1}$. As stated in Section 3, both $\mathbf{I} - \mathbf{P}_{\mathbf{X}}$ and $\mathbf{P}_{\widetilde{\mathbf{X}}}$ are projection matrices with $rank(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) = N - p_1$ and $rank(\mathbf{P}_{\widetilde{\mathbf{X}}}) = g_1$ with high probability. Assuming $N_2 = p_1$, $N_1 = g_1$ and $M_1 = p_2$ in Theorem 2.1, we can develop the following corollary for $\lambda_1^{\mathbf{M}_1}$.

Corollary 3. *Assume that $\mathbf{Z}$ in the multivariate linear model (5.1) can be written as $\mathbf{Z} = \mathbf{W}\mathbf{T}^{\frac{1}{2}}$ for some positive definite matrix $\mathbf{T}_{p_2\times p_2}$ and the matrix $\mathbf{W}_{N\times p_2}$ satisfies Condition 1. Suppose that*

$\liminf\limits_{N\to\infty}\frac{N}{p_2+p_1}>1$, $\frac{g_1}{p_1}$ *and* $\frac{p_2}{N-p_1}$ *are both bounded away from* 0 *and* $\frac{g_1}{p_2}$ *is bounded away from* 0 *and* ∞*. Denote the largest eigenvalue of* $\mathbf{M}_1=\mathbf{E}_1^{-1}\mathbf{H}_1$ *by* $\lambda_1^{\mathbf{M}_1}$*. Then under the null hypothesis (5.2), there exist* μ_N *and* σ_N *such that*

$$\lim_{N\to\infty}P(\sigma_N g_1^{2/3}(\lambda_1^{\mathbf{M}_1}-\mu_N)\le s)=F_1(s),$$

where $F_1(s)$ *is the Type-1 Tracy-Widom distribution. Denote the LSD of* $(\mathbf{W}^T(\mathbf{I}-\mathbf{P}_\mathbf{X})\mathbf{W})^{-1}$ *by* $F(\lambda)$ *and suppose that* $c_N\in[0,(1-\sqrt{\frac{p_2}{N-g_1}})^2]$*. Then the mean* μ_N *and the variance* σ_N *can be decided in the same way as in Theorem 2.1 by replacing* N_2 *with* p_1*,* N_1 *with* g_1 *and* M_1 *with* p_2*.*

Remark 10. *One should notice that* $\mathbf{Z}$ *in this Corollary and Corollary 4 corresponds to* $\mathbf{Z}^T$ *in Theorem 2.1. To see this, one may compare (5.4) with (2.1).*

By Corollary 3, we can use $\lambda_1^{\mathbf{M}_1}$ for the linear hypothesis testing (5.2) and reject H_0 if the rescaled $\lambda_1^{\mathbf{M}_1}$ is larger than the theoretical critical point obtained from Type-1 Tracy-Widom distribution. In Section 6.4, we consider the special testing of whether a certain part of $\mathbf{B}$, say $\mathbf{B}_2$, equals a zero matrix. And as in MANOVA, with regard to the pattern under the alternative, both DWA and SSA are applied, i.e. when many entries of $\mathbf{B}_2$ are nonzero but the values are small, see the third combination $(\mathbf{C}_1^{(a)},\mathbf{B}_2^{(d)},\mathbf{\Gamma}_1^{(a)})$, as well as when only two entries of $\mathbf{B}_2$ are nonzero but the values are significant, see the last combination $(\mathbf{C}_1^{(a)},\mathbf{B}_2^{(s)},\mathbf{\Gamma}_1^{(a)})$. The numerical results in Table 4 show that $\lambda_1^{\mathbf{M}_1}$ performs well under both alternatives.

We next consider the intra-subject hypothesis testing of the form

$$H_0:\quad \mathbf{CBD}=\mathbf{\Gamma},\tag{5.5}$$

where $\mathbf{C}$ is a $g_1\times p_1$ known matrix of rank g_1, $\mathbf{D}$ is a $p_2\times g_2$ known matrix of rank g_2 and $\mathbf{\Gamma}$ is a $g_1\times g_2$ known matrix of rank $\min\{g_1,g_2\}$. The hypothesis and error SSCPs for (5.5) can be obtained from $\mathbf{H}_1$ and $\mathbf{E}_1$ by modifying the multivariate linear model (5.1) to the following expression

$$\mathbf{YD}=\mathbf{XBD}+\mathbf{ZD}.$$

Replacing $\mathbf{Y},\mathbf{B}$ and $\mathbf{Z}$ by $\mathbf{YD},\mathbf{BD}$ and $\mathbf{ZD}$ respectively, we can then conclude that the SSCPs for conducting the hypothesis testing (5.5) are

$$\mathbf{H}=(\mathbf{ZD})^T\mathbf{P}_{\widetilde{\mathbf{X}}}(\mathbf{ZD}),\quad \mathbf{E}=(\mathbf{ZD})^T[\mathbf{I}-\mathbf{P}_\mathbf{X}](\mathbf{ZD}),\tag{5.6}$$

where $\widetilde{\mathbf{X}}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T$, $\mathbf{P}_{\widetilde{\mathbf{X}}}=\widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T$ and $\mathbf{P}_\mathbf{X}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. It is easy to check that $\mathbf{P}_{\widetilde{\mathbf{X}}}(\mathbf{I}-\mathbf{P}_\mathbf{X})=0$. Denote the largest eigenvalue of $\mathbf{M}=\mathbf{E}^{-1}\mathbf{H}$ by $\lambda_1^{\mathbf{M}}$. The only difference between the analysis of $\lambda_1^{\mathbf{M}_1}$ and $\lambda_1^{\mathbf{M}}$ is that $\mathbf{Z}_{N\times p_2}$ in (5.3) is replaced by $(\mathbf{ZD})_{N\times g_2}$ in (5.6). So assigning $p_2=g_2$ in Corollary 3, we can obviously obtain the following conclusion for $\lambda_1^{\mathbf{M}}$.

Corollary 4. *For the known matrix* $\mathbf{D}$ *and the error matrix* $\mathbf{Z}$ *in the multivariate linear model (5.1), assume that* $\mathbf{ZD}$ *can be written as* $\mathbf{ZD} = \mathbf{W}\mathbf{T}^{\frac{1}{2}}$ *for some positive definite matrix* $\mathbf{T}_{g_2\times g_2}$ *and the matrix* $\mathbf{W}_{N\times g_2}$ *satisfies Condition 1. Suppose that* $\liminf_{N\to\infty} \frac{N}{g_2+p_1} > 1$, $\frac{g_1}{p_1}$ *and* $\frac{g_2}{N-p_1}$ *are both bounded away from* 0 *and* $\frac{g_1}{g_2}$ *is bounded away from* 0 *and* ∞*. Denote the largest eigenvalue of* $\mathbf{M} = \mathbf{E}^{-1}\mathbf{H}$ *by* $\lambda_1^{\mathbf{M}}$*. Then under the null hypothesis (5.2), there exist* μ_N *and* σ_N *such that*

$$\lim_{N\to\infty} P(\sigma_N g_1^{2/3}(\lambda_1^{\mathbf{M}} - \mu_N) \leq s) = F_1(s),$$

where $F_1(s)$ *is the Type-1 Tracy-Widom distribution. The mean* μ_N *and the variance* σ_N *can be decided in the same way as in Corollary 3 by replacing* p_2 *with* g_2*.*

6 Numerical studies

This section is to investigate the accuracy of our proposed asymptotic Tracy-Widom distribution (Section 6.1) as well as its numerical performance in various applications (Sections 6.2-6.4). Before proceeding to the simulation results, we first introduce an asymptotic substitution of the limiting distribution for the largest eigenvalue in Theorem 2.1. The formulae for calculating μ_N and σ_N in (2.4)-(2.6) are difficult to work with. Referring to [14] and [12], we facilitate the computation by using an approximation in terms of the log transform of λ_1 in Theorem 2.1 as

$$\lim_{N\to\infty} P(\frac{\ln\lambda_1 - \widetilde{\mu}}{\widetilde{\sigma}} \leq s) = F_1(s), \tag{6.1}$$

where $F_1(s)$ still indicates the Type-1 Tracy-Widom distribution and the new mean $\widetilde{\mu}$ and variance $\widetilde{\sigma}$ are defined by

$$\widetilde{\mu} = 2\ln\tan(\frac{\phi+\varphi}{2}), \quad \widetilde{\sigma}^3 = \frac{16}{(N-N_2+N_1-1)^2}\frac{1}{\sin^2(\phi+\varphi)\sin\phi\sin\varphi}.$$

The angle parameters ϕ and φ are defined by

$$\sin^2(\frac{\varphi}{2}) = \frac{\min(M_1,N_1)-1/2}{N-N_2+N_1-1}, \quad \sin^2(\frac{\phi}{2}) = \frac{\max(M_1,N_1)-1/2}{N-N_2+N_1-1}.$$

The asymptotic equivalence between the approximation (6.1) and the one in Theorem 2.1 have been proved in [14] and [12]. All simulations in this section are conducted by adopting this $\ln\lambda_1$'s asymptotic expression. In the sequel, we also use the word "rescaled λ_1" to denote the term $\frac{\ln\lambda_1-\widetilde{\mu}}{\widetilde{\sigma}}$ in (6.1). The values of $\widetilde{\mu}$ and $\widetilde{\sigma}$ in the applications can be obtained simply by replacing N, N_1, N_2, M_1 with their corresponding notations in Sections 3-5. All simulated results below are recorded based on 10000 replications of such a re-scaled largest eigenvalue.

6.1 Approximation accuracy

This subsection is to investigate the Tracy-Widom approximation accuracy for the unified matrix $\mathbf{\Omega}$ in Section 2. Since the positive definite matrix $\mathbf{T}$ does not influence λ_1, we simply let $\mathbf{T} = \mathbf{I}_{M_1}$. Other settings to be used in the simulation are summarized below.

(1). Data distribution: Three data distributions will be used to generate the entries of $\mathbf{X}$ in the model (2.1).

- Data 1: Standard Normal distribution $N(0,1)$.
- Data 2: Discrete distribution with probability mass function $P(x=-\sqrt{3}) = P(x=\sqrt{3}) = 1/6$ and $P(x=0)=2/3$.
- Data 3: Standardized Gamma distribution $Gamma(4, 0.5)$.

The three distributions are used to verify Condition 1, i.e. for the data distribution, we do not need other restrictions except for the first two moments match and all moments are finite. Data 2 supports that the distribution can be a discrete one, while Data 3 is a skewed one with the third and fourth moments different from those of the standard normal distribution.

(2). Dimensions (M_1, N_1, N_2, N)**:** Considering the restrictions on the dimensions, we set two initial choices: $M^{(1)} = (M_1, N_1, N_2, N) = (5, 8, 10, 30)$ and $M^{(2)} = (M_1, N_1, N_2, N) = (15, 8, 10, 50)$, with M_1 being smaller than (N_1, N_2) and larger than (N_1, N_2), respectively. Then we change the magnification factor attached to the initial choices to investigate the performance when the dimensions increase. See the second row of Table 1.

(3). Matrices $\mathbf{U}_1$ and $\mathbf{U}_2$: We randomly generate two matrices $\mathbf{L}_{N\times N_2}$ and $\mathbf{D}_{N_2\times N_1}$ with entries from standard normal distribution. Let $\mathbf{U}_1\mathbf{U}_1^T = (\mathbf{LD})(\mathbf{D}^T\mathbf{L}^T\mathbf{LD})^{-1}(\mathbf{LD})^T$ and $\mathbf{U}_2\mathbf{U}_2^T = \mathbf{I}_N - \mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T$ in the model (2.1). It is easy to check that such settings satisfy the conditions on $\mathbf{U}_1$ and $\mathbf{U}_2$, taking into account the properties of projection matrices.

Simulated results based on above settings are recorded in Table 1. The column titled "Percentile" lists the percentiles of Tracy-Widom distribution corresponding to quantiles in the column "TW". The next ten columns record our estimated cumulative probabilities (i.e. estimated quantiles) for the rescaled λ_1 under various settings stated above, i.e. repeating 10000 times and finding 10000 rescaled λ_1's, then the proportion of values that are less than corresponding percentiles are recorded in Table 1. i.e. $\frac{\#\{\text{rescaled} \quad \lambda_1 \leq \text{“Percentile”}\}}{10000}$. Comparing the empirical results (the last ten columns) with the theoretical ones (the "TW" column), we can see that the rescaled λ_1 matches with the Tracy-Widom law quite well, which supports the accuracy of approximation in Theorem 2.1. Moreover, although our theoretical result is developed for large dimensions, Table 1 indicates that such approximation also works well even when the dimensions are small.

6.2 Performance in the independence testing

This subsection is to investigate the performance of our proposed largest eigenvalue $\lambda_1^{\mathbf{S}}$ in the independence testing of Section 3. For ease of construction, we let $M_1 = M_2$ and consider a series of settings for the two random vectors $\mathbf{z}$ and $\mathbf{y}$ in the following way:

$$\mathbf{z} = \sqrt{1-\tau}\mathbf{x} + \sqrt{\tau}\mathbf{y}, \quad 0 \leq \tau \leq 1,$$

where two $(M_1 \times 1)$ random vectors $\mathbf{x}$ and $\mathbf{y}$ are independent and τ is a parameter determining the level of dependence between $\mathbf{z}$ and $\mathbf{y}$. When $\tau = 0$, $\mathbf{z}$ and $\mathbf{y}$ are independent, which is the null hypothesis (3.1) in Section 3. Otherwise, as $\tau > 0$ becomes larger, the dependence between $\mathbf{z}$ and $\mathbf{y}$ increases.

Considering the conditions on the dimensions, as in Section 6.1, we also set an initial choice for (M_1, M_2, N) as $M^{(0)} = (M_1, M_2, N) = (10, 10, 40)$ and then change the magnification factor to check the influence of dimensionality. The nominal significance level is set to be $\alpha = 0.05$. According to Table 1, the corresponding theoretical quantile value is $c_\alpha = 0.98$. That is to say, we compare the rescaled $\lambda_1^{\mathbf{S}}$ introduced in Section 3 with c_α. If it is smaller than c_α, then the null hypothesis (3.1) is accepted, i.e. $\mathbf{z}$ and $\mathbf{y}$ are independent. Otherwise, we conclude that they are dependent. We use discrete distribution or Gamma distribution, stated in above Section 6.1 **(1)**, to generate N samples for $\mathbf{x}$ and $\mathbf{y}$. Repeating 10000 times, we can find 10000 rescaled $\lambda_1^{\mathbf{S}}$'s and the proportion of values that are larger than c_α are recorded in Table 2. i.e. $\frac{\#\{\text{rescaled} \quad \lambda_1^{\mathbf{S}} > c_\alpha\}}{10000}$. So when $\tau = 0$, the fourth row of Table 2 records the estimated sizes, which are close to 0.05. When τ changes from 0.1 to 0.4, the corresponding rows give the estimated powers. We can observe that as the dependence between $\mathbf{z}$ and $\mathbf{y}$ becomes stronger and as the dimensions become larger, the power values increase. We do not attach the results when $\tau > 0.4$ here because the powers are always around 1. One can also expect such a phenomenon according to the trend in Table 2.

6.3 Performance in MANOVA

This subsection is to investigate the performance of our proposed largest eigenvalue $\lambda_1^{\mathbf{V}}$ in the MANOVA approach of Section 4. The nominal significance level is set to be $\alpha = 0.05$. Consider $g = 3$ groups with mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3$ and common covariance matrix Σ. We select Σ as the covariance matrix of MA(1) model with the parameter $\theta_1 = 0.2$ and use Gamma distribution stated in Section 6.1 **(1)** to generate the data. Other settings that will be used in the simulation are summarized below.

(1). Mean vectors: Let $\boldsymbol{\mu}_1 = 0_p$, a p-dimensional zero vector, $\mathbf{a}_1 = (\tau_1, \cdots, \tau_1)^T$, a p-dimensional vector with all entries being τ_1 and $\mathbf{a}_2 = (\tau_2, 0, \cdots, 0)^T$, a p-dimensional vector with only the first entry having a nonzero value τ_2. Three different settings on the mean vectors are considered.

- H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = 0_p$. This setting corresponds to the null hypothesis (4.1) in Section 4. It is used to check the empirical size performance when the null hypothesis is true.

Both of the following two settings are under the alternative hypothesis, i.e. the three groups do not share the same mean vector.

- $H_1^{(1)}$ and $H_1^{(1)'}$: $\boldsymbol{\mu}_1 = 0_p$, $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + \mathbf{a}_1$ and $\boldsymbol{\mu}_3 = \boldsymbol{\mu}_2 + \mathbf{a}_1$. This setting reflects the dense but weak alternative (DWA), which means that there are many different entries, but these differences are faint. We choose $\tau_1 = 0.2$ for $H_1^{(1)}$ and a larger $\tau_1 = 0.5$ for $H_1^{(1)'}$. The magnitude of the difference vector $\mathbf{a}_1$ is $\|\mathbf{a}_1\|^2 = \tau_1^2 p = 0.04p$ or $0.25p$.
- $H_1^{(2)}$: $\boldsymbol{\mu}_1 = 0_p$, $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + \mathbf{a}_2$ and $\boldsymbol{\mu}_3 = \boldsymbol{\mu}_2 + \mathbf{a}_2$. This setting reflects the sparse but strong alternative (SSA), which means that the differences are rare, but significant where they appear. We choose $\tau_2 = 1$. Then the magnitude of the difference vector $\mathbf{a}_2$ is always 1.

(2). Dimensions (n_0, p)**:** For simplicity, let $n_1 = n_2 = n_3 \coloneqq n_0$. Then $n = 3n_0$. We select two initial choices for (n_0, p) as $M^{(1)} = (p, n_0) = (5, 8)$ and $M^{(2)} = (p, n_0) = (8, 5)$, with $p < n_0$ and $p > n_0$, respectively. Then we change the magnification factor for the initial choices from 1 to 100 (see the first and sixth columns of Table 3) to investigate the influence of dimensions on the numerical performance.

As in the above Section 6.2, by repeating 10000 times, we can find 10000 rescaled $\lambda_1^{\mathbf{V}}$'s and the proportion of values that are larger than c_α are recorded in Table 3. i.e. $\frac{\#\{\text{rescaled} \quad \lambda_1^{\mathbf{V}} > c_\alpha\}}{10000}$. The two columns titled "H_0" record estimated sizes, from which we can see that the size performance becomes better as the dimensions become larger. This matches with our theoretical conclusion, which relies on $n \to \infty$. Other columns report estimated powers under different mean vectors' settings. Generally speaking, the powers increase fast as the dimensions become larger, say the power values of the $8M^{(i)}$ row already all exceed 0.8. And for small dimensions, the $M^{(1)}$ domain shows better performance than $M^{(2)}$, which indicates that $\lambda_1^{\mathbf{V}}$ prefers $p < n_0$ when both p and n_0 are small. However, for moderate and large dimensions, such preference will be weakened since all the power values are close to 1.

6.4 Performance in multivariate linear model

This subsection is to investigate the performance of our proposed largest eigenvalue $\lambda_1^{\mathbf{M}_1}$ in the multivariate linear model of Section 5. The nominal significance level is set to be $\alpha = 0.05$. The covariance matrix Σ of the error matrix $\mathbf{Z}$ is selected to be a Toeplize matrix with first row $(1, 0.5, 0.5^2, 0.5^3, \cdots, 0.5^{p-1})$, i.e. the covariance matrix for the AR(1) model with the parameter $\sigma_1 = 0.5$. And we use Gamma distribution stated in Section 6.1 **(1)** to generate the data $\mathbf{Z}$. According to Section 5, the distribution of $\mathbf{X}$ does not influence the result. So we simply obtain the entries $\mathbf{X}$ from a uniform distribution $U(-2, 2)$. Considering the conditions on the dimensions, we set an initial choice for (p_1, p_2, N) as $M^{(0)} = (p_1, p_2, N) = (10, 6, 25)$ and then change the magnification factor from 1 to 20 to check the influence of dimensionality. Other settings for the model (5.1) that will be used in the simulation are summarized below.

(1). Parameter matrix B: Set $\mathbf{B} = \begin{pmatrix} (\mathbf{B}_1)_{g_1 \times p_2} \\ (\mathbf{B}_2)_{(p_1-g_1)\times p_2} \end{pmatrix}_{p_1 \times p_2}$. For ease of matrix construction, we let $g_1 = \frac{1}{2}p_1$ in the simulation. $\mathbf{B}_1$ is chosen to be a $(g_1 \times p_2)$ zero matrix, i.e. $\mathbf{B}_1 = 0_{g_1 \times p_2}$. $(\mathbf{B}_2)_{g_1 \times p_2}$ has two different settings.

- $\mathbf{B}_2^{(d)}$: All entries of $\mathbf{B}_2^{(d)}$ are generated from a discrete distribution with probability mass function $P(x = 0.1) = P(x = 0.2) = P(x = 0.3) = 1/3$. Then this $\mathbf{B}_2^{(d)}$ consists of nonzero small components. This corresponds to the DWA (dense but weak alternative) stated in the mean vectors' setting of Section 6.3.
- $\mathbf{B}_2^{(s)}$: The entries of $\mathbf{B}_2^{(s)}$ are all zeros except for the first 2 diagonal elements being ones, i.e. $\mathbf{B}_2^{(s)} = \begin{pmatrix} \mathbf{I}_2 & \\ & 0 \end{pmatrix}$. This corresponds to the SSA (sparse but strong alternative) stated in the mean vectors' setting of Section 6.3.

The two different settings of $\mathbf{B}_2$ are to investigate the power performance of $\lambda_1^{\mathbf{M}_1}$ in testing (5.2) under different alternatives.

(2). Matrix $\mathbf{C}_1$: We consider two special cases: $\mathbf{C}_1^{(a)} = [0, \mathbf{I}_{g_1}]$ and $\mathbf{C}_1^{(b)} = [\mathbf{I}_{g_1}, 0]$.

(3). Matrix $\mathbf{\Gamma}_1$: $\mathbf{\Gamma}_1$ is selected to be $\mathbf{\Gamma}_1^{(a)} = 0$ or $\mathbf{\Gamma}_1^{(b)} = \mathbf{B}_2$.

Four combinations of $(\mathbf{C}_1, \mathbf{B}_2, \mathbf{\Gamma}_1)$ are used in Table 4. For each combination, as in previous sections, by repeating 10000 times, we can find 10000 rescaled $\lambda_1^{\mathbf{M}_1}$'s and the proportion of values that are larger than c_α are recorded in Table 4. i.e. $\frac{\#\{\text{rescaled} \quad \lambda_1^{\mathbf{M}_1} > c_\alpha\}}{10000}$.

The first two combinations are used for size testing. Since the two settings of $\mathbf{B}_2$ are constructed to investigate power performance under different alternatives, for size purpose, we just adopt one of them–$\mathbf{B}_2^{(d)}$. The first combination $(\mathbf{C}_1^{(b)}, \mathbf{B}_2^{(d)}, \mathbf{\Gamma}_1^{(a)})$ is to test whether the first $(g_1 \times p_2)$ block of $\mathbf{B}$ is a zero block, i.e. $H_0 : \mathbf{B}_1 = 0$. The second combination $(\mathbf{C}_1^{(a)}, \mathbf{B}_2^{(d)}, \mathbf{\Gamma}_1^{(b)})$ is to test whether the second $((p_1 - g_1) \times p_2)$ block of $\mathbf{B}$ equals to a given matrix, i.e. $H_0 : \mathbf{B}_2 = \mathbf{\Gamma}_1^{(b)}$. One can observe that the sizes are always close to 0.05, confirming the asymptotic distribution developed for $\lambda_1^{\mathbf{M}_1}$ in Section 5.

The last two combinations are used for power testing. i.e. testing whether $\mathbf{B}_2 = 0$. Two alternatives are considered. The third combination $(\mathbf{C}_1^{(a)}, \mathbf{B}_2^{(d)}, \mathbf{\Gamma}_1^{(a)})$ is for DWA (dense but weak alternative) and the last one $(\mathbf{C}_1^{(a)}, \mathbf{B}_2^{(s)}, \mathbf{\Gamma}_1^{(a)})$ is for SSA (sparse but strong alternative). We can see that for small dimensions, SSA works better than DWA, while as the dimensions increase, a reversal takes place. This is reasonable because the magnitude of difference for DWA is much involved by values of dimensions. And for appropriate large dimensions, all power values are close to 1.

Table 1: Simulated quantiles for rescaled λ_1, i.e. the values $\frac{\#\{\text{rescaled}\quad \lambda_1 \leq \text{“Percentile”}\}}{10000}$ based on 10000 replications under different data distributions and different dimensions.

Standard Normal		$M^{(1)} = (M_1, N_1, N_2, N) = (5, 8, 10, 30)$					$M^{(2)} = (M_1, N_1, N_2, N) = (15, 8, 10, 50)$				
Percentile	TW	$M^{(1)}$	$2M^{(1)}$	$8M^{(1)}$	$16M^{(1)}$	$20M^{(1)}$	$M^{(2)}$	$2M^{(2)}$	$8M^{(2)}$	$16M^{(2)}$	$20M^{(2)}$
-3.90	0.01	0.0132	0.0083	0.0099	0.0080	0.0108	0.0109	0.0102	0.0085	0.0091	0.0090
-3.18	0.05	0.0546	0.0501	0.0497	0.0502	0.0491	0.0514	0.0495	0.0467	0.0450	0.0476
-2.78	0.10	0.1041	0.1011	0.0995	0.1030	0.0992	0.1028	0.0974	0.0981	0.0956	0.0975
-1.91	0.30	0.2941	0.2948	0.3024	0.3026	0.3028	0.3047	0.3049	0.2944	0.2908	0.3004
-1.27	0.50	0.5031	0.5007	0.5026	0.5114	0.5048	0.5072	0.5077	0.4987	0.4971	0.5009
-0.59	0.70	0.7101	0.7057	0.7116	0.7116	0.7081	0.7074	0.7040	0.7075	0.7037	0.7051
0.45	0.90	0.9138	0.9027	0.9050	0.9062	0.9014	0.9055	0.9019	0.9019	0.9048	0.9038
0.98	0.95	0.9610	0.9507	0.9552	0.9538	0.9519	0.9569	0.9525	0.9502	0.9560	0.9560
2.02	0.99	0.9933	0.9896	0.9898	0.9909	0.9912	0.9916	0.9906	0.9900	0.9910	0.9912

Discrete		$M^{(1)} = (M_1, N_1, N_2, N) = (5, 8, 10, 30)$					$M^{(2)} = (M_1, N_1, N_2, N) = (15, 8, 10, 50)$				
Percentile	TW	$M^{(1)}$	$2M^{(1)}$	$8M^{(1)}$	$16M^{(1)}$	$20M^{(1)}$	$M^{(2)}$	$2M^{(2)}$	$8M^{(2)}$	$16M^{(2)}$	$20M^{(2)}$
-3.90	0.01	0.0116	0.0093	0.0094	0.0099	0.0082	0.0099	0.0091	0.0098	0.0104	0.0080
-3.18	0.05	0.0523	0.0464	0.0503	0.0514	0.0477	0.0496	0.0480	0.0529	0.0495	0.0460
-2.78	0.10	0.0996	0.0943	0.1034	0.0983	0.0998	0.0951	0.0986	0.1037	0.0978	0.0974
-1.91	0.30	0.3049	0.2954	0.3054	0.2974	0.3024	0.2933	0.2915	0.3069	0.2968	0.3050
-1.27	0.50	0.5068	0.5002	0.5114	0.4961	0.4984	0.4989	0.4965	0.5069	0.5015	0.4964
-0.59	0.70	0.7124	0.7080	0.7065	0.6986	0.7045	0.7065	0.6976	0.7062	0.7042	0.6946
0.45	0.90	0.9102	0.9098	0.9035	0.9014	0.9021	0.9065	0.9031	0.9058	0.9067	0.8966
0.98	0.95	0.9583	0.9565	0.9537	0.9508	0.9512	0.9559	0.9540	0.9515	0.9546	0.9494
2.02	0.99	0.9931	0.9917	0.9905	0.9911	0.9894	0.9921	0.9903	0.9915	0.9912	0.9894

Gamma(4,0.5)		$M^{(1)} = (M_1, N_1, N_2, N) = (5, 8, 10, 30)$					$M^{(2)} = (M_1, N_1, N_2, N) = (15, 8, 10, 50)$				
Percentile	TW	$M^{(1)}$	$2M^{(1)}$	$8M^{(1)}$	$16M^{(1)}$	$20M^{(1)}$	$M^{(2)}$	$2M^{(2)}$	$8M^{(2)}$	$16M^{(2)}$	$20M^{(2)}$
-3.90	0.01	0.0109	0.0091	0.0099	0.0093	0.0104	0.0098	0.0069	0.0107	0.0096	0.0104
-3.18	0.05	0.0507	0.0502	0.0500	0.0501	0.0507	0.0494	0.0452	0.0503	0.0486	0.0495
-2.78	0.10	0.1025	0.1011	0.0996	0.1013	0.0991	0.1006	0.0957	0.1021	0.0965	0.1002
-1.91	0.30	0.3024	0.2953	0.3008	0.2993	0.2934	0.2985	0.2965	0.2970	0.2972	0.2983
-1.27	0.50	0.4992	0.4994	0.5013	0.4967	0.4865	0.5009	0.4890	0.5028	0.4995	0.5010
-0.59	0.70	0.7033	0.7097	0.6994	0.6935	0.6923	0.7100	0.7006	0.7015	0.7040	0.7080
0.45	0.90	0.9065	0.9062	0.9018	0.9005	0.9023	0.9052	0.9045	0.8970	0.9027	0.9037
0.98	0.95	0.9546	0.9531	0.9503	0.9509	0.9503	0.9515	0.9531	0.9499	0.9503	0.9523
2.02	0.99	0.9908	0.9912	0.9906	0.9895	0.9888	0.9910	0.9901	0.9904	0.9904	0.9900

Table 2: Simulated values for $\frac{\#\{\text{rescaled} \quad \lambda_1^{\mathbf{S}} > c_\alpha\}}{10000}$ based on 10000 replications. So "$\tau = 0$" row records estimated sizes and other rows record estimated powers. The significance level is $\alpha = 0.05$.

$M^{(0)} = (M_1, M_2, N) = (10, 10, 40)$										
	Discrete distribution					Gamma distribution				
τ	$M^{(0)}$	$2M^{(0)}$	$4M^{(0)}$	$8M^{(0)}$	$10M^{(0)}$	$M^{(0)}$	$2M^{(0)}$	$4M^{(0)}$	$8M^{(0)}$	$10M^{(0)}$
0	0.0663	0.0618	0.0622	0.0608	0.0559	0.0672	0.0663	0.0591	0.0589	0.0563
0.1	0.2766	0.5049	0.8428	0.9978	0.9998	0.2932	0.5117	0.8540	0.9981	1.0000
0.15	0.4533	0.7754	0.9915	1.0000	1.0000	0.4641	0.7887	0.9909	1.0000	1.0000
0.2	0.6280	0.9396	0.9999	1.0000	1.0000	0.6483	0.9463	1.0000	1.0000	1.0000
0.25	0.7828	0.9911	1.0000	1.0000	1.0000	0.7959	0.9934	1.0000	1.0000	1.0000
0.3	0.8959	0.9997	1.0000	1.0000	1.0000	0.9113	0.9997	1.0000	1.0000	1.0000
0.4	0.9908	1.0000	1.0000	1.0000	1.0000	0.9920	1.0000	1.0000	1.0000	1.0000

Table 3: Simulated values for $\frac{\#\{\text{rescaled} \quad \lambda_1^{\mathbf{V}} > c_\alpha\}}{10000}$ based on 10000 replications. The "H_0" columns record estimated sizes and other columns record estimated powers. The significance level is $\alpha = 0.05$.

$M^{(1)} = (p, n_0) = (5, 8)$					$M^{(2)} = (p, n_0) = (8, 5)$				
	H_0	$H_1^{(1)}$	$H_1^{(1)'}$	$H_1^{(2)}$		H_0	$H_1^{(1)}$	$H_1^{(1)'}$	$H_1^{(2)}$
$M^{(1)}$	0.0375	0.0831	0.5317	0.5589	$M^{(2)}$	0.0374	0.0511	0.1502	0.1098
$2M^{(1)}$	0.0392	0.2454	0.9955	0.8693	$2M^{(2)}$	0.0399	0.1099	0.7505	0.2449
$4M^{(1)}$	0.0405	0.8535	1.0000	0.9907	$4M^{(2)}$	0.0386	0.4395	1.0000	0.5020
$8M^{(1)}$	0.0414	1.0000	1.0000	1.0000	$8M^{(2)}$	0.0375	0.9956	1.0000	0.8341
$16M^{(1)}$	0.0445	1.0000	1.0000	1.0000	$16M^{(2)}$	0.0424	1.0000	1.0000	0.9897
$32M^{(1)}$	0.0429	1.0000	1.0000	1.0000	$32M^{(2)}$	0.0432	1.0000	1.0000	0.9999
$64M^{(1)}$	0.0396	1.0000	1.0000	1.0000	$64M^{(2)}$	0.0390	1.0000	1.0000	1.0000
$100M^{(1)}$	0.0442	1.0000	1.0000	1.0000	$100M^{(2)}$	0.0452	1.0000	1.0000	1.0000

Table 4: Simulated values for $\frac{\#\{\text{rescaled} \quad \lambda_1^{\mathbf{M}_1} > c_\alpha\}}{10000}$ based on 10000 replications. The first two combinations record estimated sizes and the last two record estimated powers. The significance level is $\alpha = 0.05$.

	$M^{(0)} = (p_1, p_2, N) = (10, 6, 25)$							
$(\mathbf{C}_1, \mathbf{B}_2, \mathbf{\Gamma}_1)$	$M^{(0)}$	$2M^{(0)}$	$3M^{(0)}$	$4M^{(0)}$	$6M^{(0)}$	$8M^{(0)}$	$10M^{(0)}$	$20M^{(0)}$
$(\mathbf{C}_1^{(b)}, \mathbf{B}_2^{(d)}, \mathbf{\Gamma}_1^{(a)})$	0.0400	0.0447	0.0453	0.0469	0.0487	0.0460	0.0466	0.0468
$(\mathbf{C}_1^{(a)}, \mathbf{B}_2^{(d)}, \mathbf{\Gamma}_1^{(b)})$	0.0397	0.0467	0.0450	0.0490	0.0466	0.0470	0.0501	0.0481
$(\mathbf{C}_1^{(a)}, \mathbf{B}_2^{(d)}, \mathbf{\Gamma}_1^{(a)})$	0.2298	0.8923	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
$(\mathbf{C}_1^{(a)}, \mathbf{B}_2^{(s)}, \mathbf{\Gamma}_1^{(a)})$	0.8337	0.9451	0.9821	0.9940	0.9992	1.0000	0.9999	1.0000

7 Appendix

7.1 Outline of The Proof for Theorem 2.1

We first give an outline of the whole proof due to its complexity. Note that the matrix $\mathbf{T}$ does not influence the largest eigenvalue of $\mathbf{\Omega}$ in (2.1) and hence we can directly work on the matrix $(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1\mathbf{U}_1^T\mathbf{X}^T$. However it involves four $\mathbf{X}$ unlike sample covariance matrices. Moreover $\mathbf{X}\mathbf{U}_1\mathbf{U}_1^T\mathbf{X}^T$ is not independent of $\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T$ for general $\mathbf{X}$ (not necessarily consisting of Gaussian entries), which makes it even harder to work on this matrix directly. In view of this, we construct a Wigner-type linearization matrix

$$\mathbf{H} = \mathbf{H}(\mathbf{X}) := \begin{pmatrix} -zI & \mathbf{U}_1^T\mathbf{X}^T & 0 \\ \mathbf{X}\mathbf{U}_1 & 0 & \mathbf{X}\mathbf{U}_2 \\ 0 & \mathbf{U}_2^T\mathbf{X}^T & I \end{pmatrix}. \tag{7.1}$$

As will be seen, the linearization matrix is much more convenient when taking derivative with respect to the entries of $\mathbf{X}$ than $\mathbf{\Omega}$. By the Schur complement formula (7.5) below it turns out that the upper-left block of the 3×3 block matrix $\mathbf{H}^{-1}$ is the Steiltjes transform of $\mathbf{U}_1^T\mathbf{X}^T(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1$ (one can also refer to (8.9) below). It then suffices to consider the linearization matrix $\mathbf{H}$ instead. First the strong local law of $\mathbf{H}^{-1}$ around μ_N (Theorem 7.1 below) is developed which is the main body of the proof. The overall strategy of proving Theorem 7.1 is similar to that used in [16] and it consists of two main parts. Part one is to prove Theorem 7.1 by applying a new Linderberg's comparison approach raised by [16] under the first three moments of the entries of $\mathbf{X}$ matching those of standard Gaussian entries. This part is similar to [12]. However, in order to make this paper more self-consistent and clear, we also repeat the necessary steps but omit some parts done in [12]. Building on part one, part two further proves Theorem 7.1 when the first two moments of the entries of $\mathbf{X}$ match those of standard Gaussian entries (by dropping the 3rd moment matching condition). After that, we use this local law to prove the edge universality (i.e. (2.3) is not affected by the distribution of $\mathbf{X}$) by adopting the strategy stated in [4] and [7]. The proof of Theorem 2.1 is complete by the fact that (2.3) holds because $\mathbf{\Omega}$ becomes a F matrix when $\mathbf{X}$ consists of the Gaussian random variable (one can refer to Theorem 1 of [14] and Theorem 2.1 of [12]).

We would highlight the difference between the proof of this paper and that of [12]. The result about the edge university for F matrices (corresponding to $\mathbf{\Omega}$ in the Gaussian case) in [12] is our starting point because we need to use Linderberg's comparison approach to link the edge universality of $\mathbf{\Omega}$ in the general case to that of F matrices. However, in order to prove the strong local law, a main difficulty is that our main result about $\mathbf{\Omega}$ doesn't assume $\mathbb{E}\mathbf{Z}_{ij}^3 = 0$ (matching the Gaussian third moment), which is much different from the paper [12] when handling the dimension is bigger than the sample size there. As a consequence, the expectation of the higher moments of the variable of interest has to be evaluated by a much more complicated method. For example, in order to calculate the higher moments, we need to extract the i-th row of $\mathbf{X}$ from $\mathbf{\Pi}(z)$ defined at (7.6) below. However $\mathbf{\Pi}(z)$ is a complex function of $\mathbf{X}$, which is not easy to deal with. To handle this, we introduce a transition matrix $\mathbf{\Pi_1}(z)$ (defined at (8.55) in the supplementary file) to find out a compact and manageable expansion of $\Pi(z)$.

7.2 Strong local law

This subsection is to present the strong local law. To this end we present some necessary notations and definitions.

As in [16], we use the following definition to provides a simple way to describe the relationship between two random variables ξ and ζ.

Definition 1. *Let*

$$\xi = \{\xi^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}\}, \quad \zeta = \{\zeta^{(N)}(u) : N \in \mathbb{N}, u \in U^{(N)}\}$$

be two families of nonnegative random variables, where $U^{(N)}$ is a parameter set (can be either dependent on or independent of N). If for all small positive ϵ and σ, there exists a number $N(\epsilon, \sigma)$ only depending on ϵ and σ such that

$$\sup_{u \in U^{(N)}} \mathbb{P}\left[|\xi^{(N)}(u)| > N^{\epsilon}|\zeta^{(N)}(u)|\right] \leq N^{-\sigma}$$

for large enough $N \geq N(\epsilon, \sigma)$, then we say that ζ stochastically dominates ξ uniformly in u. We denote this relationship by $\xi \prec \zeta$ or $\xi = O_{\prec}(\zeta)$. ***If there exists a positive constant c such that $\xi \leq c\zeta$, then we write $\xi \lesssim \zeta$.***

Recall the definition of F in Theorem 2.1. If the entries of $\mathbf{X}$ are Gaussian distributed, then $\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T$ and $\mathbf{X}\mathbf{U}_1\mathbf{U}_1^T\mathbf{X}^T$ are independent and hence $(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1\mathbf{U}_1^T\mathbf{X}^T$ reduces to the F matrix in [12]. From [2] one can then see that $m(z)$ is a unique solution in $\{z \in \mathcal{C}^+\}$ to the following equation

$$\frac{1}{m(z)} = -z + \frac{M_1}{N_1}\int \frac{t}{1+tm(z)}dF(t). \tag{7.2}$$

Define $\rho(x) = \lim_{z\in\mathcal{C}^+\to x} \Im m(z)$. One can see that μ_N defined in (2.5) is the rightmost end point of the support of $\rho(x)$. For the positive constants τ and τ', we define the domains

$$D(\tau, N) := \{z = E + i\eta \in \mathbb{C}^+ : |z| \geq \tau, |E| \leq \tau^{-1}, N^{-1+\tau} \leq \eta \leq \tau^{-1}\}, \tag{7.3}$$

$$D_+ = D_+(\tau, \tau', N) := \{z \in D(\tau, N) : E \geq \mu_N - \tau'\}. \tag{7.4}$$

Let $\mathbf{G}(z) = \mathbf{H}^{-1}$. The explicit expression of $\mathbf{G}(z)$ can be calculated by the following formula

$$\begin{pmatrix} \mathbf{K} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{D}^{-1} \end{pmatrix} + \begin{pmatrix} \mathbf{I} \\ -\mathbf{D}^{-1}\mathbf{C} \end{pmatrix} (\mathbf{K} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} \begin{pmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}. \tag{7.5}$$

To characterize the limit of $\mathbf{G}(z)$ introduce $\Gamma(X, z) = (\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T + m(z)\mathbf{I})^{-1}$ and

$$\Pi(z) = \begin{pmatrix} m(z)\mathbf{I} & 0 & 0 \\ 0 & \Gamma(X, z) & 0 \\ 0 & 0 & \mathbf{I} + \mathbf{U}_2^T\mathbf{X}^T\Gamma(X, z)\mathbf{X}\mathbf{U}_2. \end{pmatrix} \tag{7.6}$$

As will be seen $\mathbf{G}(z)$ is close to $\Pi(Z)$ in D_+. Set

$$\Psi = \Psi(z) = \sqrt{\frac{\Im m(z)}{N\eta}} + \frac{1}{N\eta}.$$

Theorem 7.1. *(Strong local law) Suppose that* $\mathbf{X}$ *satisfies Condition 1. Then*

(i) For any deterministic unit vectors $\mathbf{v}$, $\mathbf{w} \in \mathbb{R}^{N_1+N+M_1-N_2}$,

$$\langle \mathbf{v}, (\mathbf{G}(z) - \Pi(z))\mathbf{w}\rangle \prec \Psi \tag{7.7}$$

uniformly $z \in D_+$ *and*

(ii)

$$|m_N(z) - m(z)| \prec \frac{1}{N\eta} \tag{7.8}$$

uniformly in $z \in D_+$, *where* $m_N(z) = \frac{1}{N_1}\sum_{i=1}^{N_1} G_{ii}$.

Proof. The proof of this theorem is delegated to the supplement. □

7.3 Fluctuation at the right edge and universality

7.3.1 Fluctuation at the right edge

Once Theorem 7.1 is ready it is not hard to show the following Lemma.

Lemma 1. *Under conditions of Theorem 7.1,*

$$\lambda_1 - \mu_N = O_{\prec}(N^{-\frac{2}{3}}).$$

Proof. The proof of this theorem is given in the supplement. □

7.3.2 Universality

We now need edge universality at the rightmost edge of the support. i.e. the limiting distribution of $P(\sigma_N N_1^{2/3}(\lambda_1 - \mu_N) \le s)$ is not affected by the distribution of $\mathbf{X}$. This guarantees Theorem 2.1. Similar to Theorem 6.3 of [7], in order to show Theorem 2.1, it suffices to show the following green function comparison theorem (one can also refer to page 48 of [12] to understand the connection between Theorem 7.2 and (2.3)). The corresponding proof is also provided in the supplement.

Theorem 7.2. *Let $\epsilon > 0$, $\eta = N^{-2/3+\epsilon}$, E_1, $E_2 \in \mathbb{R}$ satisfy $E_1 < E_2$ and*

$$|E_1 - \mu_N|, |E_2 - \mu_N| \le N^{-2/3+\epsilon}.$$

Set $K : \mathbb{R} \to \mathbb{R}$ to be a smooth function such that

$$\max_x |K^{(l)}(x)| \le C, \quad l = 1, 2, 3, 4, 5$$

for some constant C. Then there exists a constant $\phi > 0$ such that for large enough N and small enough ϵ, we have

$$|\mathbb{E}K(N\int_{E_1}^{E_2} \Im m_{X^1}(x+i\eta)dx) - \mathbb{E}K(N\int_{E_1}^{E_2} \Im m_{\mathbf{X}^0}(x+i\eta)dx)| \le N^{-\phi}, \tag{7.9}$$

where the definitions of $\mathbf{X}^0$ and $\mathbf{X}^1$ are given in Section 8.1.1 in the supplement.

8 Local law (7.7)

Throughout the proof we use c, C, K_1 and M_0 to denote some positive constants whose values may differ from line to line. We may assume that $\mathbb{E}X_{ij}^2 = \mathbb{E}X_{it}^2 = \frac{1}{N_1}$ in the sequel. Since N_1 and N are of the same order, when we calculate the upper (lower) bound of some terms, $\mathbb{E}X_{ij}^2$ is usually regarded as $1/N$ for convenience. Before starting the proof, we present some definitions and notations first.

Definition 2. *(Matrix Norms) Let* $\mathbf{A} = (A_{ij})$ *be a matrix. We define the following norms*

$$\|\mathbf{A}\| := \max_{\|\mathbf{x}\|=1} |\mathbf{A}\mathbf{x}|, \quad \|\mathbf{A}\|_\infty := \max_{i,j} |A_{ij}|, \quad \|\mathbf{A}\|_F := \sqrt{tr\mathbf{A}\mathbf{A}^*},$$

where $|\mathbf{x}|$ *is the* L_2 *norm of a vector* $\mathbf{x}$*. Notice that we have the simple inequality*

$$\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\| \leq \|\mathbf{A}\|_F.$$

Definition 3. *We say that an event* Λ *holds with high probability if for any large positive constant* D*, there exists* $n_0(D)$ *such that*

$$\mathbb{P}(\Lambda^c) \leq n^{-D}, \textit{ for any } n \geq n_0(D).$$

In the later proof, we need the following Lemma to control the smallest eigenvalue of $\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T$.

Lemma 2. *Suppose that* $\mathbf{X}$ *satisfies Condition 1 (see the main paper). Then* $\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T$ *is invertible and*

$$\|(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\| \leq M_0 \tag{8.1}$$

for some large constant M_0 *with high probability. Moreover,*

$$\|\mathbf{X}\mathbf{X}^T\| \leq M_0 \tag{8.2}$$

with high probability under Condition 1 as well.

Proof. The proof of this lemma is exactly the same as that of Theorem 3.12 in [16]. □

Moreover, we define the following smooth cutoff function

$$\mathcal{X}(x) = \begin{cases} 1 & \text{if } |x| \leq K_1 N^{-2} \\ 0 & \text{if } |x| \geq 2K_1 N^{-2}, \end{cases}$$

whose derivatives satisfy $|\mathcal{X}^{(k)}| \leq CN^{2k}$, k=1,2,... and K_1 is a positive constant. The purpose of introducing the cutoff function is to help control the minimum eigenvalue and maximum eigenvalue of the random matrices of interest when taking derivatives.

Order the eigenvalues of $\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T$ in a decreasing order as $\tilde{\lambda}_1 \geq ..., \geq \tilde{\lambda}_{M_1}$ and denote its Stieltjes transform by $\tilde{m}_N(z)$. Since

$$\Im(\tilde{m}_N(iN^{-2})) = M_1^{-1}N^{-2}\sum_{i=1}^{M_1} \frac{1}{\tilde{\lambda}_i^2 + N^{-4}}, \tag{8.3}$$

we conclude that if $|\Im(\tilde{m}_N(iN^{-2}))| \lesssim N^{-2}$, then $\tilde{\lambda}_{N-M_2} \gtrsim \frac{1}{N}$. By Lemma 2, choosing a sufficiently small constant c, we have

$$1 - o(N^{-l}) = \mathbb{P}(\tilde{\lambda}_{M_1} \geq c) \leq \mathbb{P}(\Im(\tilde{m}_N(iN^{-2})) \leq K_1N^{-2}), \quad \text{for any positive integer } l. \tag{8.4}$$

Therefore, we have

$$\mathbb{P}(\mathcal{X}(\Im(\tilde{m}_N(iN^{-2}))) \neq 1) \leq o(N^{-l}), \quad \text{for any positive integer } l. \tag{8.5}$$

Similarly, for $\mathbf{X}\mathbf{X}^T$, we have

$$\mathbb{P}(\mathcal{X}(N^{-3}\|\mathbf{X}\|_F^2) \neq 1) \leq o(N^{-l}), \quad \text{for any positive integer } l. \tag{8.6}$$

We set $\mathcal{T}_N(X) := \mathcal{X}(\Im(\tilde{m}_n(iN^{-2}))\mathcal{X}(N^{-3}\|\mathbf{X}\|_F^2)$. In view of (8.5) and (8.6), we can show

$$\mathcal{T}_N(\mathbf{X}) = 1 \tag{8.7}$$

with high probability directly. We will use this conclusion frequently without mention.

Denote the spectral decomposition of $\mathbf{U}_1^T\mathbf{X}^T(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1$ by

$$\mathbf{U}_1^T\mathbf{X}^T(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1 = \sum_{k=1}^{N_1} \lambda_k \mathbf{v}_k\mathbf{v}_k^*,$$

where $\{\mathbf{v}_k\}_{k=1}^{N_1}$ are orthogonal bases of $\mathbb{R}^{I_{N_1}}$. For $1 \leq i, j \leq N_1$ write

$$\mathbf{G}_{ij} = \sum_{k=1}^{N_1} \frac{\mathbf{v}_k(i)\mathbf{v}_k^*(j)}{\lambda_k - z}, \tag{8.8}$$

where $\mathbf{G}_{ij}$ is the entry at the i-th row and j-th column of $\mathbf{G}$ and $\mathbf{v}_k(i)$ is the i-th element of $\mathbf{v}_k$. Define a new matrix $\mathbf{G}_{N_1}$ to be $(\mathbf{G}_{ij})_{1\leq i,j\leq N_1}$.

Moreover, we define

$$\mathbf{A}_2 := \begin{pmatrix} \mathbf{I} & \mathbf{A}_4 \end{pmatrix}^T = \begin{pmatrix} \mathbf{I} & -\mathbf{U}_1^T\mathbf{X}^T(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1} & \mathbf{U}_1^T\mathbf{X}^T(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}XU_2 \end{pmatrix}^T$$

and

$$\mathbf{A}_3 := \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{A}_5 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1} & (\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_2 \\ 0 & \mathbf{U}_2^T\mathbf{X}^T(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1} & \mathbf{I} - P_{\mathbf{U}_2^T\mathbf{X}^T} \end{pmatrix},$$

Via (7.5) we then have an explicit expression of $\mathbf{G}$

$$\mathbf{G} = \mathbf{A}_3 + \sum_{k=1}^{N_1} \frac{\mathbf{A}_2\mathbf{v}_k\mathbf{v}_k^*\mathbf{A}_2^T}{\lambda_k - z} = \mathbf{A}_3 + \mathbf{A}_2\mathbf{G}_{N_1}\mathbf{A}_2^T. \tag{8.9}$$

Definition 4 (moment matching)**.** *Let $\mathbf{X}^1 = (x_{ij}^1)_{M\times N}$ and $\mathbf{X}^0 = (x_{ij}^0)_{M\times N}$ be two complex(or real) matrices satisfying Condition 1. We say that X^1 matches X^0 to order m, if for all $i \in [1, M]$, $j \in [1, N]$, $k, l \geq 0$ and $k + l \in [0, m]$, it has the relationship*

$$\mathbb{E}(\Re(\sqrt{N}x_{ij}^1)^k\Im(\sqrt{N}x_{ij}^1)^l) = \mathbb{E}(\Re(\sqrt{N}x_{ij}^0)^k\Im(\sqrt{N}x_{ij}^0)^l) + O(e^{-(\log N)^C}), \tag{8.10}$$

where C is a constant larger than 1.

We next collect some frequently used bounds. Recall the definition of $m(z)$ in (7.2). For $z \in D(\tau, n)$ one may verify that

$$1 \lesssim |m(z)| \lesssim 1 \tag{8.11}$$

and

$$\eta \lesssim \Im(m(z)). \tag{8.12}$$

(see Lemma 2.3 in [5] or Lemma 3.1 and Lemma 3.2 in [21]). It is obvious that $m(z)$ decides a unique spectral density $\rho(x)$. Recalling the definition of μ_N we write $\mathbf{c} = -\lim_{z\in\mathcal{C}^+\to\mu_N} m(z)$. By checking the proof of Lemmas 1 and 2 in [12] carefully and noting that the proof only relies on the rigidity property of $\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T$, there exists a constant c' such that

$$\limsup_N \left[\mathbf{c}\lambda_{\max}((\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1})\right] \le 1 - c',$$

with high probability. By Lemma A.4 of [16], there exists a constant c'' such that

$$|1 + m(z)\lambda_i((\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1})| \ge c'', \tag{8.13}$$

for all $z \in D_+$ with high probability. Moreover, for $z \in D_+$ it follows from Lemma 2 that

$$\|\Pi(z)\| \prec 1, \text{ and } \|\mathbf{A}_2\| + \|\mathbf{A}_3\| \prec 1. \tag{8.14}$$

To simplify notation, we introduce the following notations with bold lower indices and if the lower index of a matrix is bold, then it represents the inner product and otherwise it means the entry of the corresponding matrix. Specifically

$$\mathbf{A}_{\mathbf{v}s} = \langle \mathbf{v}, \mathbf{A}\mathbf{e}_s\rangle, \ \mathbf{A}_{s\mathbf{v}} = \langle \mathbf{e}_s, \mathbf{A}\mathbf{v}\rangle \text{ and } \mathbf{A}_{\mathbf{v}\mathbf{w}} = \langle \mathbf{v}, \mathbf{A}\mathbf{w}\rangle, \tag{8.15}$$

where $\mathbf{e}_s$ is the unit vector with the s-th coordinate equal to 1. For any $z \in D(\tau, n)$ and fixed $\tau > 0$, we claim that

$$\|\mathbf{G}(z)\mathcal{T}_n(\mathbf{X})\| \lesssim N^9\eta^{-1}, \ \|\partial_z\mathbf{G}(z)\mathcal{T}_n(\mathbf{X})\| \lesssim N^9\eta^{-2}, \tag{8.16}$$

$$\|\mathbf{G}(z)\| \prec \eta^{-1}, \quad \|\partial_z\mathbf{G}(z)\| \prec \eta^{-2}. \tag{8.17}$$

Moreover, suppose that $\mathbf{v} = (\mathbf{v}_1^T, \mathbf{v}_2^T)^T$. Then

$$\sum_{i=1}^{M_1} |\mathbf{G}_{\mathbf{v}i}|^2 = \frac{\Im\mathbf{G}_{\mathbf{v}\mathbf{v}}}{\eta}, \ \|\Pi(z)\mathcal{T}_n(\mathbf{X})\| \lesssim N^4\eta^{-1}, \tag{8.18}$$

and

$$|\mathbf{G}_{\mathbf{v}\mathbf{v}}|^2 \prec \frac{\Im\mathbf{G}_{\mathbf{v}\mathbf{v}}}{\eta} + 1, \tag{8.19}$$

Indeed, the estimates (8.17) and (8.19) follow from Lemma 2 and (8.2). The first equality in (8.18) is straightforward and the second one is from the definition of $\mathcal{T}_n(\mathbf{X})$ directly.

When the entries of $\mathbf{X}$ are Gaussian distributed Theorem 7.1 can be obtained by Theorem 3.6 of [16]. Actually, a key observation is that each block matrix of $\mathbf{G}(z)$ (3×3 block matrix) can be represented as a linear combination of the block matrices of (4.3) in [16] by (8.9) under the Gaussian case. We demonstrate this observation by checking three block matrices of $\mathbf{G}(z)$ and the other blocks can also be inspected similarly. For example, by (8.9), the upper left block of $\mathbf{G}(z)$ is

$$\mathbf{G}_{N_1} = (\mathbf{U}_1^T\mathbf{X}^T(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1 - z\mathbf{I})^{-1}.$$

Since $(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}$ is independent of $\mathbf{X}\mathbf{U}_1$ under the gaussian case $(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}$ can be regarded as a population covariance matrix "Σ". Hence $\mathbf{G}_{N_1}$ is just one block matric of (4.3) in [16]. A second block matrix of $\mathbf{G}(z)$ is $-(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1\mathbf{G}_{N_1}$. It is also a block of (4.3) in [16] by the same reason that $(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}$ is regarded as "Σ". A third block is the second diagonal block matrix of $\mathbf{G}(z)$:

$$-(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1} + (\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{U}_1\mathbf{G}_{N_1}\mathbf{U}_1^T\mathbf{X}^T(\mathbf{X}\mathbf{U}_2\mathbf{U}_2^T\mathbf{X}^T)^{-1},$$

which is also a block of (4.3) in [16]. In fact, the matrix above corresponds to $(\Sigma\mathbf{X}\mathbf{G}_N\mathbf{X}^*\Sigma - \Sigma)$ belonging to (4.3) of [16].

8.1 Proving (7.7) for general distributions under the first three moments matching condition

We now prove (7.7) for general distributions under the condition that $\mathbf{X}$ matches $\mathbf{X}^{Gauss}$ to order 3 in this section, where the entries of $\mathbf{X}^{Gauss}$ follow standard Gaussian distribution. However, the proof of this section is very similar to that of Section 7.1 of [12] (following the strategy in [16]). Hence, we below only give an outline of the arguments in order to prepare notations and tools for the proof under the first two moment matching condition in the next section. One may refer to Section 7.1.1 in [12] for more details.

It suffices to show that for any orthogonal matrix $\mathbf{B}_1$ and $\mathbf{B}_2$,

$$\|\mathbf{B}_1(\mathbf{G}(z) - \Pi(z))\mathbf{B}_2^*\|_\infty \prec \Psi, \tag{8.20}$$

for all $z \in S$, where S is an ϵ-net of D_+ with $\epsilon = N^{-10}$. Setting δ to be a sufficient small positive constant such that $N^{24\delta}\Psi \ll 1$, for any given $\eta \geq \frac{1}{N}$, we define a serial numbers $\eta_0 \leq \eta_1 \leq \eta_2 ... \leq \eta_L$ based on η, where

$$L \equiv L(\eta) := \max\{l \leq \mathbb{N} : \eta N^{l\delta} < N^{-\delta}\}.$$

So

$$\eta_l := \eta N^{l\delta}, \quad (l = 0, 1, ..., L-1), \quad \eta_L := 1.$$

We work on the net S satisfying the condition that $E + i\eta_l \in S, \quad l = 0, ..., L$, from now on. We define $S_m \coloneqq \{z \in S : \Im z \geq N^{-\delta m}\}$ corresponding to the following events:

$$A_m = \{\|\mathbf{B}_1(G(z) - \Pi(z))B_2^*\mathcal{T}_N(\mathbf{X})\|_\infty \prec 1, \text{ for any } z \in S_m\}, \tag{8.21}$$

and

$$C_m = \{\|\mathbf{B}_1(G(z) - \Pi(z))B_2^*\mathcal{T}_N(\mathbf{X})\|_\infty \prec \Psi, \text{ for any } z \in S_m\}. \tag{8.22}$$

We start the induction by considering the event A_0 first. In fact, it is not hard to prove that event A_0 holds. By the assumption that $N^{24\delta}\Psi \ll 1$, it is easy to see that the event C_m implies the event A_m. We will prove the event (A_{m-1}) implies the event (C_m) for all $1 \leq m \leq \delta^{-1}$ in the sequel, which ensure that (8.20) holds on the set S uniformly.

For the purpose, we should calculate the upper bound of the higher moments of the following functions

$$F_{st}(X, z) = (\mathbf{B}_1 G(z)\mathbf{B}_2^*)_{st} - (\mathbf{B}_1\Pi(z)\mathbf{B}_2^*)_{st}\mathcal{T}_N(\mathbf{X}), \tag{8.23}$$

$\mathbf{B}_1, \mathbf{B}_2 \in \mathcal{L} = \{1, \mathbf{\Delta}, \mathbf{V}\}$, $\mathbf{\Delta}$ is defined in (8.31) below and $\mathbf{V}$ is any deterministic orthogonal matrix. By Markov's inequality and (8.7), in order to prove (8.20), it suffices to prove Lemma 3 below.

Lemma 3. *Let p be an positive even constant and $m \leq \delta^{-1}$. Suppose (8.21) for all $z \in S_{m-1}$. Then we have*

$$\mathbb{E}|F_{st}^p(X, z)| \prec (N^{24\delta}\Psi)^p,$$

for all $1 \leq s, t \leq N + N_1 + M_1 - N_2$ and $z \in S_m$.

The proof of Lemma 3 is almost the same as that of [12] under the order 3 moment matching condition.

Lemma 4 (Lemma 5 of [12])**.** *Let ζ be a random variable satisfying $\zeta \prec \nu$ where positive ν may be random or deterministic. Suppose $|\zeta| \leq N^C$ for some positive constant C. Then*

$$\mathbb{E}\zeta \prec (E\nu + N^{C-D}), \tag{8.24}$$

where D is a sufficiently large positive constant.

8.1.1 The proof of Lemma 3 by the interpolation method

We define the interpolation matrix $\mathbf{X}^t$ between $\mathbf{X}^1 = (X_{i\mu}^1) = \mathbf{X}$ and $\mathbf{X}^0 = \mathbf{X}^{Gauss}$ consisting of standard Gaussian random variables below, where $1 \leq i \leq M_1$ and $1 \leq \mu \leq N$.

Definition 5. *For $u \in \{0,1\}$, $1 \le i \le M_1$ and $1 \le \mu \le N$, denote the distribution function of $X_{i\mu}^u$ by $F_{i\mu}^u$. For $\theta \in [0,1]$, we define the distribution function by*

$$F_{i\mu}^{\theta} = \theta F_{i\mu}^{1} + (1-\theta) F_{i\mu}^{0}.$$

The interpolation matrix $\mathbf{X}^\theta$ is $(X_{i\mu}^\theta)$ with $F_{i\mu}^\theta$ being the distribution of $X_{i\mu}^\theta$ and the entries $\{X_{i\mu}^\theta\}$ are mutually independent for all i, μ. Moreover, we introduce the matrix

$$\mathbf{X}_{(i\mu)}^{\theta,\lambda} = \mathbf{X}^\theta + (\lambda - X_{i\mu}^\theta)\mathbf{e}_i \mathbf{e}_\mu^T, \tag{8.25}$$

which differs from $\mathbf{X}^t$ at the (i,μ) position only and the corresponding green functions

$$\mathbf{G}^\theta(z) = \mathbf{G}(\mathbf{X}^\theta, z), \quad \mathbf{G}_{(i\mu)}^{\theta,\lambda}(z) = \mathbf{G}(\mathbf{X}_{(i\mu)}^{\theta,\lambda}, z), \tag{8.26}$$

by replacing $\mathbf{X}$ in $\mathbf{G}(z)$ by $\mathbf{X}^\theta$ and $\mathbf{X}_{(i\mu)}^{\theta,\lambda}$ respectively.

To calculate the difference of $\mathbb{E}|F_{st}(X^1,z)|^p$ and $\mathbb{E}|F_{st}(X^0,z)|^p$, we introduce the following Lemma.

Lemma 5 (Lemma 7.9 of [16])**.** *For any function $F : \mathbb{R}^{M\times N} \to \mathbb{C}$, we have*

$$\mathbb{E}F(\mathbf{X}^1) - \mathbb{E}F(\mathbf{X}^0) = \int_0^1 d\theta \sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\left[\mathbb{E}F(\mathbf{X}_{(i\mu)}^{\theta,X_{(i\mu)}^1}) - \mathbb{E}F(\mathbf{X}_{(i\mu)}^{\theta,X_{(i\mu)}^0})\right]. \tag{8.27}$$

To deal with the right hand side of (8.27), we need to prove the following Lemma.

Lemma 6. *Fix an even positive integer p and $m \le \delta^{-1}$. Suppose that (A_{m-1}) holds. Then there exists some function $B_{st}(.,z)$ such that for $u \in \{0,1\}$*

$$\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\left[\mathbb{E}|F_{st}^p(\mathbf{X}_{(i\mu)}^{\theta,X_{(i\mu)}^u}, z)| - \mathbb{E}|B_{st}^p(\mathbf{X}_{(i\mu)}^{\theta,0}, z)|\right] = O((N^{24\delta}\Psi)^p + \|\mathbb{E}\mathbf{L}_p(\mathbf{X}^\theta, z)\|_\infty), \tag{8.28}$$

where $\mathbf{L}_p(\mathbf{X}^\theta, z) = (|F_{st}^p(X_{(i\mu)}^{\theta,\mathbf{X}_{(i\mu)}^1}, z)|)$.

Lemma 6 concludes that

$$\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\left[\mathbb{E}|F_{st}^p(\mathbf{X}_{(i\mu)}^{\theta,X_{(i\mu)}^1}, z)| - \mathbb{E}|F_{st}^p(\mathbf{X}_{(i\mu)}^{\theta,X_{(i\mu)}^0}, z)|\right] = O((N^{24\delta}\Psi)^p + \|\mathbb{E}\mathbf{L}_p(\mathbf{X}^\theta, z)\|_\infty), \tag{8.29}$$

for all $z \in S_m$. Therefore by Gronnwall's inequality, we can prove Lemma 3. What remains to do is to prove Lemma 6. In the sequel, we only consider the case u=1(u=0 is similar to u=1). First, we calculate the rough bound below, which is a direct conclusion given the event A_{m-1}.

Lemma 7. *Suppose that (8.21) holds for all $z \in S_{m-1}$. Then*

$$\langle \mathbf{v}, (\mathbf{G}(z) - \Pi(z))\mathbf{w} \rangle = O_{\prec}(N^{2\delta})$$

for all $z \in S_m$.

From (8.25), we write

$$\mathbf{X}^{\theta,\lambda_1}_{(i\mu)} - \mathbf{X}^{\theta,\lambda_2}_{(i\mu)} = (\lambda_1 - \lambda_2)\mathbf{e}_i\mathbf{e}_\mu^T.$$

Together with (7.1), one can obtain that

$$\mathbf{H}(\mathbf{X}^{\theta,\lambda_1}_{(i\mu)}) - \mathbf{H}(\mathbf{X}^{\theta,\lambda_2}_{(i\mu)}) = \mathbf{\Delta}^{\lambda_1-\lambda_2}_{(i\mu)}, \tag{8.30}$$

where $\mathbf{H}(\mathbf{X}^{\theta,\lambda_i}_{(i\mu)})$ is obtained from $\mathbf{H}(\mathbf{X})$ in (7.1) with $\mathbf{X}$ replaced by $\mathbf{X}^{\theta,\lambda_i}_{(i\mu)}$ respectively, i=1,2 and

$$\mathbf{\Delta}^{\lambda}_{(i\mu)} = \lambda\Big(\mathbf{\Delta}\mathbf{e}_\mu\mathbf{e}_{i+N_1}^T + \mathbf{e}_{i+N_1}\mathbf{e}_\mu^T\mathbf{\Delta}^T\Big), \ \mathbf{\Delta} = \begin{pmatrix} \mathbf{U}_1^T \\ 0 \\ \mathbf{U}_2^T \end{pmatrix}, \tag{8.31}$$

where and in the following $\mathbf{e}_{i+N_1}$ is always $(M_1+N+N_1-N_2)\times 1$ and $\mathbf{e}_\mu$ is $N\times 1$ vector. **From now on, we denote $i+N_1$ by $\tilde{i}$ for simplicity.** Applying the formula $\mathbf{A}^{-1}-\mathbf{B}^{-1} = -\mathbf{A}^{-1}(\mathbf{A}-\mathbf{B})\mathbf{B}^{-1}$ repeatedly we further obtain the following resolvent formula for any $K \in \mathbb{N}_+$

$$\mathbf{G}^{\theta,\lambda_1}_{(i\mu)} = \mathbf{G}^{\theta,\lambda_2}_{(i\mu)} + \sum_{k=1}^{K}(-1)^k\mathbf{G}^{\theta,\lambda_2}_{(i\mu)}(\mathbf{\Delta}^{\lambda_1-\lambda_2}_{(i\mu)}\mathbf{G}^{\theta,\lambda_2}_{(i\mu)})^k + (-1)^{H+1}\mathbf{G}^{\theta,\lambda_1}_{(i\mu)}(\mathbf{\Delta}^{\lambda_1-\lambda_2}_{(i\mu)}\mathbf{G}^{\theta,\lambda_2}_{(i\mu)})^{K+1}, \tag{8.32}$$

recalling the definition (8.26). Here and the remaining part of this section we drop the variable z when there is no confusion but one should remember that $z \in S_m$.

Lemma 8. *Suppose that λ is a random variable and satisfies $|\lambda| \prec N^{-1/2}$. Then*

$$\|\mathbf{B}_1(G^{\theta,\lambda}_{(i\mu)} - \Pi)\mathbf{B}_2\|_\infty \prec N^{2\delta}. \tag{8.33}$$

In order to simplify the notations, we define

$$f_{(i\mu)}(\lambda) = |F^p_{st}(X^{\theta,\lambda}_{(i\mu)})| = (F_{st}(X^{\theta,\lambda}_{(i\mu)})\overline{F_{st}(X^{\theta,\lambda}_{(i\mu)})})^{\frac{p}{2}},$$

where we omit some parameters.

Lemma 9. *Suppose that λ is a random variable and it satisfies $|\lambda| \prec N^{-1/2}$. Then for any fixed integer n, we have*

$$|f^{(k)}_{(i\mu)}(\lambda)| \prec N^{2\delta(p+k)}. \tag{8.34}$$

Moreover, we have

$$f_{(i\mu)}(\lambda) = \sum_{k=1}^{4p}\frac{\lambda^k}{k!}f^{(k)}_{(i\mu)}(0) + O_{\prec}(\Psi^p) \tag{8.35}$$

by Taylor's expansion.

The proof of Lemmas 8 and 9 can be found in [12]. By Lemma 9 and Lemma 4, we have

$$\mathbb{E}|F_{st}^p(X_{(i\mu)}^{\theta,X_{i\mu}^1})| - \mathbb{E}|F_{st}^p(X_{(i\mu)}^{\theta,0})| = \mathbb{E}f_{(i\mu)}(X_{i\mu}^1) - \mathbb{E}f_{(i\mu)}(0)$$

$$= \frac{1}{2N_1}\mathbb{E}f_{(i\mu)}^{(2)}(0) + \sum_{k=4}^{4p}\frac{1}{k!}\mathbb{E}f_{(i\mu)}^{(k)}(0)\mathbb{E}(X_{i\mu}^1)^k + O_{\prec}(\Psi^p), \tag{8.36}$$

where we use the first three moment matching condition. To show (8.28), it suffices to prove that

$$N^{-k/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}f_{(i\mu)}^{(k)}(0) = O_{\prec}((N^{24\delta}\Psi)^p + \|\mathbb{E}\mathbf{L}_p(X^\theta)\|_\infty), \tag{8.37}$$

for k=4,...,4p. We now point out that $\mathbb{E}|B_{st}(\mathbf{X}_{(i\mu)}^{\theta,0})|^p$ in (8.28) equals

$$\mathbb{E}|F_{st}(\mathbf{X}_{(i\mu)}^{\theta,0})|^p + \frac{1}{2N_1}\mathbb{E}f_{(i\mu)}^{(2)}(0).$$

But we do not prove (8.37) directly. We instead prove (8.38) in order to obtain a self-consistent estimation of X^θ instead. We claim that if

$$N^{-k/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}f_{(i\mu)}^{(k)}(X_{i\mu}^\theta) = O_{\prec}((N^{24\delta}\Psi)^p + \|\mathbb{E}\mathbf{L}_p(X^\theta)\|_\infty), \tag{8.38}$$

holds for k=4,...,8p, then (8.37) holds for n=4,...,4p. The proof of this claim is the same as (7.60)-(7.61) of [12].

It then suffices to prove (8.38). Recall that

$$f_{(i\mu)}^{(k)}(X_{i\mu}^\theta) = \frac{\partial^k\left(|F_{st}(\mathbf{X}_{(i\mu)}^\theta)|^p\right)}{\partial(X_{i\mu}^\theta)^k}, \tag{8.39}$$

where $F_{st}(\cdot)$ is given in (8.23). Since $\mathbf{X}^\theta = \mathbf{X}_{(i\mu)}^{\theta,X_{i\mu}^\theta}$ is the only matrix of interest. We below use $\mathbf{X} = (X_{i\mu})$ instead of $\mathbf{X}^\theta = (X_{i\mu}^\theta)$ to simplify notation because the entries of both of them have bounded higher moments. To prove (8.38) we need to study (8.39).

8.1.2 Estimation of higher order derivatives (8.39) in (8.38)

Before starting Section 8.2, we quote some notations and necessary results from [12] about estimation of higher order derivatives (8.39) in (8.38). By dropping $\mathbf{e}_i\mathbf{e}_\mu^T$ and $\mathbf{e}_\mu\mathbf{e}_i^T$ we define the set

$$\mathcal{Q}(k) = \{\text{The matrices constructed from sum or product of (part of) } \mathbf{U}_2, \mathbf{X},\ \Gamma(\mathbf{X}, z)\}, \tag{8.40}$$

where any kth order derivative of each block of $\Pi(z)$ with respect to $\mathbf{X}_{i\mu}$ belongs to some product(s) between some matrices in $\mathcal{Q}(k)$ and $\mathbf{e}_i\mathbf{e}_\mu^T$ or $\mathbf{e}_\mu\mathbf{e}_i^T$.

To characterize the higher order derivative conveniently we define group g of size k to be the set of paired indices:

$$g = \{s_1 t_1, s_2 t_2, \cdots, s_{k+1} t_{k+1}\},$$

where each of $\{s_j, t_j, j = 1, \cdots, k+1\}$ equals one of four letters $s, t, \tilde{i}, \mu$(recalling $\tilde{i} = i + N_1$). Here we would like to remind the readers that the size of g is k instead of k+1 in order to simplify the arguments in the following proof. Denote the size of the group g by $k = k(g)$ and introduce the set $\mathfrak{G}_k = \{g : \ k(g) = k\}$ consisting of groups of size k. Moreover, we require each group in $\mathfrak{G}_k$ to satisfy three conditions specified below:

(i) $s_1 = s$ and $t_{k+1} = t$.

(ii) For $l \in [2, k+1]$ we have $s_l \in \{\tilde{i}, \mu\}$ and $t_{l-1} \in \{\tilde{i}, \mu\}$.

(iii) For $l \in [1, n]$ we have $t_{l-1} s_l \in \{\tilde{i}\mu, \mu\tilde{i}\}$.

As will be seen, groups g are connected with the high order derivatives of $(\mathbf{B}_1 \mathbf{G}(z) \mathbf{B}_2^T)_{st}$. Moreover write $\mathbf{F}(z) = \sum\limits_{j=1}^{3} \Pi_j(z)$ where each $\Pi_j(z)$ corresponds to a non-zero block of $\Pi(z)$.

Also, to characterize the higher order derivative of each block conveniently we define groups $g^{(j)}$ of size k to be the set of paired indices:

$$g^{(j)} = \{s_{j1} t_{j1}, s_{j2} t_{j2}, \cdots, s_{j(k+1)} t_{j(k+1)}\},$$

where each of s_{jm} and t_{jm} equals one of s, t, i, μ. Moreover introduce the set $\mathfrak{G}_{jk} = \{g^{(j)} : \ k(g^{(j)}) = k\}$ consisting of groups of size k. We require each group in $\mathfrak{G}_{jk}$ to satisfy conditions:

(i) $s_{j1} = s$ and $b_{j(k+1)} = t$.

(ii) For $l \in [2, k+1]$ we have $s_{jl} \in \{i, \mu\}$ and $t_{j(l-1)} \in \{i, \mu\}$.

(iii) For $l \in [1, k]$ we have $t_{j(l-1)} s_{jl} \in \{i\mu, \mu i\}$.

As will be seen groups $g^{(j)}$ are linked to the high order derivatives of $(\mathbf{B}_1 \Pi(z) \mathbf{B}_2^T)_{st}$.

We below associate a random variable $B_{s,t,i,\mu}(g, g^{(1)}, \cdots, g^{(3)})$ with each group $g, g^{(j)}, j = 1, \cdots, 3$. When $k(g) = k(g^{(j)}) = 0$ we define

$$B_{s,t,i,\mu}(g, g^{(1)}, \cdots, g^{(3)})) = (\mathbf{B}_1 \mathbf{G}(z) \mathbf{B}_2^T)_{st} - (\mathbf{B}_1 \Pi(z) \mathbf{B}_2^T)_{st}. \tag{8.41}$$

When $k(g) \geq 1$ or $k(g^{(j)}) \geq 1$, define

$$B_{s,t,i,\mu,\mathbf{R}_{2,\cdots,n},\mathcal{R}_{11,\cdots,3k+1}}(g, g^{(1)}, ..., g^{(3)}) = C_{s,t,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,3k+1}}(g, g^{(1)}, \cdots, g^{(3)})) \tag{8.42}$$

$$- \sum_{j=1}^{3} (\mathbf{B}_1 \mathcal{R}_{j1})_{(s_{j1} t_{j1})} (\mathcal{R}_{j2})_{(s_{j2} t_{j2})} ... (\mathcal{R}_{jk})_{(s_{jn} t_{jk})} (\mathcal{R}_{jk+1} \mathbf{B}_2^T)_{(s_{jk+1} t_{jk+1})},$$

with

$$C_{s,t,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,7k+1}}(g,g^{(1)},\cdots,g^{(3)})) = (\mathbf{B}_1 G\mathbf{A}_5)_{(s_1t_1)}(\mathbf{R}_2)_{(s_2t_2)}...(\mathbf{R}_k)_{(s_kt_k)}(\mathbf{A}_4\mathbf{G}\mathbf{B}_2^T)_{(s_{k+1}t_{k+1})}, \tag{8.43}$$

where $\mathbf{R}_j (2 \le j \le k)$ has the expression of $\mathbf{R}_j = \mathbf{A}_4\mathbf{G}\mathbf{A}_5$ with $\mathbf{A}_4 \in \{1, \mathbf{\Delta}\}$, $\mathbf{A}_5 \in \{1, \mathbf{\Delta}^T\}$ and the non-zero block $\mathcal{R}_{jl}$ belongs to $\mathcal{Q}(k)$ in (8.40). Moreover the selection of 1 and $\mathbf{\Delta}$ in $\mathbf{A}_4$ and $\mathbf{A}_5$ is subject to the constraint that the total number of $\mathbf{\Delta}$ and $\mathbf{\Delta}^T$ contained in $B_{s,t,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,7k+1}}(g,g^{(1)},...,g^{(3)})$ is k. One should also notice that if $k(g)=1$, the terms R_j will disappear.

$$\begin{aligned}&\frac{\partial^k}{\partial(X_{i\mu})^k}\Big([(\mathbf{B}_1\mathbf{G}(z)\mathbf{B}_2^T)_{st} - (\mathbf{B}_1\Pi(z)\mathbf{B}_2^T)_{st}]\mathcal{T}_N(\mathbf{X})\Big) \\ &= (-1)^k \sum_{\substack{g\in\mathfrak{G}_k, g^{(j)}\in\mathfrak{G}_{jk} \\ \mathbf{R}_i, i=2,\ldots,k \\ \mathcal{R}_{jl}, j=1,..3, l=1,\ldots,k+1}} B_{s,t,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,7k+1}}(g,g^{(1)},...,g^{(3)})\mathcal{T}_N(\mathbf{X}) + O_{\prec}(0).\end{aligned} \tag{8.44}$$

To simplify the notations, we furthermore omit $\mathbf{R}_{2\cdots,k}, \mathcal{R}_{11,\ldots,3k+1}, g^{(1)},...,g^{(3)}$ in the sequel and write

$$B_{s,t,i,\mu}(g) = B_{s,t,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,3k+1}}(g,g^{(1)},...,g^{(3)}), \tag{8.45}$$

$$C_{s,t,i,\mu}(g) = C_{s,t,i,\mu,\mathbf{R}_{2,\cdots,k},\mathcal{R}_{11,\cdots,3k+1}}(g,g^{(1)},...,g^{(3)}), \tag{8.46}$$

(here one should notice that the sizes of g and $g^{(j)}$ are the same according to definition (8.42)). Hence we have

$$\begin{aligned}&\frac{\partial^k}{\partial(X_{i\mu})^k}\Big(|F_{st}(\mathbf{X})|^p\Big) = (-1)^k \sum_{\substack{k_1,\ldots,k_{p/2},\tilde{k}_1,\ldots,\tilde{k}_{p/2}\in\mathbb{N} \\ \sum_r(k_r+\tilde{k}_r)=k}} \frac{k!}{\prod_r k_r!\tilde{k}_r!} \\ &\times \prod_{r=1}^{\frac{p}{2}}(\sum_{\substack{g_r\in\mathfrak{G}_{k_r}\cup\mathfrak{G}_{jk_r} \\ \mathbf{R}_i, i=2,\ldots,k \\ \mathcal{R}_{jl}, j=1,..3, l=1,\ldots,k+1}} \sum_{\substack{\tilde{g}_r\in\mathfrak{G}_{\tilde{k}_r}\cup\mathfrak{G}_{j\tilde{k}_r} \\ \bar{\mathbf{R}}_i, i=2,\ldots,k \\ \bar{\mathcal{R}}_{jl}, j=1,..3, l=1,\ldots,k+1}} B_{s,t,i,\mu}(g_r)\overline{B_{s,t,i,\mu}(\tilde{g}_r)}\mathcal{T}_N^2(\mathbf{X})) + O_{\prec}(0),\end{aligned} \tag{8.47}$$

where $g_r \in \mathfrak{G}_{k_r} \cup \mathfrak{G}_{jk_r}$ means that the groups associated with the derivatives of $\mathbf{G}(z)$ belong to $\mathfrak{G}_{k_r}$ and the groups associated with the derivatives of $\Pi(z)$ belong to $\mathfrak{G}_{jk_r}$. In view of (8.47) and (8.39) to prove (8.38) it then suffices to show that

$$N^{-k/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\left[\prod_{r=1}^{p/2} B_{s,t,i,\mu}(g_r)\overline{B_{s,t,i,\mu}(\tilde{g}_r)}\mathcal{T}_N^p(\mathbf{X})\right] = O((N^{24\delta}\Psi)^p + \|\mathbb{E}\mathbf{L}_p(\mathbf{X})\|_\infty), \tag{8.48}$$

for $4 \le k \le 8p$ and groups $g_r \in \mathfrak{G}_{k_r} \cup \mathfrak{G}_{jk_r}$ satisfying $\sum_r k(g_r) = k$ and $k(g_0) = 0$. Define

$$\mathcal{H}_i = \mathcal{H}_{1i} + \mathcal{H}_{sti}, \quad \mathcal{H}_{1i} = |(\mathbf{B}_1\mathbf{G})_{s\tilde{i}}| + |(\mathbf{G}\mathbf{B}_2^T)_{\tilde{i}t}|, \quad \mathcal{H}_{sti} = \sum_{\mathcal{R}\in\mathcal{Q}(n)}(|(\mathbf{B}_1\mathcal{R})_{si}| + |(\mathcal{R}\mathbf{B}_2^T)_{it}|),$$

$$\mathcal{H}_\mu = \mathcal{H}_{1\mu} + \mathcal{H}_{st\mu}, \quad \mathcal{H}_{1\mu} = |(\mathbf{B}_1\mathbf{G}\mathbf{\Delta})_{s\mu}| + |(\mathbf{\Delta}^T\mathbf{G}\mathbf{B}_2^T)_{\mu t}|, \quad \mathcal{H}_{st\mu} = \sum_{\mathcal{R}\in\mathcal{Q}(k)}(|(\mathbf{B}_1\mathcal{R})_{a\mu}| + |(\mathcal{R}\mathbf{B}_2^T)_{\mu t}|.$$

By the same arguments from (7.77)-(7.85) in [12], we have

$$|B_{s,t,i,\mu}(g_r)| \prec N^{2\delta(k(g)+1)}, \tag{8.49}$$

(recall $k(g)=k(g^{(j)})$ from definition (8.42)). Likewise, for $k(g)\geq 1$, we have

$$|B_{s,t,i,\mu}(g_r)| \prec (\mathcal{H}_i^2+\mathcal{H}_\mu^2)N^{2\delta(k(g_r)-1)}, \tag{8.50}$$

while k(g)=1,

$$|B_{s,t,i,\mu}(g_r)| \prec \mathcal{H}_i\mathcal{H}_\mu. \tag{8.51}$$

$$\sum_{i=1}^{M_1}\mathcal{H}_{1i}^2+\sum_{\mu=1}^{N}\mathcal{H}_{1\mu}^2 \prec N\phi_s^2+N\phi_t^2, \tag{8.52}$$

$$\sum_{i \text{ or } s \text{ or } t}^{M_1}\mathcal{H}_{sti}^2+\sum_{s \text{ or } \mu}^{N}\mathcal{H}_{s\mu}^2 \prec 1, \tag{8.53}$$

where i or s or t means the summation over either i or s or t and

$$\phi_s^2=\frac{\Im(\mathbf{BGB}^*)_{ss}+\eta}{N\eta},$$

with $\mathbf{B}\in\mathcal{L}$ defined in (8.23).

8.2 (7.7) under the condition (2.2)

This subsection is to remove the 3rd moment matching condition needed in the previous subsection. The proving strategy is similar to that in [16]. Note that we have used the first three moments matching condition when obtaining (8.36). We now have to estimate the term involving the third derivative in (7.7) (there n now starts from three). To this end, it is enough to prove (8.38) for $k=3$. This further reduces to proving that (8.48) holds for k=3 as well.

Define $\mathbf{v},\mathbf{w}\in\{\mathbf{B}_{(1)},...,\mathbf{B}_{(N-N_2+N_1+M_1)}\}$, where $\mathbf{B}_{(i)}$ represents the i-th column of $\mathbf{B}$ with $\mathbf{B}\in\mathcal{L}$, recalling $\mathcal{L}$ defined immediately below (8.23). In the sequel, we focus on $\langle\mathbf{v},(\mathbf{G}(z)-\Pi(z))\mathbf{v}\rangle$ only because the general inner product $\langle\mathbf{v},(\mathbf{G}(z)-\Pi(z))\mathbf{w}\rangle$ can be handled by the equality that

$$\langle\mathbf{v},(\mathbf{G}(z)-\Pi(z))\mathbf{w}\rangle=\frac{1}{2}\big(\langle\mathbf{v}+\mathbf{w},(\mathbf{G}(z)-\Pi(z))(\mathbf{v}+\mathbf{w})\rangle-\langle\mathbf{w},(\mathbf{G}(z)-\Pi(z))\mathbf{w}\rangle-\langle\mathbf{v},(\mathbf{G}(z)-\Pi(z))\mathbf{v}\rangle\big).$$

Here we absorb $\mathbf{e}_s^T\mathbf{B}_1$ and $\mathbf{B}_2\mathbf{e}_t$ used in (8.23) into new vectors $\mathbf{v}$ and $\mathbf{w}$. As a consequence denote $B_{s,t,i,\mu}(g)$ used in (8.48) by $A_{\mathbf{v},i,\mu}(g)$ and ignore the conjugate symbol there for simplicity. Below we take derivatives of $\mathbf{G}$ and $\mathbf{\Pi}(z)$ with respect to $X_{i\mu}$. First, we calculate the derivative of $\mathbf{G}(z)$ with respect to $X_{i\mu}$.

To this end we expand $\mathbf{G}$ in terms of the $\tilde{i}$-th row of $\mathbf{X}$. Let $\mathbf{H}^{(\tilde{i})}$ be the submatrix obtained from $\mathbf{H}$ by deleting its $\tilde{i}$-th row and $\tilde{i}$-th column (deleting the i-th row of $\mathbf{X}$ involved in $\mathbf{H}$) and define

$\mathbf{G}^{(\tilde{i})} = (\mathbf{H}^{\tilde{i}})^{-1}$. Recalling the notations (8.15) and referring to the resolvent expansion formula (8.3) of [16] we have the following expansion

$$\begin{aligned}\mathbf{G}_{\mathbf{uw}} &= \mathbf{u}(\tilde{i})\mathbf{w}(\tilde{i})\mathbf{G}_{\tilde{i}\tilde{i}} + \mathbf{G}^{(\tilde{i})}_{\mathbf{uw}} + \mathbf{G}_{\tilde{i}\tilde{i}}(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{X}^T)_{\mathbf{u}i}(\mathbf{X}\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{i\mathbf{w}} - \mathbf{u}(\tilde{i})\mathbf{G}_{\tilde{i}\tilde{i}}(\mathbf{X}\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{i\mathbf{w}} \\ &\quad - \mathbf{w}(\tilde{i})\mathbf{G}_{\tilde{i}\tilde{i}}(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{X}^T)_{\mathbf{u}i}. \end{aligned} \tag{8.54}$$

In fact, if we exchange the first with second "row" of $\mathbf{H}$ and its first "column" with its second "column" (i.e. we convert $\mathbf{H}$ to

$$\begin{pmatrix} 0 & \mathbf{X}\mathbf{U}_1 & \mathbf{X}\mathbf{U}_2 \\ \mathbf{U}_1^T\mathbf{X}^T & -zI & 0 \\ \mathbf{U}_2^T\mathbf{X}^T & 0 & I \end{pmatrix}$$

), then (8.54) is similar to the formula (8.3) of [16]. Here one may change the subscripts s and t in (8.23) when necessary, which can still be absorbed into the vector $\mathbf{v}$.

The next aim is to correspondingly extract the entries $X_{i\mu}$ of $\mathbf{X}$ from $\mathbf{\Pi}(z)$ such that taking expectation on $X_{i\mu}$ later may use the independence between $\mathbf{\Pi}^{(i)}$ (defined below) and $X_{i\mu}$, which serves the same function as (8.54). Since it is complicated to extract $X_{i\mu}$ from $\mathbf{\Pi}(z)$ we construct a proxy matrix $\mathbf{\Pi}_1$ below. Define

$$\mathbf{\Pi}_1 = \mathbf{H}_1^{-1} = \begin{pmatrix} m(z)\mathbf{I} & \mathbf{X}\mathbf{U}_2 \\ \mathbf{U}_2^T\mathbf{X}^T & -\mathbf{I} \end{pmatrix}^{-1}. \tag{8.55}$$

By (7.5) we have

$$\mathbf{\Pi}_1 = \begin{pmatrix} \Gamma(\mathbf{X},z) & \Gamma(\mathbf{X},z)\mathbf{X}\mathbf{U}_2 \\ \mathbf{U}_2^T\mathbf{X}^T\Gamma(\mathbf{X},z) & -\mathbf{I} + \mathbf{U}_2^T\mathbf{X}^T\Gamma(\mathbf{X},z)\mathbf{X}\mathbf{U}_2 \end{pmatrix}. \tag{8.56}$$

The key observations are that the first diagonal matrix of $\mathbf{\Pi}_1$ is the same as the second block of $\mathbf{\Pi}$ and that the second diagonal block of $\mathbf{\Pi}_1$ and the third block of $\mathbf{\Pi}$ differ by $2\mathbf{I}$. Since $\mathbf{\Pi}(z)$ is a 3×3 block diagonal matrix, we can split $\mathbf{v}$ into 3 parts $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$: $\mathbf{v} = (\mathbf{v}_1^T, \mathbf{v}_2^T, \mathbf{v}_3^T)^T$ corresponding to each block. In other words,

$$\Pi_{\mathbf{vv}} = m(z)\mathbf{v}_1^T\mathbf{v}_1 + \Gamma(\mathbf{X},z)_{\mathbf{v}_2\mathbf{v}_2} + (\mathbf{I} + \mathbf{U}_2^T\mathbf{X}^T\Gamma(\mathbf{X},z)\mathbf{X}\mathbf{U}_2)_{\mathbf{v}_3\mathbf{v}_3} = \Pi_{\hat{\mathbf{v}}_1\hat{\mathbf{v}}_1} + \Pi_{\hat{\mathbf{v}}_2\hat{\mathbf{v}}_2} + \Pi_{\hat{\mathbf{v}}_3\hat{\mathbf{v}}_3},$$

where $\hat{\mathbf{v}}_1 = (\mathbf{v}_1^T, 0, 0)^T$, $\hat{\mathbf{v}}_2 = (0, \mathbf{v}_2^T, 0)^T$, $\hat{\mathbf{v}}_3 = (0, 0, \mathbf{v}_3^T)^T$ and their sizes are the same as $\mathbf{v}$'s size. Moreover, we set $\tilde{\mathbf{v}}_2 = (\mathbf{v}_2^T, 0)^T$ and $\tilde{\mathbf{v}}_3 = (0, \mathbf{v}_3^T)^T$, where their sizes are both equal to the size of $(\mathbf{v}_2^T, \mathbf{v}_3^T)^T$. It follows that

$$\Pi_{\hat{\mathbf{v}}_2\hat{\mathbf{v}}_2} = (\Pi_1)_{\tilde{\mathbf{v}}_2\tilde{\mathbf{v}}_2}. \tag{8.57}$$

Let $\Pi_1^{(i)} = (\mathbf{H}_1^{(i)})^{-1}$, where $\mathbf{H}_1^{(i)}$ is the sub matrix of $\mathbf{H}_1$ by deleting its i-th row and i-th column. One can similarly define $\boldsymbol{\Pi}^{(i)}$ from $\boldsymbol{\Pi}$. Applying (8.54) (or formula (8.3) of [16]) to Π_1 and by (8.56) we have

$$(\Pi_1)_{\tilde{\mathbf{v}}_3\tilde{\mathbf{v}}_3} = (\Pi_1^{(i)})_{\tilde{\mathbf{v}}_3\tilde{\mathbf{v}}_3} + (\Pi_1)_{ii}(\Pi_1^{(i)}\mathbf{U}_2^T\mathbf{X}^T)_{\tilde{\mathbf{v}}_3 i}(\mathbf{X}\mathbf{U}_2\Pi_1^{(i)})_{i\tilde{\mathbf{v}}_3}, \quad i \in (1, M_1)$$

(in this case note that $\tilde{\mathbf{v}}_3(i) = 0$ for $i \in \{1, 2, ..., M_1\}$). Together with $\Pi_{\hat{\mathbf{v}}_3\hat{\mathbf{v}}_3} = 2\mathbf{v}_3^T\mathbf{v}_3 + (\Pi_1)_{\tilde{\mathbf{v}}_3\tilde{\mathbf{v}}_3}$ and $\Pi^{(i)}_{\hat{\mathbf{v}}_3\hat{\mathbf{v}}_3} = 2\mathbf{v}_3^T\mathbf{v}_3 + (\Pi_1^{(i)})_{\tilde{\mathbf{v}}_3\tilde{\mathbf{v}}_3}$, we conclude that

$$\Pi_{\hat{\mathbf{v}}_3\hat{\mathbf{v}}_3} = \Pi^{(i)}_{\hat{\mathbf{v}}_3\hat{\mathbf{v}}_3} + (\Pi_1)_{ii}(\Pi_1^{(i)}\mathbf{U}_2^T\mathbf{X}^T)_{\tilde{\mathbf{v}}_3 i}(\mathbf{X}\mathbf{U}_2\Pi_1^{(i)})_{i\tilde{\mathbf{v}}_3}. \tag{8.58}$$

Similarly via (8.54) and (8.57) we have

$$\begin{aligned}\Pi_{\hat{\mathbf{v}}_2\hat{\mathbf{v}}_2} &= \mathbf{v}(\tilde{i})^2\Pi_{\tilde{i}\tilde{i}} + \Pi^{(i)}_{\hat{\mathbf{v}}_2\hat{\mathbf{v}}_2} + (\Pi_1)_{ii}(\Pi_1^{(i)}\mathbf{U}_2^T\mathbf{X}^T)_{\tilde{\mathbf{v}}_2 i}(\mathbf{X}\mathbf{U}_2\Pi_1^{(i)})_{i\tilde{\mathbf{v}}_2} \\ &\quad -\mathbf{v}(\tilde{i})(\Pi_1)_{ii}(\Pi_1^{(i)}\mathbf{U}_2^T\mathbf{X}^T)_{\tilde{\mathbf{v}}_2 i} - \mathbf{v}(\tilde{i})(\Pi_1)_{ii}(\mathbf{X}\mathbf{U}_2\Pi_1^{(i)})_{i\tilde{\mathbf{v}}_2}.\end{aligned} \tag{8.59}$$

One should notice that we will subtract $\mathbf{v}(\tilde{i})^2\Pi_{\tilde{i}\tilde{i}}$ at (8.60). In the sequel we use $\mathbf{v}_2$ (also $\mathbf{v}_3$) to represent $\mathbf{v}_2$, $\hat{\mathbf{v}}_2$ or $\tilde{\mathbf{v}}_2$ depending on the dimension of the matrix we deal with if there is no confusion.

Recalling $\tilde{i} = i + N_1$, since the expectation of $(\mathbf{G}_{\tilde{i}\tilde{i}} - \Pi_{\tilde{i}\tilde{i}})$ is difficult to handle in the following proof, we replace $\mathbf{A}_{\mathbf{v},i,\mu}(g)$ by $\hat{\mathbf{A}}_{v,i,\mu}(g)$:

$$\hat{\mathbf{A}}_{\mathbf{v},i,\mu}(g) = \begin{cases} \mathbf{G}_{\mathbf{vv}} - \Pi_{\mathbf{vv}} - \mathbf{v}(\tilde{i})^2(\mathbf{G}_{\tilde{i}\tilde{i}} - \Pi_{\tilde{i}\tilde{i}}) & \text{if } k(g) = 0 \\ \mathbf{A}_{\mathbf{v},i,\mu}(g) & \text{if } k(g) \geq 1 \end{cases} \tag{8.60}$$

By the following Lemma, it suffices to show (8.48) holds for $\hat{\mathbf{A}}_{\mathbf{v},i,\mu}(g)$ when k=3:

Lemma 10. *If (8.48) holds for $\hat{\mathbf{A}}_{\mathbf{v},i,\mu}(g)$ when k=3, then it also holds for $\mathbf{A}_{\mathbf{v},i,\mu}(g)$ when k=3.*

Proof. The proof of this lemma is similar to that in [16]. □

Referring to (8.54), one can find out that the expansion of $\mathbf{G}_{\mathbf{uw}}$ can be reorganized as follows

$$\mathbf{G}_{\mathbf{uw}} = \mathbf{G}^{(0)}_{\mathbf{uw}} + \mathbf{G}_{\tilde{i}\tilde{i}}\mathbf{G}^{(1)}_{\mathbf{uw}} + \mathbf{v}(\tilde{i})G_{\tilde{i}\tilde{i}}\mathbf{G}^{(2)}_{\mathbf{uw}}, \tag{8.61}$$

where $\mathbf{u}, \mathbf{w} \in \{\mathbf{v}, \mathbf{e}_{\tilde{i}}, \Delta\mathbf{e}_\mu\}$, $\mathbf{u}$ and $\mathbf{w}$ can not be equal to $\mathbf{v}$ at the same time. Moreover, $\mathbf{G}^{(0)}_{\mathbf{uw}}$ is $\mathbf{X}^{(i)}$ measurable (obtained from $\mathbf{X}$ by deleting the i th row) and independent of the i-th row of $\mathbf{X}$, and $\mathbf{G}^{(1)}_{\mathbf{uw}}$ and $\mathbf{G}^{(2)}_{\mathbf{uw}}$ do not include $\mathbf{G}_{\tilde{i}\tilde{i}}$ and $\mathbf{v}(\tilde{i})$. We illustrate some examples as follows:

$$\mathbf{G}_{\mathbf{v}\tilde{i}} = \mathbf{v}(\tilde{i})\mathbf{G}_{\tilde{i}\tilde{i}} - \mathbf{G}_{\tilde{i}\tilde{i}}(\mathbf{G}^{(\tilde{i})}\Delta\mathbf{X}^T)_{\mathbf{v}i} = \mathbf{G}_{\tilde{i}\tilde{i}}\mathbf{G}^{(1)}_{\mathbf{v}\tilde{i}} + \mathbf{v}(\tilde{i})\mathbf{G}_{\tilde{i}\tilde{i}}\mathbf{G}^{(2)}_{\mathbf{v}\tilde{i}}, \tag{8.62}$$

$$\begin{aligned}\mathbf{G}_{\mathbf{v}\mathbf{v}_\mu} &= \mathbf{G}^{(\tilde{i})}_{\mathbf{v}\mathbf{v}_\mu} + \mathbf{G}_{\tilde{i}\tilde{i}}(\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta}\mathbf{X}^T)_{\mathbf{v}i}(\mathbf{X}\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})})_{i\mathbf{v}_\mu} - \mathbf{v}(\tilde{i})\mathbf{G}_{\tilde{i}\tilde{i}}(\mathbf{X}\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})})_{i\mathbf{v}_\mu} \\ &= \mathbf{G}^{(0)}_{\mathbf{v}\mathbf{v}_\mu} + \mathbf{G}_{\tilde{i}\tilde{i}}\mathbf{G}^{(1)}_{\mathbf{v}\mathbf{v}_\mu} + \mathbf{v}(\tilde{i})\mathbf{G}_{\tilde{i}\tilde{i}}\mathbf{G}^{(2)}_{\mathbf{v}\mathbf{v}_\mu},\end{aligned} \tag{8.63}$$

$$\mathbf{G}_{\tilde{i}\mathbf{v}_\mu} = -\mathbf{G}_{\tilde{i}\tilde{i}}(\mathbf{X}\Delta^T\mathbf{G}^{(\tilde{i})})_{i\mathbf{v}_\mu} = \mathbf{G}_{\tilde{i}\tilde{i}}\mathbf{G}^{(1)}_{\tilde{i}\mathbf{v}_\mu}, \tag{8.64}$$

where $\mathbf{v}_\mu = \Delta \mathbf{e}_\mu$.

To conveniently write down the expansions of $\mathbf{G}_{\mathbf{uw}}$ and $\mathbf{\Pi}$ in $A_{\mathbf{v},i,\mu}(g)$ such as (8.54), (8.58) and (8.59) in terms of $\mathbf{G}^{(j)}_{\mathbf{uw}}, j = 0, 1, 2,$ we below introduce the definitions of a tagged group, a refinement of the preceding definition of group $A_{\mathbf{v},i,\mu}(g)$, as in [16].

Definition 6. *A tagged group is a pair (g, σ), where $\sigma = (\sigma(l)), \cdots, \sigma(k(g)+1)$ with $\sigma(l) \in \{0, 1, 2\}$ (denote it by $\sigma = (\sigma(l))_{l=1}^{k(g)+1}$).*

(i) If $k(g) = 0$, we set

$$A_{\mathbf{v},i,\mu}(g,0) = \mathbf{G}^{(\tilde{i})}_{\mathbf{vv}} - \mathbf{\Pi}^{(\tilde{i})}_{\mathbf{vv}} = \mathbf{G}^{(\tilde{i})}_{\mathbf{vv}} - \Pi^{(i)}_{\mathbf{v}_1\mathbf{v}_1} - \Pi^{(i)}_{\mathbf{v}_2\mathbf{v}_2} - \Pi^{(i)}_{\mathbf{v}_3\mathbf{v}_3},$$

$$\begin{aligned} A_{\mathbf{v},i,\mu}(g,1) \quad &= \quad (\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{X}^T)_{\mathbf{v}i}(\mathbf{X}\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{i\mathbf{v}} - (\Pi_1)_{ii}(\Pi_1^{(i)}\mathbf{U}_2^T\mathbf{X}^T)_{\mathbf{v}_2 i}(\mathbf{X}\mathbf{U}_2\Pi_1^{(i)})_{i\mathbf{v}_2} \\ &\quad -(\Pi_1)_{ii}(\Pi_1^{(i)}\mathbf{U}_2^T\mathbf{X}^T)_{\mathbf{v}_3 i}(\mathbf{X}\mathbf{U}_2\Pi_1^{(i)})_{i\mathbf{v}_3}, \end{aligned}$$

$$A_{\mathbf{v},i,\mu}(g,2) = -(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{X}^T)_{\mathbf{v}i} - (\mathbf{X}\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{i\mathbf{v}} + (\Pi_1)_{ii}(\Pi_1^{(i)}\mathbf{U}_2^T\mathbf{X}^T)_{\mathbf{v}_2 i} + (\Pi_1)_{ii}(\mathbf{X}\mathbf{U}_2\Pi_1^{(i)})_{i\mathbf{v}_2}.$$

(ii) If $k(g) \geq 1$, we define

$$A_{\mathbf{v},i,\mu}(g,\sigma) = [\mathbf{G}_{\mathbf{vt}_1}]^{\sigma(1)}[G_{\mathbf{s}_2\mathbf{t}_2}]^{\sigma(2)}...[G_{\mathbf{s}_{n(\omega)+1}\mathbf{v}}]^{\sigma(k(g)+1)}.$$

Here one should notice that the second term at the right hand side of (8.42) is ignored since we will discuss how to deal with it later.

In the above definition, we write $\Pi_1^{(i)}\mathbf{U}_2^T\mathbf{X}^T = \Pi_1^{(i)}(0\ \mathbf{X}\mathbf{U}_2)^T$ for simplicity. When $k(g) = 0$, $A_{\mathbf{v},i,\mu}(g,j), j = 1, 2, 3$ come from the expansion (see (8.54), (8.58) and (8.59)) of $(\mathbf{G}_{\mathbf{vv}} - \mathbf{\Pi}_{\mathbf{vv}})$ by deducting the diagonal entry and one may refer to (8.41). Observe that $\mathbf{A}_{\mathbf{v},i,\mu}(g,\sigma)$ is a homogeneous polynomial of the variable $X_{i\mu}, \mu = 1, ..., N$ with the coefficients being $\mathbf{X}^{(i)}$-measurable. One should also notice that we have not considered the derivative of $\Pi(z)$ for $k(g) \geq 1$ to be discussed at the end of this section (refer to (8.42)). Also in the sequel, we omit the terms involving Π_1 or $\Pi_1^{(i)}$ for $k(g) = 0$, which can be handled similarly by checking the following arguments carefully. Below we frequently replace $\mathcal{T}_N(\mathbf{X})$ by $\mathcal{T}_N(\mathbf{X}^{(i)})$ where $\mathcal{T}_N(\mathbf{X}^{(i)})$ is obtained from $\mathcal{T}_N(\mathbf{X}))$ with $\mathbf{X}$ replaced by $\mathbf{X}^i$. The purpose of such a replacement is that we need to extract $X_{i\mu}$, $\mu = 1, ..., N$ from all $\mathbf{X}$ involved in $\mathbf{G}$, $\mathbf{\Pi}$ and $\mathcal{T}_N(\mathbf{X}^{(i)})$ so that we may use independence between $X_{i\mu}$ and $\mathbf{G}^{(\tilde{i})}$, $\mathbf{\Pi}^i$ and $\mathcal{T}_N(\mathbf{X}^i)$. The replacement starts from $\mathcal{T}_N(\mathbf{X})$ to $\mathcal{T}_N(\mathbf{X})\mathcal{T}_N(\mathbf{X}^{(i)})$ and finally to $\mathcal{T}_N(\mathbf{X}^{(i)})$. As before one should note that $\mathcal{T}_N(X^{(i)}) = 1$ with high probability (recalling (8.7)). However in order to simplify notation we do not explicitly write down such steps below and only state $\mathcal{T}_N(\mathbf{X}^{(i)})$ and $\mathcal{T}_N(\mathbf{X})$ whenever necessary.

For $\sigma = (\sigma(l))_{l=1}^{k(g)+1}$ define

$$|\sigma|_{\mathbf{i}} = \sum_l \mathbf{I}(\sigma(l) \geq 1), \quad |\sigma|_{\mathbf{v}} = \sum_l \mathbf{I}(\sigma(l) = 2).$$

Notice that $|\sigma|_{\mathbf{i}}$ and $|\sigma|_{\mathbf{v}}$ do not depend on $\mathbf{i}$ and $\mathbf{v}$ since the expansion (8.54) always works for any i and $\mathbf{v}$. From the above we may write

$$\hat{\mathbf{A}}_{\mathbf{v},i,\mu}(g) = \sum_{\sigma_r} \mathbf{A}_{\mathbf{v},i,\mu}(g,\sigma_r)(\mathbf{G}_{\tilde{i}\tilde{i}})^{|\sigma_r|_i}\mathbf{v}(\tilde{i})^{|\sigma_r|_{\mathbf{v}}}$$

where $\sigma_r = (\sigma_r(l))_{l=1}^{k(g)+1}$ with $\sigma_r(l) \in \{0,1,2\}$. In view of the arguments above, it suffices to show the following Lemma.

Lemma 11. *Suppose that for $r \leq q$, $k(g_r) \geq 1$ and $k(g_r) = 0$ for $p \geq r \geq q+1$ subject to $\sum_r k(g_r) = 3$. For r=1,...,p, set $\sigma_r = (\sigma_r(l))_{l=1}^{k(g)+1}$ with $\sigma_r(l) \in \{0,1,2\}$. Then we have for all $\mathbf{v} \in \mathcal{L}$*

$$N^{-3/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\left|\mathbb{E}\left((\mathbf{G}_{\tilde{i}\tilde{i}})^{d_i}\mathbf{v}(\tilde{i})^{d_{\mathbf{v}}}\prod_{r=1}^{p}\mathbf{A}_{\mathbf{v},i,\mu}(g_r,\sigma_r)\mathcal{T}_N(\mathbf{X})\right)\right| = O((N^{C\delta}\Psi)^p + \mathbb{E}\|\mathbf{L}_p(\mathbf{X})\|_\infty), \quad (8.65)$$

where

$$d_\blacktriangle = \sum_{r=1}^{p} |\sigma_r|_\blacktriangle \quad (8.66)$$

for $\blacktriangle = i, \mathbf{v}$.

Before proving Lemma 11, we give a rough bound of $\mathbf{A}$ first. This bound helps us to connect the left hand side of (8.65) with the desired bound $\mathbf{L}_p(\mathbf{X})$.

Lemma 12. *[Rough bounds on $\mathbf{A}_{\mathbf{v},i,\mu}(g,\sigma)$.] Assume that (A_{m-1}) holds. Then for $z \in S_m$, we have*

$$|\mathbf{A}_{\mathbf{v},i,\mu}(g,\sigma)| \prec N^{2\delta(k(g)+1)}. \quad (8.67)$$

If $k(\omega) = 0$, then

$$|\mathbf{A}_{\mathbf{v},i,\mu}(g,\sigma)| \prec N^{(C_0/2+1)\delta}\Psi + F_{\mathbf{v}}(X) + F_{\mathbf{e}_{\tilde{i}}}(X). \quad (8.68)$$

Proof of Lemma 12. By the arguments similar to Lemma 7, it is easy to get the following bound for $z \in S_m$ given A_{m-1}:

$$(\mathbf{G} - \Pi)_{\mathbf{vv}} = O_\prec(N^{2\delta}),\ \Im\mathbf{G}_{\mathbf{vv}} \prec N^{2\delta}(\Im m + N^{C_0\delta}\Psi). \quad (8.69)$$

The remaining argument is similar to that for Lemma 8.9 in [16] and we ignore details here. □

Noticing that $\mathbf{G}_{\tilde{i}\tilde{i}}$ in the left hand side of (8.65) contains $X_{i\mu}, \mu = 1, ..., N$, we need to extract $X_{i\mu}$ from it. To this end, we use the following resolvent expansion for the diagonal entry of $\mathbf{G}$

$$\mathbf{G}_{\tilde{i}\tilde{i}} = -(Y_i + Z_i)^{-1},$$

where

$$Y_i = \mathbb{E}[(\mathbf{X}\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{X}^T)_{ii}|X^{(i)}] = \frac{1}{N}\sum_j(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{jj}, \quad Z_i = (\mathbf{X}\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{X}^T)_{ii} - Y_i.$$

Using the large deviation bound, we find out that $|Z_i| \prec N^{-\tau/2+2\delta}$. This, together with Taylor's expansion, implies that there exists a constant $K = K(\tau)$ such that

$$\mathbf{G}_{\tilde{i}\tilde{i}} = \sum_{k=0}^{K}\frac{Z_i^k}{k!(-Y_i)^{k+1}} + O_\prec(N^{-10}). \tag{8.70}$$

Expanding further Z_i, we have

$$\mathbf{G}_{\tilde{i}\tilde{i}} = \sum_{k=0}^{K} Y_{i,k}(\mathbf{X}\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{X}^T)_{ii}^k + O_\prec(N^{-10}), \tag{8.71}$$

where the coefficients Y_{ik} including $1/Y_i$ and Y_i are $\mathbf{X}^{(i)}$ measurable. In order to apply Lemma 4 an upper bound of $|Y_{i,k}|$ is needed. From (8.70) and the definition of Z_i one can see that $Y_{i,k}$ is a finite order polynomial function of $(Y_i)^{-1}$ and Y_i, Therefore it suffices to develop upper and lower bounds of Y_i. Recalling the definition of $\mathbf{\Delta}$ at (8.31), we have

$$\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}, \ \mathbf{U}_2^T\mathbf{U}_2 = \mathbf{I}, \ \text{and } \mathbf{\Delta}\mathbf{\Delta}^T = \begin{pmatrix} \mathbf{I} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mathbf{I} \end{pmatrix}.$$

From (8.9) we have

$$|Y_i| \geq \Im Y_i = \frac{1}{N}tr\Im(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{\Delta}^T) = \frac{1}{N}tr\mathbf{A}_2^{(\tilde{i})}\Im\mathbf{G}_{N_1}^{(\tilde{i})}(\mathbf{A}_2^{(\tilde{i})})^T \geq \frac{1}{N}tr\Im\mathbf{G}_{N_1}^{(\tilde{i})} \geq C\eta,$$

where $\mathbf{A}_j^{(\tilde{i})}$ are respectively obtained from $\mathbf{A}_j, j = 2, 3$ by deleting the i-th row of $\mathbf{X}$. On the other hand we conclude from (8.9) that

$$|Y_i| \leq \|\mathbf{A}_3^{(\tilde{i})}\| + \frac{\|\mathbf{A}_2^{(\tilde{i})}\|^2}{\eta}.$$

These, together with Lemma 2 and an estimate similar to (8.16), implies that there exists a constant c such that

$$N^{-c} \leq |Y_i\mathcal{T}_N(\mathbf{X}^{(i)})| \leq N^c. \tag{8.72}$$

We are now in a position to replace $\mathbf{G}_{\tilde{i}\tilde{i}}$ by $\sum_{k=0}^K \mathcal{Y}_{i,k}(\mathbf{X}\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{X}^T)_{ii}^k$ because

$$\mathbb{E}\prod_{r=1}^{p}|\mathbf{A}_{\mathbf{v},i,\mu}(g_r,\sigma_r)|\left|\mathbf{G}_{ii}^{d_i} - (\sum_{k=0}^{K}\mathcal{Y}_{i,k}(\mathbf{X}\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{X}^T)_{ii}^k)^{d_i}\mathcal{T}_N(\mathbf{X}^{(i)}))\right| \prec N^{-10+2\delta d_i}\mathbb{E}\prod_{r=1}^{p}|\mathbf{A}_{\mathbf{v},i,\mu}(g_r,\sigma_r)|$$
$$\prec N^{-5}((N^{C\delta}\Psi)^p + \mathbf{L}_p(\mathbf{X})), \tag{8.73}$$

where we apply Lemma 12, Lemma 4, (8.71), (8.72) and the fact that there are at most three r such that $k(g_r) \geq 1$. In view of (8.73) proving Lemma 11 reduces to showing

$$N^{-3/2} \sum_{i=1}^{M_1} \sum_{\mu=1}^{N} \left| \mathbb{E} \left((\sum_{k=0}^{K} Y_{i,k} (\mathbf{X} \mathbf{\Delta}^T \mathbf{G}^{(\tilde{i})} \mathbf{\Delta} \mathbf{X}^T)_{ii}^k)^{d_i} \mathbf{v}(i)^{d_{\mathbf{v}}} \prod_{r=1}^{p} \mathbf{A}_{\mathbf{v},i,\mu}(g_r, \sigma_r) \mathcal{T}_N(\mathbf{X}^{(i)}) \right) \right| \tag{8.74}$$
$$= O((N^{C\delta}\Psi)^p + \mathbb{E}\|\mathbf{L}_p(\mathbf{X})\|_\infty).$$

We further expand

$$(\sum_{k=0}^{K} Y_{i,k} (\mathbf{X} \mathbf{\Delta}^T \mathbf{G}^{(\tilde{i})} \mathbf{\Delta} \mathbf{X}^T)_{ii}^k)^{d_i} = \sum_{k=0}^{Kd_i} \mathcal{Y}_{i,k} (\mathbf{X} \mathbf{\Delta}^T \mathbf{G}^{(\tilde{i})} \mathbf{\Delta} \mathbf{X}^T)_{ii}^k,$$

where the coefficient $\mathcal{Y}_{i,k}$ is $\mathbf{X}^{(i)}$ measurable and bounded by

$$|\mathcal{Y}_{i,k}| \prec N^{Cd_i\delta}.$$

Here we don't need the explicit expression of $\mathcal{Y}_{i,k}$ and its upper bound is enough by checking the following arguments carefully. For any tagged group (g, σ), we write

$$\mathbf{A}_{\mathbf{v},i,\mu}(g, \sigma) = \mathbf{A}^-_{\mathbf{v},i,\mu}(g, \sigma) \mathbf{A}^+_{\mathbf{v},i,\mu}(g, \sigma),$$

where $\mathbf{A}^-_{\mathbf{v},i,\mu}(g, \sigma)$ is measurable with respect to $\mathbf{X}^i$ and $\mathbf{A}^+_{\mathbf{v},i,\mu}(g, \sigma)$ is a product of the terms

$$(\mathbf{G}^{(\tilde{i})} \mathbf{\Delta} \mathbf{X}^T)_{\mathbf{x}i}, \text{ or } (\mathbf{X} \mathbf{\Delta}^T \mathbf{G}^{(\tilde{i})})_{i\mathbf{x}}, \ \mathbf{x} \in \{\mathbf{v}, \mathbf{\Delta} \mathbf{e}_\mu\}.$$

So the left hand side of (8.74) is bounded by

$$N^{-3/2} N^{Cd_i\delta} \sum_{i=1}^{M_1} \sum_{\mu=1}^{N} \mathbf{v}(\tilde{i})^{d_{\mathbf{v}}} E\Big[\Big|\mathbb{E} \prod_{r=1}^{p} \mathbf{A}^-_{\mathbf{v},i,\mu}(g, \sigma)\Big| \Big|\mathbb{E}_i \prod_{r=1}^{p} \mathbf{A}^+_{\mathbf{v},i,\mu}(g_r, \sigma_r) (\mathbf{X} \mathbf{\Delta}^T \mathbf{G}^{(\tilde{i})} \mathbf{\Delta} \mathbf{X}^T)_{ii}^k \mathcal{T}_N(\mathbf{X}^{(i)})\Big|\Big] \tag{8.75}$$

where $\mathbb{E}_i$ stands for taking expectation over the random variables at the i-th row of $\mathbf{X}$ (conditional expectation).

We below first show that for the inner conditional expectation and $k \leq Kd_i$,

$$\left| \mathbb{E}_i \prod_{r=1}^{p} \mathbf{A}^+_{\mathbf{v},i,\mu}(g_r, \sigma_r) (\mathbf{X} \mathbf{\Delta}^T \mathbf{G}^{(\tilde{i})} \mathbf{\Delta} \mathbf{X}^T)_{ii}^k \mathcal{T}_N(\mathbf{X}^{(i)}) \right| \prec N^{-\mathbf{1}(d_{\mathbf{v}}=0)/2} (N^{(C_0/2+C)\delta} \Psi)^{d_{\mathbf{x}} - \mathbf{1}(d_{\mathbf{x}} \geq 3)}. \tag{8.76}$$

The proof of (8.76) is similar to that of Lemma 8.11 in [16] and the transitional arguments from (8.76) to (8.74) are the same as those from (8.32) in [16] to the end of section 8.5 in [16]. We below only list some difference involved in our derivatives when proving (8.76). Define

$$d_{\mathbf{x}} = \sum_{r=1}^{p} \deg(\mathbf{A}^+_{\mathbf{v},i,\mu}(g_r, \sigma_r)) = \sum_{r=1}^{p} \deg(\mathbf{A}_{\mathbf{v},i,\mu}(g_r, \sigma_r)), \tag{8.77}$$

where $\deg(\mathbf{A}_{\mathbf{v},i,\mu}(g_r,\sigma_r))$ stands for the degree of the polynomial $\mathbf{A}_{\mathbf{v},i,\mu}(g_r,\sigma_r)$ in terms of $X_{i\mu}$. We abbreviate $d_{\mathbf{x}}$ by d. Recalling the definition of $\mathbf{A}^+$, we have

$$\prod_{r=1}^{p}\mathbf{A}^+_{\mathbf{v},i,\mu}(g_r,\sigma_r)(\mathbf{X}\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta}\mathbf{X}^T)^k_{ii}=\sum_{j_1,\ldots,j_{d+2k}=1}^{N}\mathcal{G}_{j_1,\ldots,j_d}\tilde{\mathcal{G}}_{j_{d+1},\ldots,j_{d+2k}}\prod_{l=1}^{d+2k}X_{ij_l},$$

where $\mathcal{G}_{j_1,\ldots,j_d}$ is a product of d terms in the set $\{(\boldsymbol{\Delta}\mathbf{G}^{(\tilde{i})})_{j_l\mathbf{v}},(\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta}^T)_{\mathbf{v}j_l},(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta})_{j_l\mu},(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta})_{\mu j_l}$, $l=1,2,..,d.\}$ and $\mathcal{G}_{j_{d+1},\ldots,j_{d+2k}}$ is a product of k terms in the set $\{(\boldsymbol{\Delta}\mathbf{G}^{(\mu)}\boldsymbol{\Delta}^T)_{j_lj_{l'}},j_l,j_{l'}=d+1,d+2,..,d+2k.\}$. Since $\mathbb{E}_\mu X_{ij}=0$, it is easy to see that the conditional expectation on $\prod_{l=1}^{d+2k}X_{ij_l}$ is nonzero only if each index appears at least twice. The set $\{1,...,d+2k\}$ can be reorganized by several blocks such that each block contain the same indices j_l. For example, if b_1 and b_2 are two different blocks, then the indexes belonging to b_1 (and b_2) are all equal and any two indexes $a_1\in b_1$, $a_2\in b_2$ are not equal. For a block $b\subset\{1,...,d+2k\}$, we define $d_b=|b\cap\{1,...,d\}|$ and $k_b=|b\cap\{d+1,...,d+2k\}|$. Here d_b means the number of indices equal to b from $\{1,\cdots,d\}$ and k_b means the number of indices equal to b from $\{d+1,\cdots,d+k\}$. Moreover, we suppose there are L blocks. Hence, we reorganize the summation as follows:

$$\begin{aligned}&|\mathbb{E}_i\prod_{r=1}^{p}\mathbf{A}^+_{\mathbf{v},i,\mu}(g_r,\sigma_r)(\mathbf{X}\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta}\mathbf{X}^T)^k_{ii}|\prec C_qN^{2\delta k}\max_{L}\max_{\{d_l\}}\max_{\{k_l\}}\sum_{j_1,\ldots,j_L}\times\\&\prod_{l=1}^{L}(\mathbb{E}|\mathbf{X}_{ij_l}|^{d_l+k_l}(|(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j_l\mathbf{v}}|+|(\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta})_{\mathbf{v}j_l}|+|(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta})_{j_l\mu}|+|(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta})_{\mu j_l}|)^{d_l}),\end{aligned}\tag{8.78}$$

where d_l and k_l satisfy

$$\sum_{l=1}^{L}d_l=d,\ \sum_{l=1}^{L}k_l=2k,\ \ d_l+k_l\geq 2.\tag{8.79}$$

It is straightforward to see that the right hand side of (8.78) is bounded by

$$C_qN^{2\delta k}N^{-d/2-k}\prod_{l=1}^{L}\Big(\sum_{j}\big(|(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j\mathbf{v}}|+|(\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta})_{\mathbf{v}j}|+|(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta})_{j\mu}|+|(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta})_{\mu j}|\big)^{d_l}\Big).\tag{8.80}$$

The upper bound of the above term is

$$\sum_{j}(|(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j\mathbf{v}}|+|(\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta})_{\mathbf{v}j}|+|(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta})_{j\mu}|+|(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta})_{\mu j}|)^{d_l}\prec N(N^{(C_0/2+1)\delta}\Psi)^{d_l\wedge 2}N^{2\delta[d_l-2]_+},$$

following from $\sum_j|(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j\mathbf{v}}|^2\lesssim\frac{\Im G^{(\tilde{i})}_{\mathbf{v}\mathbf{v}}}{\eta}$, $\sum_j|(\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta})_{j\mu}|^2\lesssim\frac{\Im G^{(\tilde{i})}_{\mathbf{v}_\mu\mathbf{v}_\mu}}{\eta}$. Therefore we have

$$|\mathbb{E}_\mu\prod_{r=1}^{p}\mathbf{A}^+_{v,i,\mu}(g_r,\sigma_r)(\mathbf{X}\boldsymbol{\Delta}^T\mathbf{G}^{(\tilde{i})}\boldsymbol{\Delta}\mathbf{X}^T)^k_{ii}|\prec\max_{L}\max_{\{d_l\}}\max_{\{k_l\}}N^{-d/2-k+L}(N^{(C_0/2+1)\delta}\Psi)^{\sum_l(d_l\wedge 2)}N^{2\delta\sum_l[d_l-2]_+}.$$

The remaining arguments are the same as those below (8.29) in [16] (to the end of Section 8.5 in [16]).

We next consider the derivative of $\mathbf{\Pi}(z)$ since we have only considered the derivative of $\mathbf{G}(z)$ for $k(g) \geq 1$(one can refer to the definition of $\mathbf{A}_{\mathbf{v},i,\mu}(g,\sigma)$). That is to say, we aim at proving (8.48) for $k = 3$ but we only proved

$$N^{-3/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\left[\prod_{r=1}^{q} C_{s,t,i,\mu}(g_r)\prod_{r=q+1}^{p} B_{s,t,i,\mu}(g_r)\mathcal{T}_N^p(\mathbf{X})\right] = O((N^{24\delta}\Psi)^p + \|\mathbb{E}\mathbf{L}_p(\mathbf{X})\|_\infty), \quad (8.81)$$

for $q \leq 3$, $\sum_{r=1}^{q} k(g_r) = 3$, $k(g_r) \geq 1$, $r \leq q$ and $g_r = 0$, $r \geq q+1$, where the definitions of $C_{s,t,i,\mu}$ and $B_{s,t,i,\mu}$ are given at (8.45) and (8.46). One may refer to (8.65) for (8.81) (note that $C_{s,t,i,\mu}(g_r)$ has been decomposed as the sum of the terms $\mathbf{A}_{\mathbf{v},i,\mu}(g,\sigma)$ and we in fact extract the i-th row of $\mathbf{X}$ from C and then get $\mathbf{A}_{\mathbf{v},i,\mu}(g,\sigma)$). This means that we have not considered the second term in (8.42). Recalling (8.42), the k-th derivative of each $\mathbf{\Pi}(z)$ can be written as

$$D_{s,t,i,\mu}(g_r) = \sum_{j=1}^{3}(\mathbf{B}_1\mathcal{R}_{j1})_{(s_{j1}t_{j1})}(\mathcal{R}_{j2})_{(s_{j2}t_{j2})}...(\mathcal{R}_{jk})_{(s_{jk}t_{jk})}(\mathcal{R}_{jk+1}\mathbf{B}_2^T)_{(s_{jk+1}t_{jk+1})},\ k(g_r) = k. \quad (8.82)$$

Therefore, in order to prove (8.48) for $k = 3$ what remains is to show that

$$N^{-3/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\left[\prod_{r=1}^{l} C_{s,t,i,\mu}(g_r)\prod_{r=l+1}^{q} D_{s,t,i,\mu}(g_r)\prod_{r=q+1}^{p} B_{s,t,i,\mu}(g_r)\mathcal{T}_N^p(\mathbf{X})\right] = O((N^{24\delta}\Psi)^p + \|\mathbb{E}\mathbf{L}_p(\mathbf{X})\|_\infty), \quad (8.83)$$

where $q \leq 3$, $\sum_{r=1}^{q} k(g_r) = 3$, $k(g_r) \geq 1$, $r \leq q$ and $g_r = 0$, $r \geq q+1$.

We first consider the case when $l = 0$, which implies that there is no $C_{s,t,i,\mu}(g_r)$ for $k(g_r) \geq 1$. We start with the case when $l = 0, q = 3$, which corresponds to $k(g_j) = 1, j = 1,2,3$ due to $\sum_{r=1}^{q} k(g_r) = 3$ with $k(g_r) \geq 1$. As a sequence, each summand in $D_{s,t,i,\mu}(g_r)$ becomes $(\mathbf{B}_1\mathcal{R}_{j1})_{(s_{j1}t_{j1})}(\mathcal{R}_{j2}\mathbf{B}_2^T)_{(s_{j2}t_{j2})}$. Recalling the definition of $\mathcal{H}_{1i}$, $\mathcal{H}_{1\mu}$, $\mathcal{H}_{sti}$ and $\mathcal{H}_{st\mu}$ above (8.49), the left hand side of (8.83) can be bounded by

$$\begin{aligned} &\Big|N^{-3/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\left[\prod_{r=1}^{3} D_{s,t,i,\mu}(g_r)\prod_{r=4}^{p} B_{s,t,i,\mu}(g_r)\mathcal{T}_N^p(\mathbf{X})\right]|\Big| \quad (8.84)\\ = \ &N^{-3/2}\Big|\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\left[\prod_{r=1}^{3} D_{s,t,i,\mu}(g_r)F_{st}^{p-3}\mathcal{T}_N^p(\mathbf{X})\right]\Big| \leq 3^3N^{-3/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\Big|\mathbb{E}\left[\mathcal{H}_{sti}^3\mathcal{H}_{st\mu}^3F_{st}^{p-3}\mathcal{T}_N^p(\mathbf{X})\right]\Big| \end{aligned}$$

Recalling (8.53), we have

$$\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathcal{H}_{sti}^3\mathcal{H}_{st\mu}^3 \prec 1.$$

Therefore the right hand side of (8.84) can be bounded by

$$N^{-3/2}\Big|\mathbb{E}\left[F_{st}^{p-3}\mathcal{T}_N^p(\mathbf{X})\right]\Big| = O((N^{24\delta}\Psi)^p + \|\mathbb{E}\mathbf{L}_p(\mathbf{X})\|_\infty),$$

using the fact that $N^{-1/2} \lesssim \Psi$. This ensures that (8.83) holds for $l = 0$ and $q = 3$.

We next consider the case when $l = 0, q = 2$, which forces $k(g_1) = 1, k(g_2) = 2$ due to $\sum_{r=1}^{q} k(g_r) = 3$ with $k(g_r) \geq 1$. Similar to (8.84) we have the upper bound

$$\begin{aligned} &N^{-3/2}\Big|\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\left[\prod_{r=1}^{2}D_{s,t,i,\mu}(g_r)\prod_{r=3}^{p}B_{s,t,i,\mu}(g_r)\mathcal{T}_N^p(\mathbf{X})\right]\Big| = N^{-3/2}\Big|\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\left[\prod_{r=1}^{2}D_{s,t,i,\mu}(g_r)F_{st}^{p-2}\mathcal{T}_N^p(\mathbf{X})\right]\Big| \\ &\leq 3^2N^{-3/2}\Big|\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\left[(\mathcal{H}_{sti}^3\mathcal{H}_{st\mu}+\mathcal{H}_{sti}\mathcal{H}_{st\mu}^3+\mathcal{H}_{sti}^2\mathcal{H}_{st\mu}^2)F_{st}^{p-2}\mathcal{T}_N^p(\mathbf{X})\right]\Big| \\ &\prec \mathbb{E}\left[(\Psi^3+\Psi^2)F_{st}^{p-2}\mathcal{T}_N^p(\mathbf{X})\right] = O((N^{24\delta}\Psi)^p + \|\mathbb{E}\mathbf{L}_p(\mathbf{X})\|_\infty), \end{aligned} \tag{8.85}$$

where we apply the Cauchy-Schwarz inequality together with (8.53) such that $\sum_{i=1}^{M_1}\mathcal{H}_{sti} \prec \sqrt{N}$ and $\sum_{\mu\in I_N}\mathcal{H}_{st\mu} \prec \sqrt{N}$.

As for $l = 0, q = 1$ we have $k(g_1) = 3$ due to $\sum_{r=1}^{q} k(g_r) = 3$ with $k(g_r) \geq 1$. We also have a similar upper bound

$$\begin{aligned} &N^{-3/2}\Big|\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\left[D_{s,t,i,\mu}(g_1)\prod_{r=2}^{p}B_{s,t,i,\mu}(g_r)\mathcal{T}_N^p(\mathbf{X})\right]\Big| = N^{-3/2}\Big|\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\left[D_{s,t,i,\mu}(g_1)F_{st}^{p-1}\mathcal{T}_N^p(\mathbf{X})\right]\Big| \\ &\leq 3N^{-3/2}\Big|\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\left[(\mathcal{H}_{sti}^2+\mathcal{H}_{st\mu}^2)F_{st}^{p-1}\mathcal{T}_N^p(\mathbf{X})\right]\Big|, \\ &\prec \mathbb{E}\left[\Psi F_{st}^{p-1}\mathcal{T}_N^p(\mathbf{X})\right] = O((N^{24\delta}\Psi)^p + \|\mathbb{E}\mathbf{L}_p(\mathbf{X})\|_\infty). \end{aligned} \tag{8.86}$$

Therefore (8.83) holds when $l = 0$.

When $l \neq 0$ in (8.83) we below consider the case when $l = 1, q = 2$ only and the other cases can be handled similarly. In this case we need to show that

$$N^{-3/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\left[C_{s,t,i,\mu}(g_1)D_{s,t,i,\mu}(g_2)\prod_{r=3}^{p}B_{s,t,i,\mu}(g_r)\mathcal{T}_N^p(\mathbf{X})\right] = O((N^{24\delta}\Psi)^p + \|\mathbb{E}\mathbf{L}_p(\mathbf{X})\|_\infty), \tag{8.87}$$

where $k(g_1) = 2, k(g_2) = 1$ or $k(g_1) = 1, k(g_2) = 2$. Similar to the arguments above, the left hand side of (8.87) can be bounded by

$$\Big|F_{st}^{p-2}N^{-3/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}(\mathcal{H}_{1i}^2\mathcal{H}_{sti}\mathcal{H}_{st\mu}+\mathcal{H}_{1\mu}^2\mathcal{H}_{sti}\mathcal{H}_{st\mu}+\mathcal{H}_{1i}\mathcal{H}_{1\mu}\mathcal{H}_{sti}\mathcal{H}_{st\mu})\Big|,$$

which is bounded by $|F_{st}^{p-2}\Psi^2|$ by the inequality

$$\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathcal{H}_{1i}^2\mathcal{H}_{sti}\mathcal{H}_{st\mu} \prec \sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathcal{H}_{1i}^2\mathcal{H}_{st\mu} \prec N^{3/2}\Psi^2.$$

This ensures (8.87).

9 Local law in average (7.8)

The purpose of this subsection is to prove the following ((7.8) in Theorem 7.1)

$$|m_N(z) - m(z)| \prec \frac{1}{N\eta}. \tag{9.1}$$

As pointed out in the paragraph below (8.19), (7.8) holds when the entries $\{X_{ij}\}$ of $\mathbf{X}$ are the standard Gaussian random variable. We next use the interpolation method to prove (7.8) for the general distributions as in proving (7.7). However we do not need induction on the imaginary part of z unlike before due to existence of (7.7). In order to prove (9.1), it is enough to prove that

$$|m_N(z) - m(z)|\mathcal{T}_N(\mathbf{X}) \prec \frac{1}{N\eta}. \tag{9.2}$$

We introduce the notation $\tilde{F}(X,z)$ as in the last section

$$\tilde{F}(X,z) = |m_N(z) - m(z)|\mathcal{T}_N(\mathbf{X}) = |\frac{1}{N_1}\sum_{k=1}^{N_1} G_{kk}(z) - m(z)|\mathcal{T}_N(\mathbf{X}).$$

Checking on Lemmas 3, 5, 6, (8.29) and (8.38) in the last section it suffices to show

$$N^{-k/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\Big[(\frac{\partial}{\partial \mathbf{X}_{i\mu}})^k \tilde{F}^p(\mathbf{X},z)\Big] = O((N^\delta\Psi^2)^{2q} + \|\tilde{F}^p(\mathbf{X},z)\|_\infty),\ k \geq 3 \tag{9.3}$$

where δ is a sufficiently small constant such that N^δ is much smaller than N^ε before (9.2) due to the definition of the partial order. Applying the definition of $B_{s,t,i,\mu}$ in the preceding section with $\mathbf{B}_1 = \mathbf{B}_2 = 1$ and $s = t = k$, it suffices to prove that for $\sum_r k(g_r) = k$

$$N^{-k/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}\Big(\prod_{r=1}^{p}\Big[\frac{1}{N_1}\sum_{k=1}^{N_1}B_{k,k,i,\mu}(g(r))\Big]\mathcal{T}_N(\mathbf{X})\Big) = O((N^\delta\Psi^2)^p + \|\mathbb{E}\tilde{F}^p(X)\|_\infty), k \geq 3. \tag{9.4}$$

One can verify (9.4) for $k \geq 4$ by repeating the same arguments as in (7.95)-(7.97) in [12]. The key steps are the following two inequalities:

$$\frac{1}{N_1}\sum_{k=1}^{N_1}B_{k,k,i,\mu}(g_h) \prec \Psi^2, \quad \text{for} \quad n(g_h) \geq 1, \tag{9.5}$$

and

$$\frac{1}{N_1}\sum_{k=1}^{N_1}C_{k,k,i,\mu}(g_h) \prec \Psi^2, \quad \text{for} \quad n(g_h) \geq 1. \tag{9.6}$$

Consider (9.4) for $k = 3$ now. To this end, as in (8.65) and (9.4), it suffices to prove that

$$N^{-3/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}|N_1^{-p}\sum_{v_1,\ldots,v_p=1}^{N_1}\mathbb{E}\Big(\prod_{r=1}^{p}\mathbf{A}_{\mathbf{e}_{v_r},i,\mu}(g_r,\sigma_r)G_{\tilde{i}\tilde{i}}^{d_i}\mathcal{T}_N(\mathbf{X})\Big)| = O_\prec((N^\delta\Psi^2)^p + \mathbb{E}\|\tilde{F}^p(X)\|_\infty), \tag{9.7}$$

where $\mathbf{e}_{v_r}$ is an $(N+M_1+N_1-M_2)$-dimensional unit vector with the v_r-th element being 1 and $1 \le v_r \le N_1$ (here the size of $\mathbf{e}_{v_r}$ is the same as the size of the matrix $\mathbf{G}(z)$). One should notice that we don't consider the derivatives of $\mathbf{\Pi}(z)$ any more in this subsection since we only care about the upper left $N_1 \times N_1$ block matrix of $\mathbf{\Pi}(z)$ for the purpose of proving (9.1), which is $m(z)\mathbf{I}$ (see (7.6)). Hence it suffices to consider the derivative on $\mathbf{G}(z)$ and apply its expansion (8.54). Moreover, in this case, one can see that $d_{\mathbf{v}} = 0$ since $\mathbf{e}_{v_k}(\tilde{i}) = 0, k = 1, ..., p$ recalling $\tilde{i} = i + N_1$. Hence there is no factor $\mathbf{v}(\tilde{i})^{d_v}$ in (9.7) unlike (8.65). As in (8.74) and (8.75), it then suffices to prove that

$$N^{-3/2}N^{d_i\delta}\textstyle\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}|N_1^{-p}\sum_{v_1,...,v_p=1}^{N_1}\mathbb{E}\Big[\prod_{r=1}^{p}\mathbf{A}^{-}(r)\mathbb{E}_i\Big(\prod_{r=1}^{p}\mathbf{A}^{+}(r)(\mathbf{X}\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{X}^T)_{ii}^k\mathcal{T}_N(\mathbf{X})\Big)|\Big]$$
$$= O_{\prec}((N^{\delta}\Psi^2)^p + \mathbb{E}\|\tilde{F}^p(X)\|_{\infty}), \tag{9.8}$$

for $k \le Cd_i$, where $\sum_r k(g_r) = 3$ and $\mathbf{A}^{\cdot}(r) = \mathbf{A}^{\cdot}_{\mathbf{e}_{v_r},i,\mu}(g_r, \sigma_r)$ with $\cdot = -, +$. Here each factor $\mathbf{A}^{+}(r)$ is a product of factors $(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}X^T)_{\mathbf{s}i}$ and $(X\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{i\mathbf{s}}$, $\mathbf{s} \in \{\mathbf{e}_{v_r}, \mathbf{\Delta}e_{\mu}\}$. We denote the number of factors $(\Delta^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}X^T)_{\mu i}$ and $(X\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\Delta)_{i\mu}$ by $d_{\mathbf{x},\mu,r}$(i.e. $\mathbf{s} = \Delta\mathbf{e}_{\mu}$) contained in $\mathbf{A}^{+}(r)$ and write $d_{\mathbf{x},\mu} = \sum\limits_{r=1}^{p} d_{\mathbf{x},\mu,r}$. By (7.7) and Definition 6 , it is easy to conclude that

$$\frac{1}{N_1}\sum_{v_r=1}^{N_1}|(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{v_r l}(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{v_r j}| \prec \frac{\Im(\mathbf{G}^{(\tilde{i})})_{jl}}{N\eta} \prec \Psi^2. \tag{9.9}$$

Consider $\mathbb{E}_i\Big(\prod_{r=1}^{p}\mathbf{A}^{+}(r)(\mathbf{X}\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{X}^T)_{ii}^k\Big)$. As in (8.78)-(8.80) we obtain

$$|\mathbb{E}_i\prod_{r=1}^{p}\mathbf{A}^{+}(r)(\mathbf{X}\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}\mathbf{X}^T)_{ii}^k| \prec \max_{L}\max_{\{d_l\}}\max_{\{k_l\}}\sum_{j_1,...,j_L}\times\prod_{l=1}^{L}\Bigg(N^{-d/2-k}(|(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{j_l\mu}|$$
$$+|(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{\mu j_l}|)^{d_l-\sum_{r=1}^{p}d_{l,r}}\prod_{r=1}^{p}\Big[(|(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j_l\mathbf{e}_{v_r}}| + |(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{\mathbf{e}_{v_r}j_l}|)^{d_{l,r}}\Big]\Bigg), \tag{9.10}$$

where $d_{l,r}$ denotes the number of the factors $(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j_l\mathbf{e}_{v_r}}$ and $(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{\mathbf{e}_{v_r}j_l}$ (essentially, it is the number of the factors $(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}X^T)_{\mathbf{s}i}$ and $(X\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{i\mathbf{s}}$ with $\mathbf{s} = \mathbf{e}_{v_r}$ in $\mathbf{A}^{+}(r)$). By (8.62)-(8.63), it is easy to see that $\sum_l d_{l,r} \le 2$ in $A^{+}(r)$. We have to combine $\prod_{l=1}^{L}\Big[(|(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j_l\mathbf{e}_{v_r}}| + |(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{\mathbf{e}_{v_r}j_l}|)^{d_{l,r}}\Big]$ with $A^{-}(r)$ together so that we may use (9.9). Hence we below consider the upper bound of

$$N_1^{-p}\sum_{v_1,...,v_p=1}^{N_1}\prod_{r=1}^{p}\mathbf{A}^{-}(r)\prod_{l=1}^{L}\Big[(|(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j_l\mathbf{e}_{v_r}}| + |(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{\mathbf{e}_{v_r}j_l}|)^{d_{l,\mu,r}}\Big] \tag{9.11}$$

first. One should notice that the above summation and product are only about v_r, which are independent of l.

For $r \ge q+1$ satisfing $\sigma_r = 0$, recalling $A_{\mathbf{v},i,\mu}(g, 0)$ in Definition 6, we have

$$\frac{1}{N_1}\sum_{v_r=1}^{N_1}\mathbf{A}^{-}(r) = \tilde{F}(X, z) + O_{\prec}(\Psi^2),$$

where the $O_\prec(\Psi^2)$ follows from the fact that $\frac{1}{N_1}\sum_{v_r=1}^{N_1}\mathbf{A}^-(r)=m_N(z)-m(z)-\frac{1}{N_1}\sum_{v_r=1}^{N_1}\mathbf{A}_{\mathbf{e}_{v_r},i,\mu}(g_r,1)$ and using the large deviation inequality and (9.9) to control $\frac{1}{N_1}\sum_{v_r=1}^{N_1}\mathbf{A}_{\mathbf{e}_{v_r},i,\mu}(g_r,1)$. For the remaining $\sum_{r=q+1}^{p}|\sigma_r|$ indices, we always have the trivial order $\mathbf{A}^-(r) = 1$ by the fact that $\mathbf{A}(r)=\mathbf{A}^+(r)$. When $\sigma_r=1$, $r\geq q+1$, by the expansion of Definition 6(i), we have $\sum_{l=1}^{L}d_{l,r}=2$. Thus by (9.9) we have for $\sigma_r=1$, $r\geq q+1$,

$$\frac{1}{N_1}\sum_{v_r=1}^{N_1}\prod_{l=1}^{L}\Big[(|(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j_l\mathbf{e}_{v_r}}|+|(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{\mathbf{e}_{v_r}j_l}|)^{d_{l,\mu,r}}\Big]\prec\Psi^2. \tag{9.12}$$

Furthermore, consider $r\leq q$. If there are two indices v_r (associated with $\mathbf{e}_{v_r}$ in the factors $(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta}X^T)_{\mathbf{s}i}$ and $(X\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{i\mathbf{s}}$, $\mathbf{s}\in\{\mathbf{e}_{v_r},\mathbf{\Delta}e_\mu\}$) appearing in $\mathbf{A}^-(r)$, by (9.9) then we have

$$\frac{1}{N_1}\sum_{v_r=1}^{N_1}\mathbf{A}^-(r)\prec\Psi^2.$$

If there is no index v_r appearing in $\mathbf{A}^-(r)$, then we use the bound

$$\mathbf{A}^-(r)\prec 1.$$

In this case($r\leq q$, no v_r appears in $\mathbf{A}^-(r)$) two indices v_r both appear in $\mathbf{A}^+(r)$ and hence we combine them with $\mathbf{A}^-(r)$, as in (9.11). Hence we also have as in (9.12)

$$\frac{1}{N_1}\sum_{v_r=1}^{N_1}\mathbf{A}^-(r)\prod_{l=1}^{L}\Big[(|(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j_l\mathbf{e}_{v_r}}|+|(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{\mathbf{e}_{v_r}j_l}|)^{d_{l,\mu,r}}\Big]\prec\Psi^2.$$

If there is only one v_r appearing in $\mathbf{A}^-(r)$, by Definition 6(ii) we have $\sum_{l=1}^{L}d_{l,r}=1$. Hence one v_r appears in $\mathbf{A}^+(r)$ and we combine such a term involving v_r in $\mathbf{A}^+(r)$ with $\mathbf{A}^-(r)$, as in (9.11). Therefore by (9.9) we conclude that (9.12) holds. Therefore, summarizing above arguments for $r\leq q$, we have for $r\leq q$

$$\frac{1}{N_1}\sum_{v_r=1}^{N_1}\mathbf{A}^-(r)\prod_{l=1}^{L}\Big[(|(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j_l\mathbf{e}_{v_r}}|+|(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{\mathbf{e}_{v_r}j_l}|)^{d_{l,r}}\Big]\prec\Psi^2.$$

Furthermore, together with the arguments for $r\geq q+1$, we have

$$\begin{aligned}&N_1^{-p}\sum_{v_1,\dots,v_p=1}^{N_1}\prod_{r=1}^{p}\mathbf{A}^-(r)\prod_{l=1}^{L}\Big[(|(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j_l\mathbf{e}_{v_r}}|+|(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{\mathbf{e}_{v_r}j_l}|)^{d_{l,\mu,r}}\Big]\\&\prec(\tilde{F}(X,z)+O_\prec(\Psi^2))^{p-q-\sum_{r=q+1}^{p}|\sigma_r|}\Psi^{2(q+\sum_{r=q+1}^{p}|\sigma_r|)}.\end{aligned}\tag{9.13}$$

We come back to analyze (9.10). Similar to (8.76), we can show that

$$\begin{aligned}&\sum_{j_1,\dots,j_L}\prod_{l=1}^{L}\left(N^{-d/2-k}(|(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{j_l\mu}|+|(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{\mu j_l}|)^{d_l-\sum_{r=1}^{p}d_{l,r}}\right)\\&\prec N^{-1/2}\Psi^{d_{\mathbf{x}}-\sum_{l,r}d_{l,r}-\mathbf{1}(d_{\mathbf{x},\mu}=3)},\end{aligned}\tag{9.14}$$

where $d_{\mathbf{x},\mu} = \sum_{r=1}^{p} d_{\mathbf{x},\mu,r}$. At the right hand side of (9.14), comparing to (8.76), $\mathbf{1}(d_{\mathbf{v}} = 0)$ disappears since $d_{\mathbf{v}}$ is always equal to 0 for $\mathbf{v} = \mathbf{e}_{v_1}, ..., \mathbf{e}_{v_p}$. The reason why we can replace $d_{\mathbf{x}}$ by $d_{\mathbf{x},\mu}$ is because we don't consider $(\mathbf{G}^{(\tilde{i})}\mathbf{\Delta})_{\mathbf{e}_{v_r}j_l}$ and $(\mathbf{\Delta}^T\mathbf{G}^{(\tilde{i})})_{j_l\mathbf{e}_{v_r}}$ in (9.14). Also the reason why $d_{\mathbf{x}}$ can be replaced by $d_{\mathbf{x}} - \sum_{l,r} d_{l,\mu,r}$ is because the power at the left hand side of (9.14) becomes $d_l - \sum_{r=1}^{p} d_{l,\mu,r}$.

By the arguments above, we conclude that the LHS of (9.8) is bounded by

$$N^{d_i\delta}\Psi^{d_{\mathbf{x}}-\sum_{l,r} d_{l,\mu,r}-\mathbf{1}(d_{\mathbf{x},\mu}=3)}\Psi^{2(q+\sum_{r=q+1}^{p}|\sigma_r|)}\mathbb{E}(\tilde{F}(\mathbf{X})+\Psi^2)^{p-q-\sum_{r=q+1}^{p}|\sigma_r|}.$$

So (9.7) holds if

$$d_{\mathbf{x}} - \sum_{l,r} d_{l,r} - \mathbf{1}(d_{\mathbf{x},\mu} = 3) \geq 0. \tag{9.15}$$

In order to establish (9.15), we analyze $d_{\mathbf{x},\mu,r}$ carefully. First, if $k(g_r) = 0$, then $d_{\mathbf{x},\mu,r} = 0$. Secondly, if $1 \leq k(g_r) \leq 3$, the following holds

$$d_{\mathbf{x},r} \leq 2 \Longrightarrow d_{\mathbf{x},\mu,r} \leq I(k(g_r) \geq 2),$$

where $d_{\mathbf{x},r} = deg(A^+(r))$. Hence if $d_{\mathbf{x},\mu} = 3$, then there exists an $r \leq q$ such that $d_{\mathbf{x},r} \geq 3$. Then

$$d_{\mathbf{x},r} - \sum_{l} d_{l,r} - 1 \geq 0,$$

from $\sum_l d_{l,r} \leq 2$. Therefore, (9.15) holds by the fact that

$$d_{\mathbf{x},r} - \sum_{l} d_{l,r} \geq 0.$$

Therefore, we have proved the averaged local law.

10 Proof of Lemma 1

The proof of Lemma 1 is exactly the same as the proof of Lemma 13 in [12] and thus we omit it.

11 Proof of Theorem 7.2

Proof. Unlike [16], [7] and [4] we use the interpolation method (8.27), which is succinct and powerful when proving green function comparison theorems. In view of (8.5) and (8.6) we have

$$\begin{aligned}&|\mathbb{E}K(N\int_{E_1}^{E_2}\Im m_{\mathbf{X}^1}(x+i\eta)dx) - \mathbb{E}K(N\int_{E_1}^{E_2}\Im m_{\mathbf{X}^0}(x+i\eta)dx)| = \\ &\left|\mathbb{E}K(N\int_{E_1}^{E_2}\Im m_{\mathbf{X}^1}(x+i\eta)\mathcal{T}_N(\mathbf{X}^1)dx) - \mathbb{E}K(N\int_{E_1}^{E_2}\Im m_{\mathbf{X}^0}(x+i\eta)\mathcal{T}_N(\mathbf{X}^0)dx)\right| + O(N^{-1}).\end{aligned} \tag{11.1}$$

Applying (8.27) with $F(\mathbf{X}) = K(N\int_{E_1}^{E_2}\Im m_{\mathbf{X}}(x+i\eta)\mathcal{T}_N(\mathbf{X}))$ we only need to bound the following

$$\left|\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\mathbb{E}g(X_{i\mu}^1) - \mathbb{E}g(X_{i\mu}^0))\right|, \tag{11.2}$$

where

$$g(X_{i\mu}^u) = K(N\int_{E_1}^{E_2}\Im m_{\mathbf{X}_{(i\mu)}^{t,X_{i\mu}^u}}(x+i\eta)\mathcal{T}_N(\mathbf{X}_{(i\mu)}^{t,X_{i\mu}^u})dx), \quad u=0,1. \tag{11.3}$$

As in (8.35) and (8.36), we use Taylor's expansion up to order five to expand two functions $g(X_{i\mu}^u), u=0,1$ at the point 0. Then take the difference of the Taylor's expansions of $g(X_{i\mu}^u), u=0,1$. By the first two moments matching condition it then suffices to bound the third, fourth and remainder derivatives as follows

$$N^{-3/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\sum_{r=1}^{3}\sum_{\substack{k_1,..,k_r\in\mathbb{N}_+\\ k_1+..+k_r=3}} C_r\left|N\int_{E_1}^{E_2}\mathbb{E}K^{(r)}(0)\prod_{j=1}^{r}m_{\mathbf{X}_{(i\mu)}^{t,0}}^{(k_j)}(x+i\eta)\mathcal{T}_N(\mathbf{X}_{(i\mu)}^{t,0})dx\right|, \tag{11.4}$$

$$N^{-2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\sum_{r=1}^{4}\sum_{\substack{k_1,..,k_r\in\mathbb{N}_+\\ k_1+..+k_r=4}} C_r\max_x|K^{(r)}(x)|\mathbb{E}\prod_{j=1}^{r}\left(N\int_{E_1}^{E_2}\left|m_{\mathbf{X}_{(i\mu)}^{t,0}}^{(k_j)}(x+i\eta)\mathcal{T}_N(\mathbf{X}_{(i\mu)}^{t,0})\right|dx\right), \tag{11.5}$$

and the fifth derivative corresponding to the remainder of integral form

$$N^{-5/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\sum_{r=1}^{5}\sum_{\substack{k_1,..,k_r\in\mathbb{N}_+\\ k_1+..+k_r=4}} C_r\max_x|K^{(r)}(x)|\mathbb{E}\prod_{j=1}^{r}\left(N\int_{E_1}^{E_2}\left|m_{\mathbf{X}_{(i\mu)}^{t,\theta X_{i\mu}^u}}^{(k_j)}(x+i\eta)\mathcal{T}_N(\mathbf{X}_{(i\mu)}^{t,\theta X_{i\mu}^u})\right|dx\right), \tag{11.6}$$

where C_r is a constant depending on r only, $m_{\mathbf{X}_{(i\mu)}^{t,0}}^{(k_i)}(\cdot)$ denotes the k_ith derivative with respect to $X_{i\mu}^u$ and $0\le\theta\le1$. Here we ignore the terms involving the derivatives of $\mathcal{T}_N(\mathbf{X}_{(i\mu)}^{t,\theta X_{i\mu}^u})$ due to (8.5), (8.6) and (8.17).

We focus on (11.5) and (11.6) first. To investigate (11.5) and (11.6) we claim that it suffices to prove that

$$\left(N\int_{E_1}^{E_2}\left|m_{\mathbf{X}_{(i\mu)}^{u,X_{i\mu}^1}}^{(k)}(x+i\eta)\mathcal{T}_N(\mathbf{X}_{(i\mu)}^{u,X_{i\mu}^1})\right|dx\right) \prec (N^{\frac{1}{3}+\epsilon}\Psi^2), \tag{11.7}$$

where $k\ge1$. Indeed, if (11.7) holds then (11.7) still holds when $X_{i\mu}^1$ is replaced by $\theta X_{i\mu}^1$ by checking on the argument of (11.7). We then conclude that the facts that (11.5) $\prec (N^{\frac{1}{3}+\epsilon}\Psi^2)$ and that (11.6) $\prec (N^{-\frac{1}{2}+\frac{1}{3}+\epsilon}\Psi^2)$ follow from Lemma 4, (8.17) and an application of (8.35).

By (8.43) and (9.6) we have for $k\ge1$

$$\left|m_{\mathbf{X}_{(i\mu)}^{u,X_{i\mu}^1}}^{(k)}(x+i\eta)\mathcal{T}_N(\mathbf{X}_{(i\mu)}^{u,X_{i\mu}^1})\right| \prec \Psi^2, \tag{11.8}$$

which implies that (11.7) $\prec (N^{\frac{1}{3}+\epsilon}\Psi^2)$. Here we would point out that the derivatives $m^{(k)}_{\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)}}(\cdot)$ are of the form $\frac{1}{M_1}\sum_{k=1}^{M_1} C_{k,k,i,\mu}(g_h)$ from (8.42), (8.43), (8.44), (8.46), (9.3) and (9.4). By Lemma 2.3 of [5] we have

$$\Psi^2 \asymp \frac{1}{N\sqrt{\eta}} = O(N^{-\frac{2}{3}+\epsilon/2}). \tag{11.9}$$

From now on we consider (11.4). One should notice that we do not extract the summation $\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}$ outside the expectation like (11.5) in order to make it easier, compared with the proof of (9.7). Since $K(.)$ involved in the above expectation is non random and does not affect the order of the expectation, we can ignore $K^{(r)}(0)$ in the sequel. Similar to the claim (11.7), it suffices to find the upper bound of

$$N^{-3/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\sum_{r=1}^{3}\sum_{\substack{k_1,..,k_r\in\mathbb{N}_+\\ k_1+..+k_r=3}}\left|N\int_{E_1}^{E_2}\mathbb{E}\prod_{j=1}^{r} m^{(k_j)}_{\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)}}(x+i\eta)\mathcal{T}_N(\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)})dx\right|. \tag{11.10}$$

First of all, (11.8)-(11.9) always hold, which concludes that for $r\geq 2$

$$\begin{aligned}&N^{-3/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\sum_{r=2}^{3}\sum_{\substack{k_1,..,k_r\in\mathbb{N}_+\\ k_1+..+k_r=3}}\left|N\int_{E_1}^{E_2}\mathbb{E}\prod_{j=1}^{r} m^{(k_j)}_{\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)}}(x+i\eta)\mathcal{T}_N(\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)})dx\right|\\ &\prec N^{1/2+1/3+\epsilon}\Psi^4 \prec N^{-\frac{1}{2}+2\epsilon}.\end{aligned} \tag{11.11}$$

Therefore, referring to (11.10), it remains to consider the case $r=1$, i.e. we need to find the upper bound of

$$N^{-3/2}\sum_{i=1}^{M_1}\sum_{\mu=1}^{N}\left|N\int_{E_1}^{E_2}\mathbb{E}m^{(3)}_{\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)}}(x+i\eta)\mathcal{T}_N(\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)})dx\right|.$$

By checking (11.1)-(11.9) carefully one can find if we can extract one more $\frac{1}{\sqrt{N}}$ from the expectation above then the proof of this theorem is complete. In other words, the aim is to prove that

$$\left|\mathbb{E}m^{(3)}_{\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)}}(x+i\eta)\mathcal{T}_N(\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)})dx\right| \prec N^{-1/2}\Psi^2.$$

This can be proved by (8.43) and (9.9) as in (8.76) and (9.14) and the details are ignored here. Here we would comment that $N^{-1/2}$ comes from counting the number of the i-th row of $\mathbf{X}$ in the expansion of $m^{(3)}_{\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)}}(x+i\eta)\mathcal{T}_N(\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)})$ and one can also refer to the arguments above (9.8) to see $d_{\mathbf{v}}=0$. In addition Ψ^2 follows from (9.9) and the fact that there are always two indices $\mathbf{e}_{v_r}$ involved in $m^{(3)}_{\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)}}(x+i\eta)\mathcal{T}_N(\mathbf{X}^{u,X^1_{i\mu}}_{(i\mu)})$.

Summarizing the above we have shown that

$$|\mathbb{E}K(N\int_{E_1}^{E_2}\Im m_{\mathbf{X}^1}(x+i\eta)dx)-\mathbb{E}K(N\int_{E_1}^{E_2}\Im m_{\mathbf{X}^0}(x+i\eta)dx)| \prec N^{-\frac{1}{3}+2\epsilon}. \tag{11.12}$$

The proof is complete by choosing an appropriate ϵ.

□

References

[1] Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*, 2nd ed. Wiley, New York.

[2] Bai, Z. D. and Silverstein, J. W. (2006). *Spectral analysis of large dimensional random matrices*, 1st ed. Springer, New York.

[3] Bao, Z. G., Hu, J. Pan, G. M. and Zhou, W. (2015). Canonical correlation coefficients of high-dimensional normal vectors: finite rank case. http://arxiv.org/abs/1407.7194.

[4] Bao, Z. G., Pan, G. M. and Zhou, W. (2015). Universality for the largest eigenvalue of sample covariance matrices with general population. *Ann. Statist.* **43**(1), 382–421.

[5] Bao, Z. G., Pan, G. M. and Zhou, W. Local density of the spectrum on the edge for sample covariance matrices with general population. *Preprint.* Available at http://www. ntu. edu. sg/home/gmpan/publications. html.

[6] EL Karoui, N. (2007). Tracy-Widom Limit for the Largest Eigenvalue of a Large Class of Complex Sample Covariance Matrices, *Ann. Probab.* **35**,663-714.

[7] Erdös, L., Yau, H.-T., and Yin, J.(2011). Rigidity of Eigenvalues of Generalized Wigner Matrices , *Advances in Mathematics*, **229(3)**, 1435-1515.

[8] Féral, D., Péché, S.(2009). The largest eigenvalues of sample covariance matrices for a spiked population: Diagonal case. J. Math. Phys. **50**, 073302.

[9] Fujikoshi, Y., Ulyanov, V. V., and Shimizu, R. (2009). Multivariate Statistics : High-Dimensional and Large-Sample Approximations. Wiley.

[10] Gao, C., Ma, Z. and Ren, Z., H. Zhou. (2016). Minimax Estimation in Sparse Canonical Correlation Analysis. To appear in Annals of Statistics.

[11] H. Hotelling. (1936). Relations between two sets of variates. Biometrika, 321-377.

[12] Han, X, Pan, G. M. and Zhang, B(2016). The Tracy-Widom law for the Largest Eigenvalue of F Type Matrix. To appear in Annals of Statistics.
http://www3.ntu.edu.sg/home/gmpan/Fmatrix_30052015.pdf

[13] Johnstone, I.M. (2001). On the Distribution of the Largest Eigenvalue in Principal Component Analysis, *Ann. Statist.* **29**, 295-327.

[14] Johnstone, I. M. (2008). Multivatiate analysis and Jacobi ensembles:Largest eigenvalue,Tracy-Widom limits and rates of convergence. *Ann. Statist.* **36** 2638–2716.

[15] Johnstone, I. M. (2009). Approximation null distribution of the largest root in multivariate analysis. *Ann. Appl. Statist.* **3** No.4 1616–1633.

[16] Knowles, A. and Yin, J. (2015). Anisotropic local laws for random matrices. *arXiv:1410.3516v3.*

[17] Lee, J. O. and Schnelli, K. (2014). Tracy-Widom Distribution for the Largest Eigenvalue of Real Sample Covariance Matrices with General Population. *arXiv:1409.4979v1.*

[18] Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sb. Math.* **4** 457–483.

[19] Muirhead, R. J. (1982). Aspects of Multivariate Statistical Theory. *Wiley, New York.* MR0652932.

[20] Pillali, N. S. and Yin, J. (2011). Universality of covariance matrices. *Ann. Appl. Prob.* **24** No.3,935–1001.

[21] Silverstein, J. W. and Choi,S.-I (1995). Analysis of the Limiting Spectral Distribution of Large Dimensional Random Matrices. *Journal of Multivariate Analysis*, 54(2), 295C309.

[22] Soshnikov, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. Jour. Stat. Phys. **108**(5), 1033-1056.

[23] Tao, T. and Vu,V. (2011). Random matrices: Universality of local eigenvalue statistics. Acta Mathematica, **206**(1), 127-204.

[24] Tao, T. and Vu, V. (2012). Random covariance matrices: Universality of local statistics of eigenvalues. Ann. Probab. **40**(3), 1285-1315.

[25] Tracy, C. A. and Widom, H. (1994). Level-spacing distributions and the Airy kernel. Comm. Math. Phys. **159**, No. 1, 151-174.

[26] Tracy, C. A. and Widom, H. (1996). On orthogonal and symplectic matrix ensembles. Comm. Math. Phys. **177**, No. 3, 727-754.

[27] Wang, K. (2012). Random covariance matrices: Universality of local statistics of eigenvalues up to the edge. Random matrices: Theory and Applications, **1**(1), 1150005.

[28] Yang, Y. R. and Pan, G. M. (2015). Independence test for high dimensional data based on regularized canonical correlation coefficients. *Ann. Statist.* **43(2)**, 467-500.

[29] Zheng, S. R. (2012). Central Limit Theorem for Linear Spectral Statistics of Large Dimensional F Matrix. Ann. Institut Henri Poincare Probab. Statist. 48, 444-476.