A Note on the Likelihood Ratio Test in High-Dimensional Exploratory Factor Analysis

Yinqiu He, Zi Wang, and Gongjun Xu

Department of Statistics, University of Michigan

Abstract

The likelihood ratio test is widely used in exploratory factor analysis to assess the model fit and determine the number of latent factors. Despite its popularity and clear statistical rationale, researchers have found that when the dimension of the response data is large compared to the sample size, the classical chi-square approximation of the likelihood ratio test statistic often fails. Theoretically, it has been an open problem when such a phenomenon happens as the dimension of data increases; practically, the effect of high dimensionality is less examined in exploratory factor analysis, and there lacks a clear statistical guideline on the validity of the conventional chi-square approximation. To address this problem, we investigate the failure of the chi-square approximation of the likelihood ratio test in high-dimensional exploratory factor analysis, and derive the *necessary and sufficient* condition to ensure the validity of the chi-square approximation. The results yield simple quantitative guidelines to check in practice and would also provide useful statistical insights into the practice of exploratory factor analysis.

Keywords: Exploratory factor analysis, likelihood ratio test, chi-square approximation.

This research is partially supported by NSF CAREER SES-1846747, DMS-1712717, and SES-1659328.

1 Introduction

Exploratory factor analysis serves as a popular statistical tool to gain insights into latent structures underlying the observed data (Gorsuch, 1988; Fabrigar and Wegener, 2011; Bartholomew et al., 2011). It is widely used in many application areas such as psychological and social sciences (Fabrigar et al., 1999; Preacher and MacCallum, 2002; Thompson, 2004; Finch and Finch, 2016). In factor analysis, the relationship among observed variables in data are explained by a smaller number of unobserved underlying variables, called common factors. To understand the underlying scientific patterns, one fundamental problem in factor analysis is to decide the minimum number of latent common factors that is needed to describe the statistical dependencies in data.

In order to determine the number of factors in exploratory factor analysis, a wide variety of procedures have been proposed; see reviews and discussions in Costello and Osborne (2005), Barendse et al. (2015) and Luo et al. (2019). For instance, one broad class of criteria are based on the eigenvalues of the sample correlation matrix of the observed data. Examples include the Kaiser criterion (Kaiser, 1960), the scree test (Cattell, 1966), the parallel analysis method (Horn, 1965; Keeling, 2000; Dobriban, 2020), and testing linear trend of eigenvalues (Bentler and Yuan, 1998) among many others. Another class of methods propose various goodness-of-fit indexes to select the number of factors, such as AIC (Akaike, 1987), BIC (Schwarz, 1978), the reliability coefficient (Tucker and Lewis, 1973), and the root mean square error of approximation (Steiger, 2016). Moreover, the likelihood ratio test provides another popularly used approach in practice (Bartlett, 1950; Anderson, 2003).

Among the various criteria to determine the number of factors, the likelihood ratio test plays a unique role, as it is based on a formal hypothesis testing procedure with a clear statistical rationale and also has a solid theoretical foundation with guaranteed statistical properties. In particular, the likelihood ratio test examines how a factor analysis model fits the data using a hypothesis testing framework based on the likelihood theory. The classical statistical theory shows that under the null hypothesis, the likelihood ratio test statistic (after proper scaling) asymptotically converges to a chi-square distribution, with the degrees of freedom equal to the difference in the number of free

parameters between the null and alternative hypothesis models (see, e.g., Anderson, 2003, Section 14.3.2).

In the modern big data era, it is of emerging interest to analyze high-dimensional data (Finch and Finch, 2016; Harlow and Oswald, 2016; Chen et al., 2019), where throughout this paper we refer to the dimension of the observed response variables as the dimension of data. Classical asymptotic theory, despite its importance, often replies on the assumption that the data dimension is fixed as the sample size increases. Such an assumption often fails in high-dimensional data analysis with large data dimension, and therefore the corresponding asymptotic theory is no longer directly applicable to modern high-dimensional applications. In fact, it has been found in the recent statistical literature that the chi-square approximations for the likelihood ratio test statistics can become inaccurate as the dimension of data increases with the sample size (e.g. Bai et al., 2009; Jiang and Yang, 2013; He et al., 2020a). In factor analysis, although considerable high-dimensional statistical analysis results have been recently developed (Bai and Ng, 2002; Bai and Li, 2012; Sundberg and Feldmann, 2016; Ait-Sahalia and Xiu, 2017; Chen and Li, 2020), less attention has been paid to the statistical properties of the popular likelihood ratio test under high dimensions. Particularly, it remains an open problem when the conventional chi-square approximation of the likelihood ratio test starts to fail as the data dimension grows. In other words, for a dataset with sample size N, how large the data dimension p can be to still ensure the validity of the chi-square approximation of the likelihood ratio test?

To better understand this issue, this paper investigates the influence of the data dimensionality on the likelihood ratio test in high-dimensional exploratory factor analysis. Specifically, under the null hypothesis, we derive the necessary and sufficient condition for the chi-square approximation to hold. The results consider both the likelihood ratio test without and with the Bartlett correction, and provide useful quantitative guidelines that are easy to check in practice. Our simulation results are consistent with the theoretical conclusions, suggesting good finite-sample performance of the developed theory.

The rest of the paper is organized as follows. In Section 2.1, we give a brief review of the exploratory factor analysis and the likelihood ratio test, and in Section 2.2, we present our theoretical

and numerical results on the performance of the chi-square approximation under high dimensions. Several extensions are discussed in Section 3, and the technical proofs and additional simulation studies are deferred to the appendix.

2 Likelihood Ratio Test under High Dimensions

2.1 Likelihood ratio test for exploratory factor analysis

In this section, we briefly review the likelihood ratio test in exploratory factor analysis (see, e.g., Anderson, 2003, Section 14). Suppose X_i, $i = 1, \ldots, N$ are independent and identically distributed p-dimensional random vectors. The exploratory factor analysis considers the following common-factor model

$$X_i = \mu + \Lambda F_i + U_i, \tag{1}$$

where μ a the p-dimensional mean parameter vector, Λ is a $p \times k_0$ loading matrix with $\text{rank}(\Lambda) = k_0 < p$, F_i is a k_0-dimensional random vector containing the common factors, and U_i is a p-dimensional error vector. It is well known that the factor model (1) is not identifiable without additional constraints, and there are many ways to impose identifiablity restrictions (Anderson, 2003; Bai and Li, 2012). In this paper, we focus on the following identification conditions which have been popularly used in exploratory factor analysis. In particular, we assume that F_i and U_i are independent latent random vectors with $\text{E}(F_i) = \mathbf{0}_{k_0}$, $\text{cov}(F_i) = \text{I}_{k_0}$, $\text{E}(U_i) = \mathbf{0}_p$, and $\text{cov}(U_i) = \Psi$, where $\mathbf{0}_{k_0}$ denotes a k_0-dimensional all-zero vector, I_{k_0} represents a $k_0 \times k_0$ identity matrix, and Ψ is a $p \times p$ diagonal matrix with $\text{rank}(\Psi) = p$. It follows that the population covariance matrix $\Sigma = \text{cov}(X_i)$ can be expressed as

$$\Sigma = \Lambda\Lambda^\top + \Psi. \tag{2}$$

Typically, the true number of common factors k_0 is unknown. In exploratory factor analysis,

to determine the number of factors in model (1), various procedures have been developed. Among them, the likelihood ratio test plays a unique role due to its solid theoretical foundation and nice statistical properties. The common practice utilizes the model's likelihood function assuming both F_i and U_i to be normally distributed. In such case, X_i follows a multivariate normal distribution with mean vector $\mathbf{0}_p$ and covariance matrix Σ as in (2), and we write $X_i \sim \mathcal{N}(\mathbf{0}_p, \Sigma)$. Then, the likelihood ratio test is used to sequentially test the factor analysis model with a specified number of factors against the saturated model (e.g., Hayashi et al., 2007). Specifically, for each $k = 0, 1, \ldots, p$, we consider the following null and alternative hypotheses:

$$H_{0,k} : \Sigma = \Lambda\Lambda^\top + \Psi \text{ with (at most) } k \text{ factors}, \quad \text{versus} \quad H_{A,k} : \Sigma \text{ is any positive definite matrix.}$$

In practice without a priori knowledge, a typical procedure examines the above hypotheses in a forward stepwise manner. Specifically, we first consider $k = 0$ and examine $H_{0,0} : k_0 = 0$ versus $H_{A,0}$ using the likelihood ratio test, that is, testing whether there is any factor in model (1). If $H_{0,0}$ is rejected, we then consider $k = 1$, that is, a 1-factor model in the null hypothesis $H_{0,1}$. If $H_{0,1}$ is rejected, we proceed with $k = 2$, and test a 2-factor model for $H_{0,2}$. This testing procedure continues until we fail to reject $H_{0,\hat{k}}$ for some $\hat{k}$. Then $\hat{k}$ is taken as an estimate of the true number of factors based on the likelihood ratio test.

We next introduce the details on the abovementioned likelihood ratio test. For $k = 0$, $H_{0,0}$ examines the existence of any significant factors, which is an important problem in psychology applications (e.g., Mukherjee, 1970). This test can be written as

$$H_0 : \Sigma = \Psi \quad \text{versus} \quad H_A : \Sigma \neq \Psi,$$

that is, testing whether Σ is a diagonal matrix. Statistically, this is also equivalent to the following hypothesis test

$$H_0 : R = \mathrm{I}_p \quad \text{versus} \quad H_A : R \neq \mathrm{I}_p,$$

where R denotes the population correlation matrix of the response variables $\{X_i, i = 1, \ldots, N\}$.

Under the normality assumption of X, $H_{0,0}$ then tests for the complete independence between p dimensions of X. The likelihood ratio test statistic for $H_{0,0}$ with the chi-square limit is $T_0 = -(N-1)\log(|\hat{R}_N|)$, where $\hat{R}_N$ denotes the sample correlation matrix of the observations $\{X_i, i = 1, \ldots, N\}$, and $|\hat{R}_N|$ denotes the determinant of $\hat{R}_N$; see, e.g., Bartlett (1950). When the dimension p is fixed and the sample size $N \to \infty$, under $H_{0,0}$,

$$T_0 \xrightarrow{D} \chi^2_{f_0}, \quad \text{with } f_0 = p(p-1)/2, \tag{3}$$

where $\xrightarrow{D}$ represents the convergence in distribution, and $\chi^2_{f_0}$ represents a random variable following the chi-square distribution with degrees of freedom f_0. To improve the finite-sample performance, researchers have proposed using the Bartlett correction for the likelihood ratio test (Bartlett, 1950). The corrected test statistic is $\rho_0 T_0$ with the Bartlett correction term $\rho_0 = 1 - (2p+5)/\{6(N-1)\}$, and under $H_{0,0}$ with fixed p and $N \to \infty$, we still have the chi-square approximation:

$$\rho_0 \times T_0 \xrightarrow{D} \chi^2_{f_0}, \tag{4}$$

while it improves the convergence rate of the chi-square approximation (3) from $O(N^{-1})$ to $O(N^{-2})$.

For $k \geq 1$, $H_{0,k}$ examines whether the k-factor model fits the observed data. Under the k-factor model, let $\hat{\Lambda}_k$ and $\hat{\Psi}_k$ denote the maximum likelihood estimators of Λ and Ψ, respectively, and define $\hat{\Sigma}_k = \hat{\Lambda}_k \hat{\Lambda}_k^\top + \hat{\Psi}_k$. Then to test $H_{0,k}$, the likelihood ratio test statistic can be written as

$$T_k = -(N-1)\log(|\hat{\Sigma}| \times |\hat{\Sigma}_k|^{-1}) + (N-1)\{\mathrm{tr}(\hat{\Sigma}\hat{\Sigma}_k^{-1}) - p\}, \tag{5}$$

where $\hat{\Sigma}$ is the unbiased sample covariance matrix of the observations $\{X_i, i = 1, \cdots, N\}$, and $\mathrm{tr}(A)$ denotes the trace of a matrix A; see, e.g., Lawley and Maxwell (1962). Under the null hypothesis with $k_0 = k$, p fixed and $N \to \infty$, we have the following chi-square approximation:

$$T_k \xrightarrow{D} \chi^2_{f_k}, \quad \text{where } \; f_k = \{(p-k)^2 - p - k\}/2. \tag{6}$$

Moreover, applying the Bartlett correction for this test, we have

$$\rho_k \times T_k \xrightarrow{D} \chi^2_{f_k}, \quad \text{where } \rho_k = 1 - \frac{2p+5+4k}{6(N-1)}. \tag{7}$$

Despite the usefulness of the above chi-square approximations, classical large sample theory assumes that the data dimension p is fixed, and therefore many conclusions are not directly applicable to high-dimensional data when p increases with the sample size N. As analyzing high-dimensional data is of emerging interest in modern data science, it imposes new challenges to understanding the statistical performance of the likelihood ratio test in the exploratory factor analysis, which will be investigated in the next section.

2.2 Main results

In high-dimensional exploratory factor analysis, it is important to understand the limiting behavior of the likelihood ratio test, as applying an inaccurate limiting distribution would lead to misleading scientific conclusions. This section focuses on the limiting distribution of the likelihood ratio test under the null hypothesis, and investigates the influence of the data dimension p and the sample size N on the chi-square approximation.

Recent statistical literature has shown that the chi-square approximation for the likelihood ratio test can become inaccurate in various testing problems (Bai et al., 2009; Jiang and Yang, 2013; He et al., 2020a), while this inaccuracy issue is still less studied in the exploratory factor analysis. To demonstrate that similar phenomena exist for the exploratory factor analysis, we first present a numerical example, before showing our theoretical results.

Numerical Example 1. *Consider $H_{0,0}$ in Section 2.1 with $N = 1000$ and $p \in \{20, 100, 300, 500\}$. Under each combination of (N, p), we generate $X_i, i = 1, \ldots, N$ from $\mathcal{N}(\mathbf{0}_p, \mathrm{I}_p)$ independently, and then compute the likelihood ratio test statistics T_0 in (3) and its Bartlett corrected version $\rho_0 T_0$ in (4). We repeat the procedure 5000 times, and present the histograms of T_0 and $\rho_0 T_0$ in the first and second rows, respectively, of Figure 1. For comparison, in each histogram, we add the theoretical density curve of the limiting distribution $\chi^2_{f_0}$ in (3) and (4) (the red curves in Figure 1).*

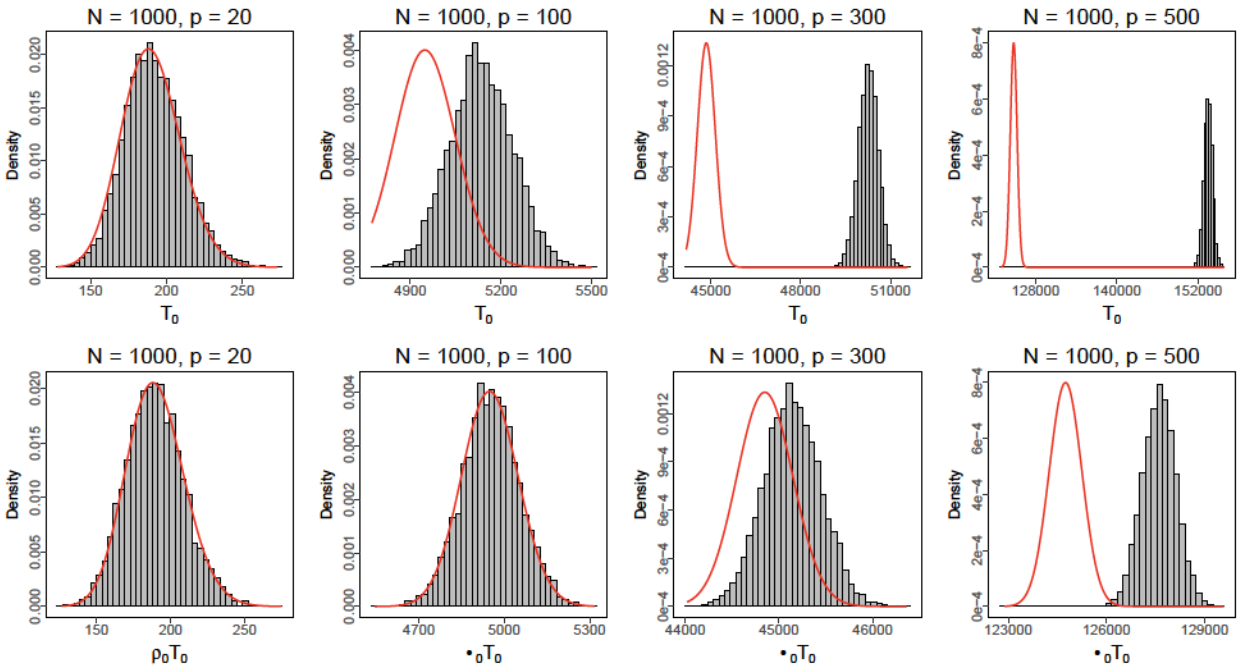

Figure 1: Histograms of T_0 and $\rho_0 T_0$ with the density curves of $\chi^2_{f_0}$

From the two figures in the first column of Figure 1, we can see that when p is small ($p = 20$) compared to N, the density curve of $\chi^2_{f_0}$ approximates the histograms of T_0 and $\rho_0 T_0$ well. This is consistent with the classical large sample theory in (3) and (4). However, as p increases from 20 to 500, the density curve of $\chi^2_{f_0}$ moves farther away from the sample histograms of T_0 and $\rho_0 T_0$, indicating the failure of the chi-square approximation as p increases. It is also interesting to note that the likelihood ratio test statistics without and with the Bartlett correction behave differently as p increases, despite their similarity when p is small. For instance, when $p = 100$, $\chi^2_{f_0}$ already fails to approximate the distribution of T_0, but it can still well approximate that of the corrected statistic $\rho_0 T_0$. Nevertheless, when $p = 300$ and 500, $\chi^2_{f_0}$ fails to approximate the distributions of both T_0 and $\rho_0 T_0$, while the approximation biases differ. These numerical observations bring the following question in practice: how large the dimension p with respect to the sample size N can be so that we can still apply the classic chi-square approximation for the likelihood ratio test?

To provide a statistical insight into this important practical issue, we derive the necessary and sufficient condition to ensure the validity of the chi-square approximation for the likelihood ratio test, as p increases with N. Particularly, we first consider $H_{0,0} : k_0 = 0$ in Section 2.1, and provide

the following Theorem 1.

Theorem 1 *Suppose $N \geq p+5$. Let $\chi^2_{f_0}(\alpha)$ denote the upper-level α-quantile of the $\chi^2_{f_0}$ distribution. Under $H_{0,0} : k_0 = 0$, as $N \to \infty$,*

(i) $\sup_{\alpha\in(0,1)} |\Pr\{T_0 > \chi^2_{f_0}(\alpha)\} - \alpha| \to 0$, *if and only if* $\lim_{n\to\infty} p/N^{1/2} = 0$;

(ii) $\sup_{\alpha\in(0,1)} |\Pr\{\rho_0 \times T_0 > \chi^2_{f_0}(\alpha)\} - \alpha| \to 0$, *if and only if* $\lim_{n\to\infty} p/N^{2/3} = 0$.

In Theorem 1, $N \geq p+5$ is required for the technical proof. This condition is mild as $N \geq p+1$ is required for the existence of the likelihood ratio test statistic with probability one (Jiang and Yang, 2013). Theorem 1 (i) suggests that the chi-square approximation for T_0 in (3) starts to fail when the dimension p approaches $N^{1/2}$, and (ii) shows that the chi-square approximation for $\rho_0 T_0$ in (4) starts to fail when p approaches $N^{2/3}$. To further demonstrate the validity of Theorem 1, we conduct a simulation study as follows.

Numerical Example 2. *We take $p = \lfloor N^\varepsilon \rfloor$, where $N \in \{100, 500, 1000, 2000\}$ and $\varepsilon \in \{3/24, 4/24, \ldots, 23/24\}$. For each combination of (N, p), we generate X_i from $\mathcal{N}(\mathbf{0}_p, \mathrm{I}_p)$ for $i = 1, \ldots, N$ independently, and conduct the likelihood ratio test with two chi-square approximations in* (3) *and* (4)*, respectively. We repeat the procedure 1000 times to estimate the type I error rates with significance level 0.05, and then plot estimated type I error rates versus ε in Figure 2. The left figure in Figure 2 presents the results of the chi-square approximation for T_0 in* (3)*, where the estimated type I error begins to inflate when ε approaches $1/2$. In addition, the right figure in Figure 2 presents the results of the chi-square approximation for $\rho_0 T_0$ in* (4)*, where the estimated type I error begins to inflate when ε approaches $2/3$. The two theoretical boundaries on ε in Theorem 1 are denoted by two vertical dashed lines in Figure 2. For each approximation, the theoretical and empirical values of ε where the approximation begins to fail are consistent.*

We next investigate the sequential test for $H_{0,k}$ when $k \geq 1$. Under $H_{0,k}$, assume the true factor number is k, and $\Lambda_k\Lambda_k^\top$ and Ψ_k are the true values such that (2) holds with $\Lambda\Lambda^\top = \Lambda_k\Lambda_k^\top$ and $\Psi = \Psi_k$, where Λ_k is a matrix of size $p \times k$, and Ψ_k is a diagonal matrix. In classical multivariate analysis with fixed dimension and certain regularity conditions, it can be shown that $\hat{\Lambda}_k\hat{\Lambda}_k^\top \xrightarrow{P} \Lambda_k\Lambda_k^\top$

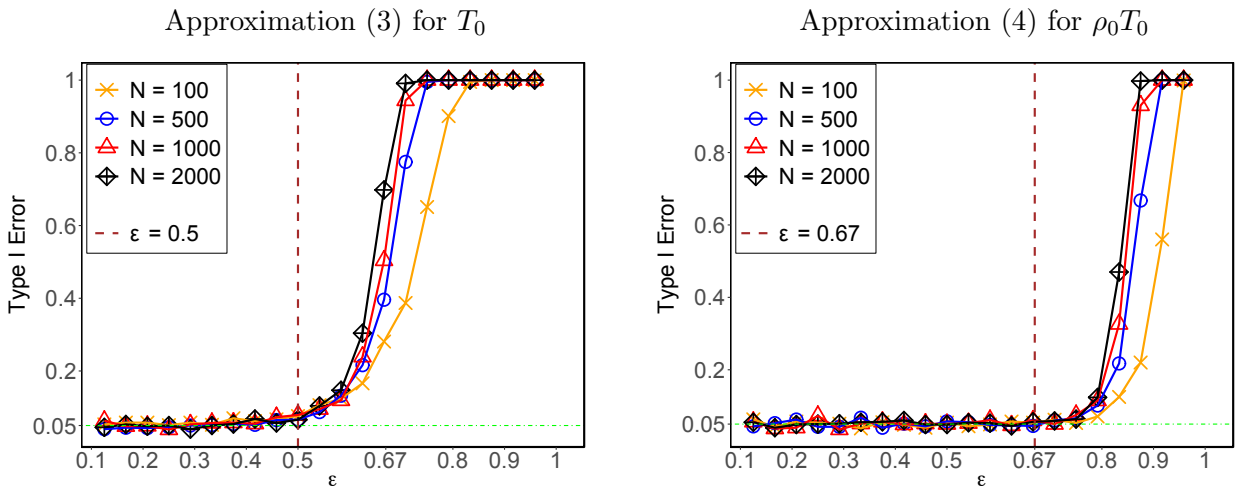

Figure 2: Estimated type I error versus ε when $k_0 = 0$

and $\hat{\Psi}_k \xrightarrow{P} \Psi_k$, where $\xrightarrow{P}$ represents the convergence in probability; see, e.g., Theorem 14.3.1 in Anderson (2003). To facilitate the following theoretical analysis, we consider a simplified version of the test by assuming $\Lambda_k \Lambda_k^\top$ and Ψ_k are given, and define $\Sigma_k = \Lambda_k \Lambda_k^\top + \Psi_k$. Then we consider testing $H'_{0,k} : \Sigma = \Sigma_k$, and the likelihood ratio test statistic can be expressed as

$$T' = -(N-1)\log(|\hat{\Sigma}| \times |\Sigma_k|^{-1}) + (N-1)\{\mathrm{tr}(\hat{\Sigma}\Sigma_k^{-1}) - p\};$$

see Section 8.4 of Muirhead (2009). The test statistic T' and T_k in (5) are the same except that T' is based on the true value $\Sigma_k = \Lambda_k \Lambda_k^\top + \Psi_k$, while T_k is based on $\hat{\Sigma}_k = \hat{\Lambda}_k \hat{\Lambda}_k^\top + \hat{\Psi}_k$, with $\hat{\Lambda}_k \hat{\Lambda}_k^\top$ and $\hat{\Psi}_k$ being the maximum likelihood estimators of $\Lambda_k \Lambda_k^\top$ and Ψ_k, respectively, under the k-factor model. Under the classical setting with p fixed, the chi-square approximation of T' is $T' \xrightarrow{D} \chi^2_{f'}$, where $f' = p(p+1)/2$, and by the Bartlett correction with $\rho' = 1 - \{6(N-1)(p+1)\}^{-1}(2p^2+3p-1)$, we have $\rho' T' \xrightarrow{D} \chi^2_{f'}$. For this simplified testing problem $H'_{0,k}$, the test statistic T' and its limit do not depend on the number of factors k, as the true $\Lambda_k \Lambda_k^\top$ and Ψ_k are assumed to be given.

Considering $H'_{0,k}$ and the statistic T', we next provide the necessary and sufficient condition on when the chi-square approximation for the likelihood ratio test fails as the data dimension p increases under $H'_{0,k}$.

Theorem 2 *Suppose $N \geq p+2$. Under $H'_{0,k}$: $\Sigma = \Lambda_k \Lambda_k^T + \Psi_k$, with given Λ_k and Ψ_k, and $k = k_0$,*

as $N \to \infty$,

(i) $\sup_{\alpha\in(0,1)} |\Pr\{T' > \chi^2_{f'}(\alpha)\} - \alpha| \to 0$, *if and only if* $\lim_{n\to\infty} p/N^{1/2} = 0$;

(ii) $\sup_{\alpha\in(0,1)} |\Pr\{\rho' \times T' > \chi^2_{f'}(\alpha)\} - \alpha| \to 0$, *if and only if* $\lim_{n\to\infty} p/N^{2/3} = 0$.

Remark 1 *For the more general testing problem* $H_{0,k}$*, we need to obtain the maximum likelihood estimators* $\hat{\Lambda}_k$ *and* $\hat{\Psi}_k$*, and then conduct the likelihood ratio test with chi-square approximations* (6) *or* (7)*. When the number of latent factors* k *is fixed compared to* N *and* p*, we note that* ρ_k/ρ' *and* f_k/f' *asymptotically converge to 1. Furthermore, if* $\hat{\Lambda}_k\hat{\Lambda}_k^\top + \hat{\Psi}_k$ *approximates the true* $\Lambda_k\Lambda_k^\top + \Psi_k$ *sufficiently well, we expect that the conclusions in Theorem 2 would hold for the likelihood ratio test under the null hypothesis* $H_{0,k}$ *similarly. In particular, when* k *is fixed as* $N \to \infty$*, consistent estimation of* Λ_k *and* Ψ_k *has been discussed under both fixed* p *in the classical literature (see, e.g., Anderson, 2003, Theorem 14.3.1) and* $p \to \infty$ *in recent literature on high-dimensional factor analysis model (see, e.g., Bai and Li, 2012). When* k *also diverges with* N *and* p*, an asymptotic regime that is less investigated in the literature, deriving a similar condition for the chi-squared approximation would require accurate characterizations of the biases of estimating* Λ_k *and* Ψ_k*, which, however, would be challenging and need new developments of high-dimensional theory and methodology.*

We next demonstrate the theoretical results through the following numerical study.

Numerical Example 3. *We consider the likelihood ratio test under* $H_{0,k}$ *with* $k = k_0 \in \{1, 3\}$*. (I) When* $k_0 = 1$*, under* $H_{0,1}$*, we set* $\Lambda = \rho \times \mathbf{1}_p$ *and* $\Psi = (1-\rho^2)\mathrm{I}_p$*, with* $\rho = 0.3$*. (II) When* $k_0 = 3$*, under* $H_{0,3}$*, we set* $\Psi = (1-\rho^2)\mathrm{I}_p$ *and*

$$\Lambda = \begin{bmatrix} \rho \times \mathbf{1}_{p_1} & \mathbf{0}_{p_1} & \mathbf{0}_{p_1} \\ \mathbf{0}_{p_1} & \rho \times \mathbf{1}_{p_1} & \mathbf{0}_{p_1} \\ \mathbf{0}_{p-2p_1} & \mathbf{0}_{p-2p_1} & \rho \times \mathbf{1}_{p-2p_1} \end{bmatrix},$$

where $p_1 = \lfloor p/3 \rfloor$*,* $\rho = 0.6$*, and* $\mathbf{1}_{p_1}$ *represents a* p_1*-dimensional vector with all one entries. For both cases, we set* $p = \lfloor N^\varepsilon \rfloor$*, where* $N \in \{100, 500, 1000, 2000\}$ *and* $\varepsilon \in \{8/24, 7/24, ..., 23/24\}$*. And*

we generate each observation $X_i, i = 1, \ldots, N$, *from* $\mathcal{N}(\mathbf{0}, \Lambda\Lambda^\top + \Psi)$ *independently, and conduct the likelihood ratio test with the function* factanal() *in R. Similarly to Figure 2, we plot the estimated type I error rates (based on 1000 replications) versus* ε *for two approximations* (6) *and* (7)*, where the results of case (I) are in Figure 3, and the results of case (II) are in Figure 4.*

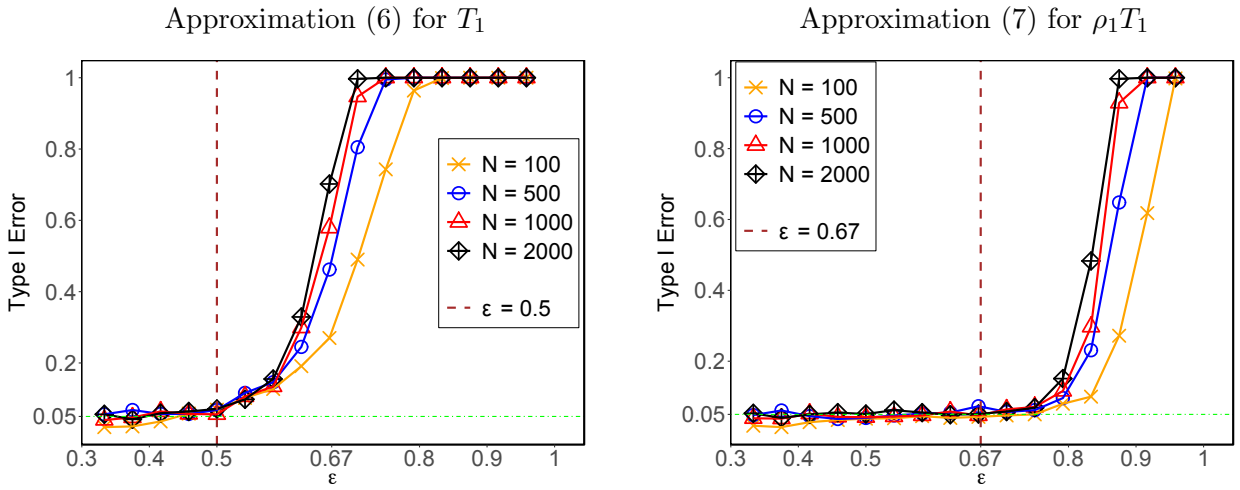

Figure 3: Estimated type I error versus ε when $k_0 = 1$

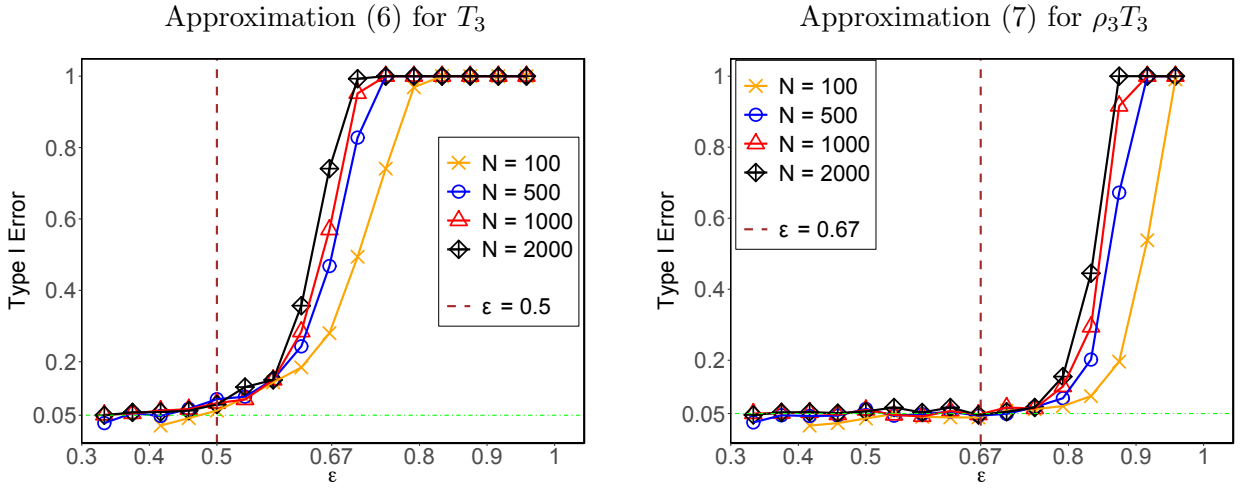

Figure 4: Estimated type I error versus ε when $k_0 = 3$

Similarly to Numerical Example 2, Numerical Example 3 also demonstrates that the empirical values of ε, where the chi-square approximations start to fail, are consistent with the corresponding

theoretical results. The necessary and sufficient conditions therefore would provide simple quantitative guidelines to check in practice. In addition, it is worth mentioning that the conditions in Theorems 1 and 2 also reflect the biases of the chi-square approximations. For instance, considering the likelihood ratio test for $H_{0,0}$, by the proof of Theorem 1, when $p/N \to 0$, we obtain that $\mathrm{E}(T_0 - \chi^2_{f_0}) \times \{\mathrm{var}(\chi^2_{f_0})\}^{-1/2}$ is approximately $C_1 p^2/N$, and $\mathrm{E}(\rho_0 \times T_0 - \chi^2_{f_0}) \times \{\mathrm{var}(\chi^2_{f_0})\}^{-1/2}$ is approximately $C_2 p^3/N^2$, where C_1 and C_2 are positive constants. This suggests that the mean of the chi-square limit will become smaller than the means of T_0 and $\rho_0 T_0$ as p increases, which is consistent with the observed phenomenon in Figure 1.

Moreover, Figures 2–4 show that the estimated type I error of the likelihood ratio test increases as ϵ increases. This can provide one possible explanation for the well-known finding that the likelihood ratio test tends to overestimate the number of factors (Hayashi et al., 2007). In particular, let $\hat{k}$ denote the number of factors estimated by the sequential procedure described in Section 2.1, and let k_0 denote the true number of factors. Note that in the sequential procedure, rejecting H_{0,k_0} leads to an overestimation of the number of factors, i.e., $\hat{k} > k_0$. Thus, when the type I error of testing H_{0,k_0} inflates as in Figures 2–4, the probability of rejecting H_{0,k_0} would also increase, which consequently suggests an inflation of the probability of overestimating the number of factors, $\hat{k} > k_0$. We also conduct simulation studies in Section B.2 to demonstrate the performance of estimating the number of factors using the likelihood ratio test. The numerical results are consistent with the above theoretical analyses and show that the procedure begins to overestimate the number of factors when the type I error begins to inflate.

Furthermore, Theorems 1 and 2 indicate that given the same sample size, the chi-square approximation with the Bartlett correction can hold for a larger p than the one without the Bartlett correction. This explains the patterns observed in Figure 1. Under the classical settings where p is fixed, researchers have shown that the Bartlett correction can improve the convergence rate of the likelihood ratio test statistic from $O(N^{-1})$ to $O(N^{-2})$; however, this result does not apply to the high-dimensional setting with p increasing with N. Our theoretical results in Theorems 1 and 2 provide a more precise description on how the Bartlett correction improves the chi-square approximations for high-dimensional data, in terms of the failing boundary of p with respect to N.

Remark 2 *Similar phase transition phenomena were discussed in He et al. (2020b). However, we point out that this paper considers different problem settings. In particular, He et al. (2020b) discussed several problems on testing mean vectors and covariances, whereas Theorem 1 examines testing correlation matrices. Moreover, Theorem 2 considers a problem of testing the covariance equal to a given k-factor matrix, which was not discussed in He et al. (2020b). To establish the result, it is required to derive a new high-dimensional asymptotic result given as Lemma 3 in the Appendix of this paper.*

3 Discussions

This paper investigates the influence of the data dimension on the popularly used likelihood ratio test in high-dimensional exploratory factor analysis. For the likelihood ratio test without or with the Bartlett correction, we derive the necessary and sufficient conditions to ensure the validity of the chi-square approximations under the corresponding null hypothesis. The developed theoretical conditions only depend on the relationship between the data dimension and the sample size, and would provide simple quantitative guidelines to check in practice.

The theoretical results in this paper are established under the common normality assumption of the observations $\{X_i, i = 1, \ldots, N\}$. To illustrate the robustness of the theoretical results to the normality assumption, we conduct additional simulation studies with X_i's following a discrete distribution or a heavy-tailed t-distribution in Appendix. Similar numerical findings are observed when detecting the existence of factors, which suggests that the validity of the theoretical results and the usefulness of the developed conditions in practice. Please see Section B.1 in Appendix.

Moreover, this paper focuses on controlling the type I error when testing a given null hypothesis, whereas deciding the number of factors would involve multiple steps of hypothesis testing in the sequential procedure. When the derived phase transition conditions are satisfied, our theoretical results suggest that the type I error of testing corresponding null hypothesis can be asymptotically controlled. However, the probability of correctly deciding the true number of factors relies on not only the type I error but also the power of testing each hypothesis in the sequential procedure.

The power of the likelihood ratio test depends on certain complicated hypergeometric functions (Muirhead, 2009), which would be very challenging to investigate under high dimensions. We would like to leave this interesting problem as a future study. In addition to the likelihood ratio test, it is also of interest to develop other efficient methods for deciding the number of factors in high-dimensional settings (see, e.g., Bai and Ng, 2002; Chen and Li, 2020).

When applying the likelihood ratio test in the exploratory factor analysis, it is worth noting that the data dimension p is not the only condition to consider. Researchers have discussed various other regularity conditions such as small sample size (MacCallum et al., 1999; Mundfrom et al., 2005; Winter et al., 2009; Winter and Dodou, 2012), nonnormality (Yuan et al., 2002; Barendse et al., 2015), and rank deficiency (Hayashi et al., 2007). The results in this paper only provide one necessary requirement to check in the high-dimensional exploratory factor analysis.

The results in this paper are also related to the important design problem on minimum sample size requirement for the exploratory factor analysis (Velicer and Fava, 1998; Mundfrom et al., 2005). The existing literature have conducted extensive simulation studies to explore what is the minimum sample size N required or how large the ratio N/p should be. In this paper, we derive theoretical results suggesting that we may also consider the polynomial relationship between N and p. Specifically, given the number of variables p to consider, the sample size should be at least p^2 to apply the likelihood ratio test, and at least $p^{3/2}$ to apply the likelihood ratio test with the Bartlett correction. This may provide helpful statistical insights into the practice of exploratory factor analysis.

Although this paper focuses on the exploratory factor analysis, we expect that the failure of chi-square approximations under high dimensions can happen generally in other latent factor modeling problems such as the confirmatory factor analysis (Thompson, 2004; Koran, 2020) and the exploratory item factor analysis (Reckase, 2009; Chen et al., 2019). Moreover, the phenomena introduced in this paper may also occur for other fit indexes that involve certain chi-square limit, such as the root mean square error of approximation (Steiger, 2016). New high-dimensional theory and methodology for these problems would need to be further investigated.

Acknowledgement

The authors are grateful to the Editor-in-Chief Professor Matthias von Davier, an Associate Editor, and three referees for their valuable comments and suggestions. This research is partially supported by NSF CAREER SES-1846747, DMS-1712717, and SES-1659328.

References

Ait-Sahalia, Y. and Xiu, D. (2017). Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics*, 201(2):384–399.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52(3):317–332.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley, New York, NY.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436–465.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, Z., Jiang, D., Yao, J.-F., and Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.*, 37(6B):3822–3840.

Barendse, M., Oort, F., and Timmerman, M. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(1):87–101.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, 3(2):77–85.

Bentler, P. M. and Yuan, K.-H. (1998). Tests for linear trend in the smallest eigenvalues of the correlation matrix. *Psychometrika*, 63(2):131–144.

Cattell, R. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276.

Chen, Y. and Li, X. (2020). Determining the number of factors in high-dimensional generalised latent factor models. *arXiv preprint arXiv:2010.02326*.

Chen, Y., Li, X., and Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84(1):124–146.

Costello, A. B. and Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation*, 10(1):7.

Dobriban, E. (2020). Permutation methods for factor analysis and PCA. *The Annals of Statistics*.

Fabrigar, L. R. and Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3):272–299.

Finch, W. H. and Finch, M. E. H. (2016). Fitting exploratory factor analysis models with high dimensional psychological data. *Journal of Data Science*, 14(3):519–537.

Gorsuch, R. L. (1988). *Exploratory Factor Analysis*, pages 231–258. Springer US, Boston, MA.

Harlow, L. L. and Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, 21(4):447.

Hayashi, K., Bentler, P., and Yuan, K.-H. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3):505–526.

He, Y., Jiang, T., Wen, J., and Xu, G. (2020a). Likelihood ratio test in multivariate linear regression: from low to high dimension. *Statistica Sinica*.

He, Y., Meng, B., Zeng, Z., and Xu, G. (2020b). On the phase transition of Wilks' phenomenon. *Biometrika*, to appear.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

Jiang, T. and Qi, Y. (2015). Likelihood ratio tests for high-dimensional normal distributions. *Scandinavian Journal of Statistics*, 42(4):988–1009.

Jiang, T. and Yang, F. (2013). Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *Ann. Statist.*, 41(4):2029–2074.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151.

Keeling, K. B. (2000). A regression equation for determining the dimensionality of data. *Multivariate Behavioral Research*, 35(4):457–468.

Koran, J. (2020). Indicators per factor in confirmatory factor analysis: More is not always better. *Structural Equation Modeling: A Multidisciplinary Journal*, 0(0):1–8.

Lawley, D. N. and Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229.

Luo, L., Arizmendi, C., and Gates, K. M. (2019). Exploratory Factor Analysis (EFA) Programs in R. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(5):819–826.

MacCallum, R. C., Widaman, K. F., Zhang, S., and Hong, S. (1999). Sample size in factor analysis. *Psychological methods*, 4(1):84.

Muirhead, R. J. (2009). *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons.

Mukherjee, B. N. (1970). Likelihood ratio tests of statistical hypotheses associated with patterned covariance matrices in psychology. *British Journal of Mathematical and Statistical Psychology*, 23(2):89–120.

Mundfrom, D. J., Shaw, D. G., and Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, 5(2):159–168.

Preacher, K. J. and MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior genetics*, 32(2):153–161.

Reckase, M. (2009). *Multidimensional Item Response Theory*. Statistics for Social and Behavioral Sciences. Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Steiger, J. H. (2016). Notes on the Steiger–Lind (1980) handout. *Structural equation modeling: A multidisciplinary journal*, 23(6):777–781.

Sundberg, R. and Feldmann, U. (2016). Exploratory factor analysis – Parameter estimation and scores prediction with high-dimensional data. *Journal of Multivariate Analysis*, 148:49–59.

Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*. American Psychological Association.

Tucker, L. R. and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1):1–10.

Velicer, W. F. and Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological methods*, 3(2):231.

Winter, J. C. F. and Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, 39(4):695–710.

Winter, J. C. F., Dodou, D., and Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, 44(2):147–181.

Yuan, K.-H., Marshall, L. L., and Bentler, P. M. (2002). A unified approach to exploratory factor analysis with missing data, nonnormal data, and in the presence of outliers. *Psychometrika*, 67(1):95–121.

Appendix

This appendix presents the technical proofs in Section A and additional simulations in Section B.

A Proofs

We prove Theorems 1 and 2 in Sections A.1 and A.2, respectively, and provide a required lemma and its proof in Section A.3. In the following proofs, for two sequences of number $\{a_N : N \geq 1\}$ and $\{b_N : N \geq 1\}$, $a_N = O(b_N)$ denotes $\limsup_{n\to\infty} |a_N/b_N| < \infty$, and $a_N = o(b_N)$ denotes $\lim_{N\to\infty} a_N/b_N = 0$.

A.1 Proof of Theorem 1

To derive the necessary and sufficient condition on the dimension of data, it is required to correctly understand the limiting behavior of the likelihood ratio test statistic under both low- and high-dimensional settings. In particular, we examine the limiting distribution of the likelihood ratio test statistic based on its moment generating function. For easy presentation in the technical proof, we let $n = N - 1$ below. Then we can write $T_0 = -n \log |\hat{R}_n|$. Under the conditions of Theorem 1, by Theorem 5.1.3 in Muirhead (2009) and Lemma 5.10 in Jiang and Yang (2013), we know that there exists a small constant $\delta_0 > 0$ such that for $h \in (-\delta_0, \delta_0)$,

$$\mathrm{E}\{\exp(h \times T_0)\} = \mathrm{E}\{|\hat{R}_n|^{-hn}\} = \left\{ \frac{\Gamma(n/2)}{\Gamma(n/2 - hn)} \right\}^p \times \frac{\Gamma_p(n/2 - hn)}{\Gamma_p(n/2)},$$

where $\Gamma(z)$ denotes the Gamma function, and $\Gamma_p(z)$ denotes the multivariate Gamma function satisfying $\Gamma_p(z) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\{z - (j-1)/2\}$.

Part (i) The chi-square approximation. When p is fixed compared to N, by applying Stirling's approximation to the Gamma function, it can be shown that as $N \to \infty$, for any $h \in (-\delta_0, \delta_0)$, $\mathrm{E}\{\exp(h \times T_0)\}$ converges to $(1-2h)^{-f_0/2}$, which is the moment generating function of $\chi^2_{f_0}$; see, e.g., Bartlett (1950) and Section 5.1.2 of Muirhead (2009). It follows that $T_0 \xrightarrow{D} \chi^2_{f_0}$ by the continuity theorem. When $p \to \infty$, Jiang and Yang (2013) and Jiang and Qi (2015) derived an approximate expansion of the multivariate Gamma function $\Gamma_p(\cdot)$ when p increases with the sample size N, and then showed that for any $h \in (-\delta_0, \delta_0)$,

$$\mathrm{E}[\exp\{h(T_0 + n\mu_{n,0})/(n\sigma_{n,0})\}] \to \exp(h^2/2), \tag{8}$$

where δ_0 is a constant that is sufficiently small, $\exp(h^2/2)$ is the moment generating function of the standard normal random variable $\mathcal{N}(0,1)$, and

$$\mu_{n,0} = (p - n + 1/2)\log\left(1 - \frac{p}{n}\right) - \frac{n-1}{n}p, \quad \sigma^2_{n,0} = -2\left\{\frac{p}{n} + \log\left(1 - \frac{p}{n}\right)\right\}.$$

This suggests $(T_0 + n\mu_{n,0})/(n\sigma_{n,0}) \xrightarrow{D} \mathcal{N}(0,1)$ by the continuity theorem. Note that $\chi^2_{f_0}$ can be viewed as a summation of the squares of f_0 independent standard normal random variables, and $f_0 \to \infty$ when $p \to \infty$. By applying the central limit theorem to $\chi^2_{f_0}$ when $p \to \infty$, we obtain $(\chi^2_{f_0} - f_0)/\sqrt{2f_0} \xrightarrow{D} \mathcal{N}(0,1)$, giving $\mathrm{E}[\exp\{h(\chi^2_{f_0} - f_0)/\sqrt{2f_0}\}] \to \exp(h^2/2)$. Therefore, if the chi-square approximation for T_0 holds, we know $\mathrm{E}[\exp\{h(T_0 - f_0)/\sqrt{2f_0}\}] \to \exp(h^2/2)$ for $h \in (-\delta_0, \delta_0)$, which, given (8), is equivalent to

$$\sqrt{2f_0} \times (n\sigma_{n,0})^{-1} \to 1, \tag{9}$$

$$(f_0 + n\mu_{n,0}) \times (n\sigma_{n,0})^{-1} \to 0. \tag{10}$$

We next examine (9) and (10) by discussing two cases $\lim_{n\to\infty} p/n = 0$ and $\lim_{n\to\infty} p/n = C \in (0, 1]$, respectively.

Case (i.1): $\lim_{n\to\infty} p/n = 0$. Under this case, we show that (9) holds. By Taylor's expansion, $\log(1-x) = -x - x^2/2 + O(x^3)$ for $x \in (0, 1)$, and then

$$\sigma_{n,0}^2 = -2\left\{-\frac{p^2}{2n^2} + O\left(\frac{p^3}{n^3}\right)\right\} = \frac{p^2}{n^2}\{1 + o(1)\}. \tag{11}$$

Recall that $f_0 = p(p-1)/2$, and it follows that (9) holds. Next we prove (10) holds if and only if $p/n^{1/2} \to 0$. Similarly by Taylor's expansion and $p/n = o(1)$, we have

$$\begin{aligned}
\mu_{n,0} &= (-n + p + 1/2)\left\{-\sum_{k=1}^{3}\frac{1}{k}\left(\frac{p}{n}\right)^k + O\left(\frac{p^4}{n^4}\right)\right\} - \frac{n-1}{n}p \\
&= p + \frac{p^2}{2n} + \frac{p^3}{3n^2} - \frac{p(p+1/2)}{n} - \frac{p^2(p+1/2)}{2n^2} + O\left(\frac{p^4}{n^3}\right) - p + \frac{p}{n} \\
&= \frac{p}{2n} - \frac{p^2}{2n} - \frac{p^3}{6n^2} + O\left(\frac{p^4}{n^3}\right) + o\left(\frac{p}{n}\right),
\end{aligned}$$

and then $f_0 + n\mu_{n,0} = -p^3/(6n) + O(p^4/n^2) + o(p)$. Given that (9) holds under this case and $\sqrt{2f_0}/p \to 1$, we obtain $(f_0 + n\mu_{n,0}) \times (n\sigma_{n,0})^{-1} = -p^2/(6n) + O(p^3/n^2) + o(1)$, which converges to 0 if and only if $p^2/n \to 0$ under this case.

Case (i.2): $\lim_{n\to\infty} p/n = C \in (0, 1]$. Under this case, we show that (9) does not hold. Note that

$$\frac{2f_0}{n^2\sigma_{n,0}^2} \to \frac{C^2}{-2\{C + \log(1-C)\}}.$$

If $C = 1$, $2f_0/(n^2\sigma_{n,0}^2) \to 0$, and thus (9) does not hold. We next consider $C \in (0, 1)$. If (9) holds, we shall have $g_1(C) = 0$ with $g_1(C) = C^2 + 2\{C + \log(1-C)\}$. By taking derivative of $g_1(C)$, we obtain

$$g_1'(C) = 2C + 2 - \frac{2}{1-C} = -\frac{2C^2}{1-C} < 0$$

when $C \in (0, 1)$. This suggests that $g_1(C)$ is strictly decreasing on $C \in (0, 1)$. As $g_1(0) = 0$, we

know $g_1(C) < 0$ for $C \in (0,1)$, and thus (9) does not hold.

Finally, we consider a general sequence $p/n \in (0,1]$, and write $p_n = p$ and $f_{n,0} = f_0$ below to emphasize that p and f_0 change with n. For the bounded sequence $\{p_n/n\}$, by the Bolzano-Weierstrass theorem, we can further take a subsequence $\{p_{n_k}/n_k\}$ such that $p_{n_k}/n_k \to C \in [0,1]$. If $C \in (0,1]$, the analysis in Case (i.2) applies, and we know $\sqrt{2f_{n_k,0}} \times (n_k \sigma_{n_k,0})^{-1}$ does not converge to 1. Since a sequence converges if and only if every subsequence converges, we know (9) does not converge to 1 under this case. Alternatively, if all the subsequences of $\{p/n\}$ converge to 0, we know $p/n \to 0$, and the analysis in Case (i.1) applies. In summary, the chi-square approximation holds if and only if $p^2/n \to 0$.

Part (ii) The chi-square approximation with the Bartlett correction. Similarly to the proof of Part (i), when p is fixed, it has been shown that $\mathrm{E}\{\exp(h \times \rho_0 \times T_0)\} \to (1-2h)^{-f_0/2}$ for $h \in (-\delta_0, \delta_0)$ and $\rho_0 = 1 - (2p+5)/(6n)$ (see, e.g., Bartlett, 1950); when $p \to \infty$, we also have (8) holds. If the chi-square approximation with the Bartlett correction holds, $\mathrm{E}[\exp\{h(\rho_0 T_0 - f_0)/\sqrt{2f_0}\}] \to \exp(h^2/2)$ for $h \in (-\delta_0, \delta_0)$, which, given (8), is equivalent to

$$\sqrt{2f_0} \times (n\rho_0 \times \sigma_{n,0})^{-1} \to 1, \tag{12}$$

$$(f_0 + n\rho_0 \times \mu_{n,0}) \times (n\rho_0 \times \sigma_{n,0})^{-1} \to 0. \tag{13}$$

Case (ii.1): $\lim_{n\to\infty} p/n = 0$. Under this case, we have (12) holds given $\rho_0 \to 1$ and (9) proved above. We next prove (13) holds if and only if $p^3/n^2 \to 0$. Similarly to the proof in Case (i.1), by Taylor's expansion and $p/n = o(1)$, we have

$$\begin{aligned}
\mu_{n,0} &= (-n+p+1/2)\left\{-\sum_{k=1}^{4}\frac{1}{k}\left(\frac{p}{n}\right)^k + O\left(\frac{p^5}{n^5}\right)\right\} - \frac{n-1}{n}p \\
&= p + \frac{p^2}{2n} + \frac{p^3}{3n^2} + \frac{p^4}{4n^3} - \frac{p(p+1/2)}{n} - \frac{p^2(p+1/2)}{2n^2} \\
&\quad -\frac{p^3(p+1/2)}{3n^3} + O\left(\frac{p^5}{n^4}\right) - p + \frac{p}{n} \\
&= \frac{p}{2n} - \frac{p^2}{2n} - \frac{p^3}{6n^2} - \frac{p^4}{12n^3} + O\left(\frac{p^5}{n^4}\right) + o\left(\frac{p}{n}\right).
\end{aligned}$$

By $n\rho_0 = n - (2p+5)/6$, we obtain

$$\begin{aligned} f_0 + n\rho_0 \times \mu_{n,0} &= f_0 + n \times \mu_{n,0} - p \times \mu_0/3 + o(p) \\ &= f_0 + \frac{p - p^2}{2} - \frac{p^3}{6n} - \frac{p^4}{12n^2} + O\left(\frac{p^5}{n^3}\right) + o(p) + \frac{p^3}{6n} + \frac{p^4}{18n^2} \\ &= -\frac{p^4}{36n^2} + O\left(\frac{p^5}{n^3}\right) + o(p). \end{aligned}$$

Given that (12) holds under this case and $\sqrt{2f_0}/p \to 1$, we obtain $(f_0 + n\rho_0\mu_{n,0}) \times (n\rho_0\sigma_{n,0})^{-1} = -p^3/(36n^2) + O(p^4/n^3) + o(1)$, which converges to 0 if and only if $p^3/n^2 \to 0$ under this case.

Case (ii.2): $\lim_{n\to\infty} p/n = C \in (0, 1]$. Under this case, we show that (12) does not hold. Note that

$$\frac{2f_0}{n^2\rho_0^2\sigma_{n,0}^2} \to \frac{C^2}{-2(1-C/3)^2\{C + \log(1-C)\}}. \tag{14}$$

If $C = 1$, $2f_0/(n^2\rho_0^2\sigma_{n,0}^2) \to 0$ and thus (12) does not hold. We next consider $C \in (0, 1)$. If (12) holds, we shall have $g_2(C) = 0$ with $g_2(C) = C^2 + 2(1-C/3)^2\{C + \log(1-C)\}$. By taking derivative of $g_2(C)$, we obtain $g_2'(0) = 0$, $g_2''(0) = 0$, and

$$g_2'''(C) = -\frac{4C(3C^2 - 8C + 9)}{9(1-C)^3} < 0$$

when $C \in (0, 1)$. Similarly to the analysis in Case (i.1), we obtain that $g_2'(C) < 0$ for $C \in (0, 1)$. It follows that $g_2(C)$ is strictly decreasing on $C \in (0, 1)$ with $g_2(0) = 0$. Therefore $g_2(C) < 0$ on $C \in (0, 1)$, which suggests that (12) does not hold.

Finally, for a general sequence $p/n \in (0, 1]$, following the analysis of taking subsequences in Part (i), we know that the chi-square approximation with the Bartlett correction holds if and only if $p^3/n^2 \to 0$. Recall that $N = n + 1$. Thus, the same conclusions hold asymptotically by replacing n with N, that is, the chi-square approximations without and with the Bartlett correction hold if and only if $p^2/N \to 0$ and p^3/N^2, respectively.

A.2 Proof of Theorem 2

Similarly to the proof of Theorem 1, we next examine the limiting distribution of T' based on its moment generating function. In Theorem 2, testing $H'_{0,k} : \Sigma = \Lambda_k \Lambda_k^\top + \Psi_k$ when Λ_k and Ψ_k are given is equivalent to testing the null hypothesis $H_0 : \Sigma = \mathrm{I}_p$ by applying the data transformation $\Sigma_k^{-1/2} X_i$ with $\Sigma_k = \Lambda_k \Lambda_k^\top + \Psi_k$. Then by Corollary 8.4.8 in Muirhead (2009), under the null hypothesis, we have

$$\mathrm{E}\{\exp(h \times T')\} = \left(\frac{2e}{n}\right)^{-pnh} (1-2h)^{-pn(1-2h)/2} \times \frac{\Gamma_p\{n(1-2h)/2\}}{\Gamma_p(n/2)}, \tag{15}$$

where $n = N - 1$.

Part (i) The chi-square approximation. When p is fixed compared to the sample size N, by applying Stirling's approximation to the Gamma function, it has been shown that as $N \to \infty$, (15) converges to $(1-2h)^{-f'/2}$, which is the moment generating function of $\chi^2_{f'}$ (Muirhead, 2009, Section 8.4.4), and therefore $T' \xrightarrow{D} \chi^2_{f'}$. When $p \to \infty$, by the proof of Lemma 3 in Section A.3, we have $\mathrm{E}[\exp\{h(T' + n\mu_n)/(n\sigma_n)\}] \to \exp(h^2/2)$, where

$$\mu_n = -p + (p - n + 1/2)\log\left(1 - \frac{p}{n}\right), \quad \sigma_n^2 = -2\left\{\frac{p}{n} + \log\left(1 - \frac{p}{n}\right)\right\}. \tag{16}$$

Similarly to the proof of Theorem 1, we know that the chi-square approximation for T' holds if and only if

$$\sqrt{2f'} \times (n\sigma_n)^{-1} \to 1, \tag{17}$$

$$(f' + n\mu_n) \times (n\sigma_n)^{-1} \to 0. \tag{18}$$

Case (i.1): $\lim_{n\to\infty} p/n = 0$. Under this case, similar to (11), by Taylor's expansion, $\sigma_n^2 = p^2 n^{-2}\{1 + o(1)\}$. As $\sqrt{2f'}/p \to 1$, we have (17) holds. We next show that (18) holds if and only if

$p^2/n \to 0$. Particularly, by Taylor's expansion and $p/n \to 0$,

$$\begin{aligned} \mu_n &= -p + (-n + p + 1/2)\left\{-\frac{p}{n} - \frac{p^2}{2n^2} - \frac{p^3}{3n^3} + O\left(\frac{p^4}{n^4}\right)\right\} \\ &= -p + p + \frac{p^2}{2n} + \frac{p^3}{3n^2} - \frac{p^2}{n} - \frac{p^3}{2n^2} - \frac{p}{2n} + O\left(\frac{p^4}{n^3}\right) + o\left(\frac{p}{n}\right). \end{aligned} \tag{19}$$

It follows that $f' + n\mu_n = -p^3/(6n^2) + O(p^4/n^3) + o(p/n)$. Given (17) and $\sqrt{2f'}/p \to 1$, (18) holds if and only if $p^2/n \to 0$.

Case (i.2): $\lim_{n\to\infty} p/n = C \in (0, 1]$. Under this case, $2f'/(n^2\sigma_n^2) \to -C^2/[2\{C + \log(1 - C)\}]$. We then know (17) does not hold following the proof of Theorem 1, and therefore the chi-square approximation fails.

Finally, for a general sequence $p/n \in (0, 1]$, following the same analysis of taking subsequences as in the proof of Theorem 1, we know that the chi-square approximation holds if and only if $p^2/n \to 0$.

Part (ii) The chi-square approximation with the Bartlett correction. Similarly to the proof of Theorem 1 and the analysis above, we know that the chi-square approximation with the Bartlett correction holds if and only if

$$\sqrt{2f'} \times (n\rho' \times \sigma_n)^{-1} \to 1, \tag{20}$$

$$(f' + n\rho' \times \mu_n) \times (n\rho' \times \sigma_n)^{-1} \to 0. \tag{21}$$

Case (ii.1): $\lim_{n\to\infty} p/n = 0$. As $\rho' \to 1$ under this case, we know (20) holds given (17) proved in Part (i). We next prove (21) holds if and only if $p^3/n^2 \to 0$. Similarly to (19), by Taylor's expansion and $p/n \to 0$,

$$\begin{aligned} \mu_n &= -p + (-n + p + 1/2)\left\{-\sum_{j=1}^{4} \frac{p^j}{j \times n^j} + O\left(\frac{p^5}{n^5}\right)\right\} \\ &= -\frac{p(p+1)}{2n} - \frac{p^3}{6n^2} - \frac{p^4}{12n^3} + O\left(\frac{p^5}{n^4}\right) + o\left(\frac{p}{n}\right). \end{aligned}$$

By $n\rho' = n - p/3 + O(1)$ and $p/n \to 0$,

$$\begin{aligned} n\rho'\mu_n &= \{n - p/3 + O(1)\}\left\{-\frac{p(p+1)}{2n} - \frac{p^3}{6n^2} - \frac{p^4}{12n^3} + O\left(\frac{p^5}{n^4}\right) + o\left(\frac{p}{n}\right)\right\} + o(p) \\ &= -\frac{p(p+1)}{2} - \frac{p^3}{6n} - \frac{p^4}{12n^2} + \frac{p^3}{6n} + \frac{p^4}{18n^2} + O\left(\frac{p^5}{n^3}\right) + o(p). \end{aligned}$$

It follows that $f' + n\rho'\mu_n = -p^4/(36n^2) + O(p^4/n^3) + o(p)$. Given (13) and $\sqrt{2f'}/p \to 1$, (21) holds if and only if $p^3/n^2 \to 0$.

Case (ii.2): $\lim_{n\to\infty} p/n = C \in (0, 1]$. Under this case, $\rho' \to 1 - C/3$ and $2f'/(n\rho'\sigma_n)^2$ converges to the limit same as the right hand side of (14). Thus the same analysis applies and we know that the chi-square approximation with the Bartlett correction fails.

Finally, for a general sequence $p/n \in (0, 1]$, following the same analysis of taking subsequences as in the proof of Theorem 1, we know that the chi-square approximation with the Bartlett correction holds if and only if $p^3/n^2 \to 0$. Recall that $N = n + 1$. Thus, the same conclusions hold asymptotically by replacing n with N, that is, the chi-square approximations without and with the Bartlett correction hold if and only if $p^2/N \to 0$ and p^3/N^2, respectively.

A.3 Lemma

Lemma 3 *Under the conditions of Theorem 2, when $p \to \infty$ as $n = N - 1 \to \infty$, we have $(T' + n\mu_n)/(n\sigma_n) \xrightarrow{D} \mathcal{N}(0, 1)$ with μ_n and σ_n^2 in (16).*

Proof. It suffices to show that there exists a constant $\delta' > 0$ such that $\mathrm{E}[\exp\{h(T'+n\mu_n)/(n\sigma_n)\}] \to \exp(h^2/2)$ for all $|h| < \delta'$. Particularly, we let $s = h/(n\sigma_n)$, and prove $\log[\mathrm{E}\{\exp(sT')\}] \to h^2/2 - h\mu_n/\sigma_n$. By the moment generating function of T' in (15), we have

$$\begin{aligned} & \log\left[\mathrm{E}\{\exp(s \times T')\}\right] \qquad (22) \\ = \; & -pns\log(2e/n) - \frac{pn}{2}(1-2s)\log(1-2s) + \log\left\{\frac{\Gamma_p(n/2 - ns)}{\Gamma_p(n/2)}\right\}. \end{aligned}$$

We next derive the approximate expansion of (22) by discussing two cases.

Case 1: $\lim p/n \to C \in (0, 1]$. Under this case, we utilize the approximate expansion of multivari-

ate gamma function in Lemma 5.4 of Jiang and Yang (2013). To apply the result, we first show that the conditions are satisfied. Specifically, define $r_n^2 = -\log(1-p/n)$, and we have

$$(-ns)^2 \times r_n^2 = -\frac{h^2}{\sigma_n^2}\log(1-p/n) \to \begin{cases} \dfrac{h^2}{2} \times \dfrac{\log(1-C)}{C+\log(1-C)}, & \text{if } C \in (0,1); \\ \dfrac{h^2}{2}, & \text{if } C = 0. \end{cases}$$

Therefore, $-ns = O(1/r_n)$, and then Lemma 5.4 in Jiang and Yang (2013) can be applied to expand (22). It follows that

$$\begin{aligned} (22) &= -pns\log(2e/n) - \frac{pn}{2}(1-2s)\log(1-2s) \\ &\quad -pns\log\{n/(2e)\} + r_n^2\left\{(-ns)^2 - (p-n+1/2)(-ns)\right\} + o(1). \end{aligned}$$

By Taylor's expansion $(1-2s)\log(1-2s) = -2s + 2s^2 + O(s^3)$ for $s \in (0,1)$, we obtain

$$\begin{aligned} (22) &= -\frac{pn}{2}\left\{-2s + 2s^2 + O(s^3)\right\} \\ &\quad -\log\left(1-\frac{p}{n}\right)\left\{n^2s^2 + (p-n+1/2)ns\right\} + o(1) \\ &= s^2\left\{-pn - n^2\log\left(1-\frac{p}{n}\right)\right\} + s\left\{pn - (p-n+1/2)\log\left(1-\frac{p}{n}\right)\right\} + o(1). \end{aligned}$$

With $s = h/(n\sigma_n)$, we have $\log(\mathrm{E}[\exp\{hT'/(n\sigma_n)\}]) = h^2/2 - h\mu_n/\sigma_n + o(1)$.

Case 2: $\lim p/n = 0$. Under this case, we utilize the approximate expansion of multivariate gamma function in Proposition A.1 of Jiang and Qi (2015). To apply the result, we first show that the conditions are satisfied. Particularly, as $\sigma_n^2 = p^2n^{-2}\{1+o(1)\}$, we have $-ns \times p/n = -ph(n\sigma_n)^{-1} = h\{1+o(1)\}$. Therefore, $-ns = O(n/p)$, and we can apply Proposition A.1 in Jiang and Qi (2015) to expand (22). It follows that

$$\log\left\{\frac{\Gamma_p(n/2-ns)}{\Gamma_p(n/2)}\right\} = \gamma_{n,1}(-ns) + \gamma_{n,2}(-ns)^2 + \gamma_{n,3} + o(1),$$

where

$$\begin{aligned}
\gamma_{n,1} = & -\{2p + (n - p - 1/2)\log(1 - p/n)\}, \\
\gamma_{n,2} = & -\{p/n + \log(1 - p/n)\}, \\
\gamma_{n,3} = & \, p\{(n/2 - ns)\log(n/2 - ns) - (n/2)\log(n/2)\}.
\end{aligned}$$

Note that $\gamma_{n,3} = (pn/2)(1-2s)\log(1-2s) - pns\log(n/2)$. Then we have

$$\begin{aligned}
(22) = & -pns\log\left(\frac{2e}{n}\right) - \frac{pn}{2}(1-2s)\log(1-2s) - \gamma_{n,1}ns + \gamma_{n,2}n^2s^2 + \gamma_{n,3} + o(1) \\
= & -(p + \gamma_{n,1})ns + \gamma_{n,2}n^2s^2 + o(1),
\end{aligned}$$

which gives $\log(\mathrm{E}[\exp\{hT'/(n\sigma_n)\}]) = h^2/2 + \mu_n h/\sigma_n + o(1)$ by $s = h/(n\sigma_n)$.

Finally, for a general sequence $\{p/n\}$, to prove that $(T' + n\mu_n)/(n\sigma_n)$ converges in distribution to $\mathcal{N}(0,1)$, it suffices to show that every subsequence has a further subsequence that converges in distribution to $\mathcal{N}(0,1)$. By the boundedness of p/n and the Bolzano-Weierstrass theorem, we can further take a subsequence such that p/n has a limit and the arguments above can be applied. In summary, Lemma 3 is proved. □

B Supplementary simulation studies

B.1 Simulations on the Type I Error

In this section, we provide additional simulation studies when the data is not normally distributed. Particularly, we focus on the likelihood ratio test under the null hypothesis $H_{0,0}$, which detects the existence of any factors or not.

Simulations with heavy-tailed *t*-distributed data. Similarly to previous simulations, we consider $p = \lfloor N^{\varepsilon} \rfloor$, where $N \in \{100, 500, 1000, 2000\}$ and $\varepsilon \in \{3/24, 4/24, \ldots, 23/24\}$. Under each combination of (N, p), we generate the entries of data matrix X_i as independent and identical

random variables following t_{d_0} distribution, where d_0 denotes the degrees of freedom and we take $d_0 \in \{5, 10\}$. Then we conduct the likelihood ratio test for $H_{0,0}$ with approximations (3) and (4). We repeat the procedure 1000 times, and estimate the type I error rates with significance level 0.05. We present the results of t_5 and t_{10} distributed data in Figures 5 and 6, respectively. In each figure, we draw the estimated type I error rates versus ε values for approximations (3) and (4) in the left and right plots, respectively. Similarly to Numerical Example 2, we can see that the chi-square approximation for T_0 starts to fail when ε approaches 1/2, and the chi-square approximation for $\rho_0 T_0$ starts to fail when ε approaches 2/3.

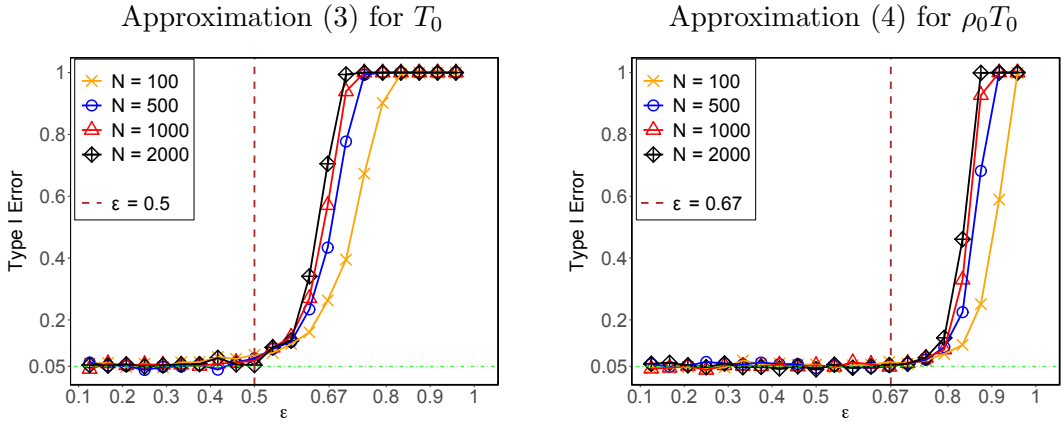

Figure 5: Estimated type I error versus ε of t_5-distributed data

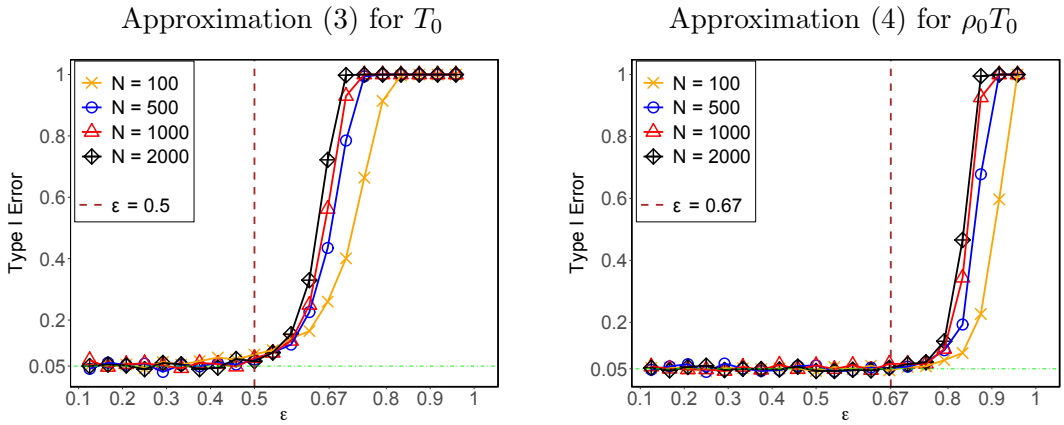

Figure 6: Estimated type I error versus ε of t_{10}-distributed data

Setting (I)

$z_{i,j}$	$(-\infty, 0)$	$[0, \infty)$
$x_{i,j}$	-1	1

Setting (II)

$z_{i,j}$	$(-\infty, -1)$	$[-1, 0)$	$[0, 1)$	$[1, \infty)$
$x_{i,j}$	-2	-1	1	2

Setting (III)

$z_{i,j}$	$(-\infty, -1)$	$[-1, -0.4)$	$[-0.4, 0)$	$[0, 0.4)$	$[0.4, 1)$	$[1, \infty)$
$x_{i,j}$	-3	-2	-1	1	2	3

Table 1: Three settings of correspondence between $x_{i,j}$ and $z_{i,j}$

Simulations with discrete multinomial data. The simulations are conducted same as above, except that we generate the entries in the data matrix X_i from a discrete multinomial distribution. Specifically, for each entry $x_{i,j}$ within the matrix X_i, where $i = 1, \ldots, N$ and $j = 1, \ldots, p$, we first sample $z_{i,j} \sim \mathcal{N}(0, 1)$, and then set discrete value of $x_{i,j}$ according to the range of $z_{i,j}$ considering three settings (I)–(III) in Table 1. The results under settings (I)–(III) are given in Figures 7–9, respectively. Similarly to Numerical Example 2, under each setting, we observe that the chi-square approximation (3) for T_0 starts to fail when ε approaches 1/2, and the chi-square approximation (4) for $\rho_0 T_0$ starts to fail when ε approaches 2/3.

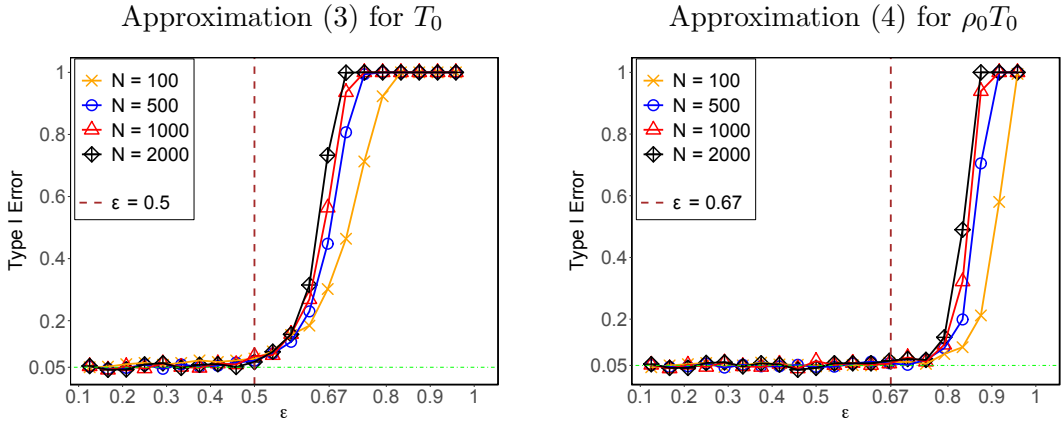

Figure 7: Discrete data (I): Estimated type I error versus ε

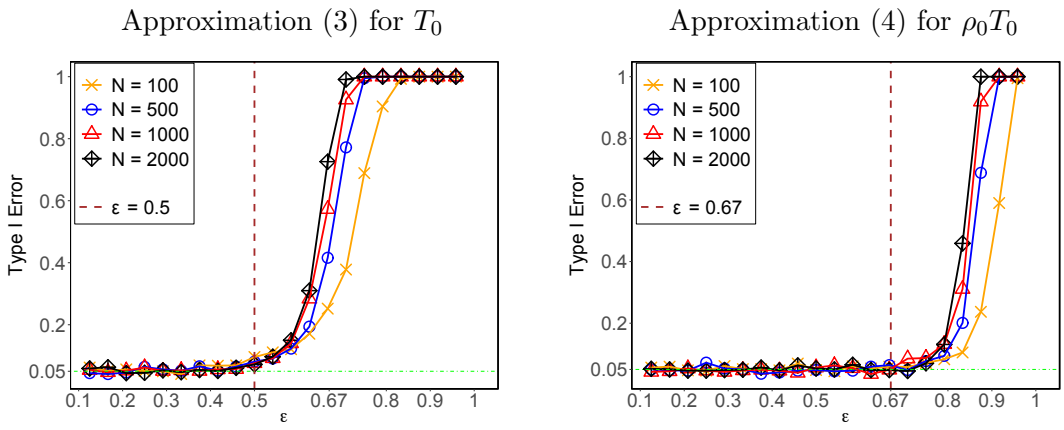

Figure 8: Discrete data (II): Estimated type I error versus ε

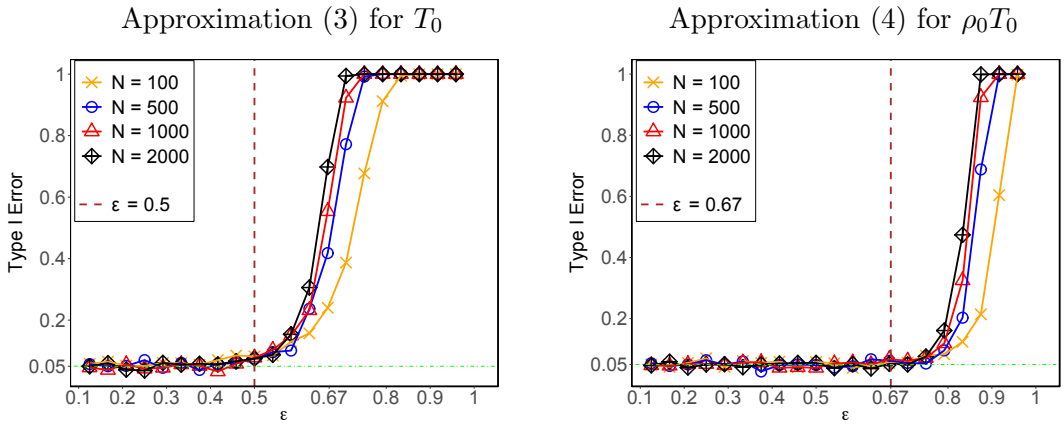

Figure 9: Discrete data (III): Estimated type I error versus ε

B.2 Simulations on Estimating the Number of Factors

In this section, we demonstrate the performance of estimating the number of factors using the sequential procedure described in Section 2.1. In particular, we consider the simulation setting similar to that in Numerical Example 3, where we take the true number of factors $k_0 \in \{1, 3\}$, sample size $N \in \{500, 1000\}$ and data dimension $p = \lfloor N^{\epsilon} \rfloor$ for different ϵ values. When conducting the likelihood ratio tests in the sequential procedure, the nominal significance level is set as $\alpha = 0.05$. For each combination of (k_0, N), we use the sequential procedure to estimate the number of factors, denoted as $\hat{k}$. We repeat the procedure 1000 times and estimate the proportions of correct estimation ($\hat{k} = k_0$) and overestimation ($\hat{k} > k_0$), respectively. We present the results for $k_0 = 1, 3$ in Figures 10 and 11 , respectively, where the results based on the likelihood ratio test without and

with the Bartlett correction are given in the left and right columns, respectively.

The numerical results in Figures 10 and 11 show that (I) using the likelihood ratio test, the procedure begins to overestimate the number of factors when ϵ approaches 1/2; (II) using the likelihood ratio test with the Bartlett correction, the procedure begins to overestimate the number of factors when ϵ approaches 2/3. These observations, compared with Figures 2–4, suggest that the sequential procedure begins to overestimate the number of factors when the corresponding type I error begins to inflate, which is consistent with our discussions in Section 2.2. Moreover, in Figures 10 and 11, when ϵ is small and does not pass the corresponding phase transition boundary, the proportion of overestimation ($\hat{k} > k_0$) is around 0.05. This is because that rejecting H_{0,k_0} suggests $\hat{k} > k_0$, and the probability of rejecting H_{0,k_0} (type I error of testing H_{0,k_0}) can be asymptotically controlled at the level $\alpha = 0.05$ under the asymptotic regimes derived in Theorems 1 and 2.

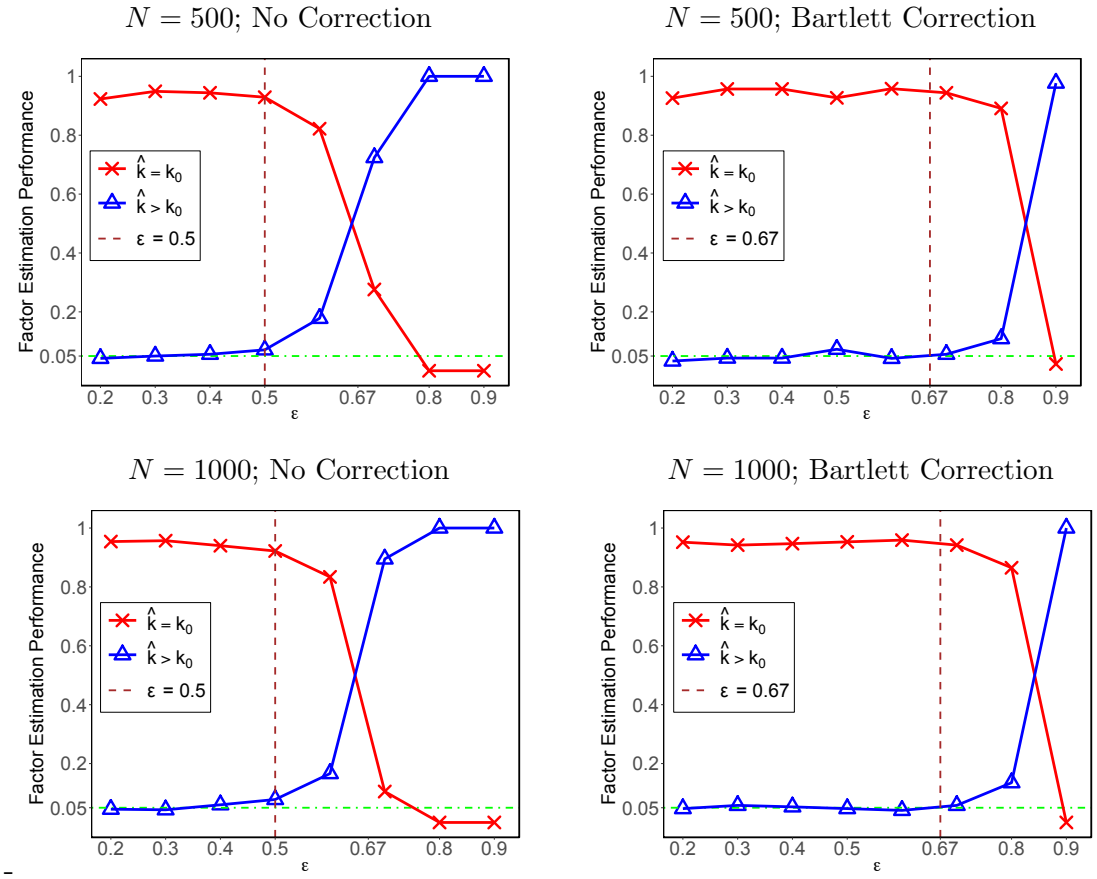

Figure 10: Estimating the number of factors when $k_0 = 1$

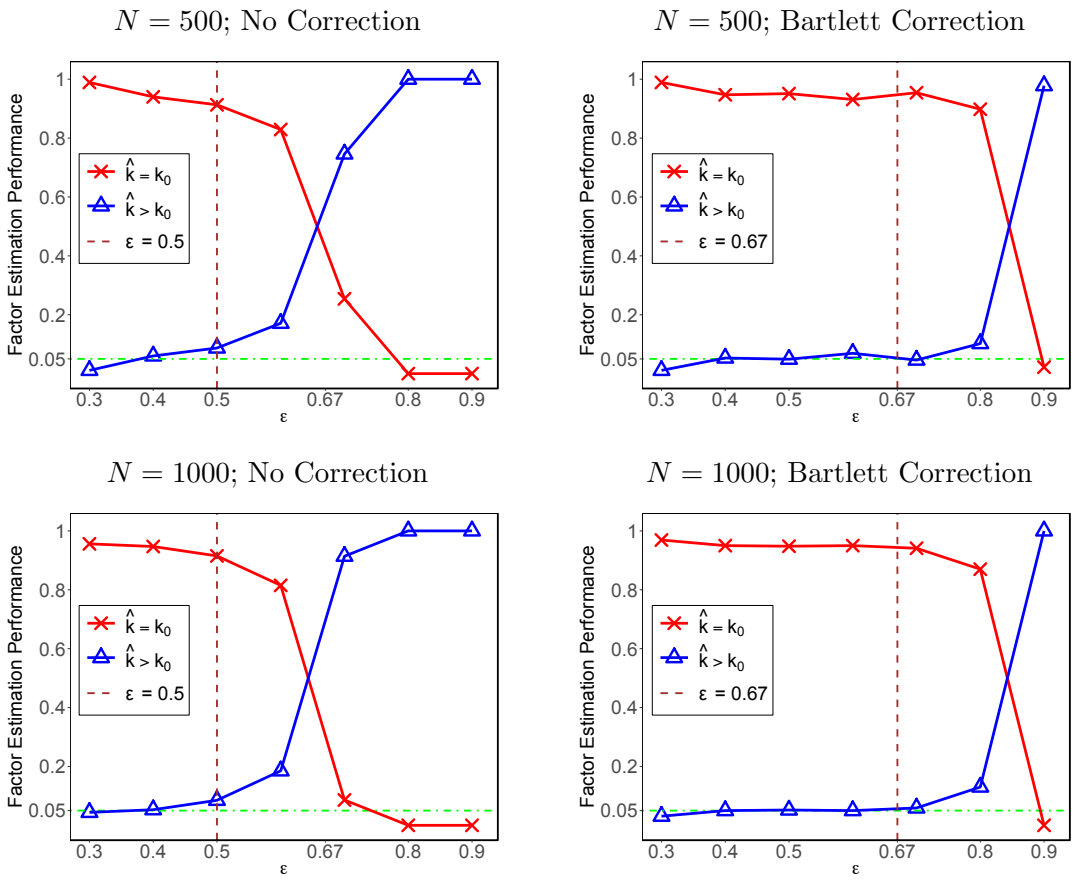

Figure 11: Estimating the number of factors when $k_0 = 3$