Higher Moment Estimation for Elliptically-distributed Data: Is it Necessary to Use a Sledgehammer to Crack an Egg?

Zheng Tracy Ke[*], Koushiki Bose[†], Jianqing Fan[‡]

Abstract

Multivariate elliptically-contoured distributions are widely used for modeling economic and financial data. We study the problem of estimating moment parameters of a semi-parametric elliptical model in a high-dimensional setting. Such estimators are useful for financial data analysis and quadratic discriminant analysis.

For low-dimensional elliptical models, efficient moment estimators can be obtained by plugging in an estimate of the precision matrix. Natural generalizations of the plug-in estimator to high-dimensional settings perform unsatisfactorily, due to estimating a large precision matrix. Do we really need a sledgehammer to crack an egg? Fortunately, we discover that moment parameters can be efficiently estimated without estimating the precision matrix in high-dimension.

We propose a marginal aggregation estimator (MAE) for moment parameters. The MAE only requires estimating the diagonal of covariance matrix and is convenient to implement. With mild sparsity on the covariance structure, we prove that the asymptotic variance of MAE is the same as the ideal plug-in estimator which knows the true precision matrix, so MAE is asymptotically efficient. We also extend MAE to a block-wise aggregation estimator (BAE) when estimates of diagonal blocks of covariance matrix are available. The performance of our methods is validated by extensive simulations and an application to financial returns.

1 Introduction

The classical multivariate statistics is largely motivated by relaxing the Gaussian assumption, which is not satisfied in many applications. There is an extensive literature in finance on the tail-index estimates of stock returns; while being unimodal and symmetric, the empirical returns exhibit leptokurtosis, which means that they have heavier tails and flatter peaks than those of normal data

[*]Department of Statistics, Harvard University, Cambridge, MA 02138.

[†]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544.

[‡]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544. Fan's research is supported by NSF grants DMS-1712591 and DMS-1662139 and NIH grant R01-GM072611.

(Fama, 1965; Bollerslev and Wooldridge, 1992; Eberlein and Keller, 1995; Frahm et al., 2003; Cizek et al., 2005). Empirical evidence of the violation of Gaussian assumption has also been observed in genomics (Liu et al., 2003; Posekany et al., 2011; Hardin and Wilson, 2009) and in bioimaging (Ruttimann et al., 1998). The family of multivariate elliptically contoured distributions (Kelker, 1970), which we shall call elliptical distributions in short, provides a natural generalization of multivariate Gaussian distributions. Recently, many statistical methods for elliptically distributed data have been proposed, including works on covariance matrix estimation (Fan et al., 2018), graphical modeling (Han and Liu, 2012), classification (Fan et al., 2015b), etc.

The elliptical distributions are typically used as a semi-parametric model. Given a mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}} \in \mathbb{R}^p$, a covariance matrix $\boldsymbol{\Sigma} = (\sigma_{jk})_{1 \le j,k \le p} \in \mathbb{R}^{p \times p}$ and a probability characteristic function $\phi : [0, \infty) \to \mathbb{R}$, we say a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^{\mathrm{T}}$ has an elliptical distribution $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ if

$$\boldsymbol{Y} = \boldsymbol{\mu} + \xi \, \boldsymbol{\Sigma}^{1/2} \boldsymbol{U}, \tag{1}$$

where $\boldsymbol{U}$ is a random vector that is uniformly distributed on the unit sphere $\mathbb{S}^{p-1}$, and independent of $\boldsymbol{U}$, ξ is a nonnegative random variable whose characteristic function is ϕ. For model identifiability, we normalize ξ such that

$$\mathbb{E}(\xi^2) = p. \tag{2}$$

Under (1)-(2), $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and covariance matrix of $\boldsymbol{Y}$, respectively. The variable ξ determines which sub-family the distribution belongs to. When ξ^2 is a chi-square random variable, it belongs to the multivariate Gaussian sub-family, and when ξ^2 follows an F-distribution, it belongs to the multivariate t sub-family or multivariate Cauchy sub-family. For most applications, the sub-family of the elliptical distribution is unknown, leaving the distribution of ξ unspecified.

Although full knowledge of the distribution of ξ is often not required, an estimate of its moment parameters is useful to statistical analysis and for understanding the tail of the distributions. One application is in quadratic classification. When data from two classes both follow elliptical distributions but have unequal covariance matrices, Fan et al. (2015b) showed that an estimate of $\mathbb{E}(\xi^4)$ is desired for building a quadratic classifier. Another application is to capture the tail behavior of financial returns by estimating the leptokurtosis. Modeling the returns of a set of financial assets by an elliptical distribution, the leptokurtosis equals to $\{p(p+2)\}^{-1}\mathbb{E}(\xi^4) - 1$, so the problem reduces to estimating $\mathbb{E}(\xi^4)$.

For any $m \ge 1$, define the m-th scaled even moment of ξ by

$$\theta_m \equiv p^{-m}\mathbb{E}(\xi^{2m}). \tag{3}$$

The first scaled even moment θ_1 is 1. In this paper, we are interested in estimating θ_m for any fixed $m \ge 2$, given independent and identically distributed (i.i.d.) samples $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ from (1).

1.1 The plug-in estimators

We consider an ideal case where $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are known. Given *iid* samples $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ from an unknown elliptical distribution, each $\boldsymbol{Y}_i$ has a decomposition $\boldsymbol{Y}_i = \boldsymbol{\mu} + \xi_i \boldsymbol{\Sigma}^{1/2} \boldsymbol{U}_i$, and $\xi_1, \ldots, \xi_n$ are *iid* copies of ξ. Using the fact that $\boldsymbol{U}_i$ takes values on the unit sphere, we observe $\xi_i^2 = (\boldsymbol{Y}_i - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Omega} (\boldsymbol{Y}_i - \boldsymbol{\mu})$ for $i = 1, \ldots, n$, where $\boldsymbol{\Omega} \equiv \boldsymbol{\Sigma}^{-1}$. Hence, in the ideal case, $\xi_1, \ldots, \xi_n$ are directly observed. It motivates the following estimator of θ_m:

$$\widehat{\theta}_m^{\mathrm{I}}(\boldsymbol{\mu}, \boldsymbol{\Omega}) = \frac{1}{np^m} \sum_{i=1}^{n} (\xi_i^2)^m = \frac{1}{np^m} \sum_{i=1}^{n} \{(\boldsymbol{Y}_i - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Omega} (\boldsymbol{Y}_i - \boldsymbol{\mu})\}^m. \tag{4}$$

We call $\widehat{\theta}_m^{\mathrm{I}}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ the *Ideal Estimator*. The ideal estimator is not feasible in practice, and a natural modification is to plug in estimates of $(\boldsymbol{\mu}, \boldsymbol{\Omega})$. This gives rise to the plug-in estimator:

$$\widehat{\theta}_m^{\mathrm{I}}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Omega}}) = \frac{1}{np^m} \sum_{i=1}^{n} \{(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}})^{\mathrm{T}} \widehat{\boldsymbol{\Omega}} (\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}})\}^m, \tag{5}$$

This estimator was proposed by Maruyama and Seo (2003) in the setting of a fixed dimension, where they used the sample mean to estimate $\boldsymbol{\mu}$ and the inverse of the sample covariance matrix to estimate $\boldsymbol{\Omega}$. In the modern high-dimensional settings where p grows with n, one can on longer use the inverse of sample covariance matrix to estimate $\boldsymbol{\Omega}$; Fan et al. (2015b) proposed plugging in an estimator of $\boldsymbol{\Omega}$ from high-dimensional sparse precision matrix estimation methods, with stringent structural assumptions on $\boldsymbol{\Omega}$.

However, the plug-in estimators perform unsatisfactorily for high-dimensional settings due to the difficulty of estimating $\boldsymbol{\Omega}$. Existing methods of estimating $\boldsymbol{\Omega}$ only perform well under stringent conditions, such as the sub-Gaussian assumption on the distribution and/or structural assumptions on $\boldsymbol{\Omega}$ (e.g., sparsity). Especially, the structural assumption on $\boldsymbol{\Omega}$ is critical for the success of these methods. Figure 1 shows the performance of the plug-in estimator when the structural assumption required by $\widehat{\boldsymbol{\Omega}}$ is violated. We consider two estimators of $\boldsymbol{\Omega}$, the CLIME estimator (Cai et al., 2011) which requires sparsity of $\boldsymbol{\Omega}$, and the POET estimator (Fan et al., 2013) which assumes a factor structure with sparse covariance of the idiosyncratic component. On the left panel of Figure 1, we generate elliptical data with a sparse covariance matrix, $\boldsymbol{\Sigma}_{i,j} = a^{|i-j|}$, $1 \le i, j \le p$, where a controls the sparsity level and varies in $\{0.5, 0.55, \ldots, 0.85, 0.9\}$. Here, the structural assumption of POET is not satisfied, and the associated plug-in estimator of θ_2 performs unsatisfactorily. On the right panel, we generate data with a sparse precision matrix $\boldsymbol{\Omega}$, where each entry of the upper triangle of $\boldsymbol{\Omega}$ has a probability of a to be nonzero,[1] with a chosen from $\{0.5, 0.55, \ldots, 0.85, 0.9\}$.

[1] We generate $\boldsymbol{\Omega}$ using *fastclime.generator*$(\cdot)$ in the R package *clime*, where the *graph* argument is set "random".

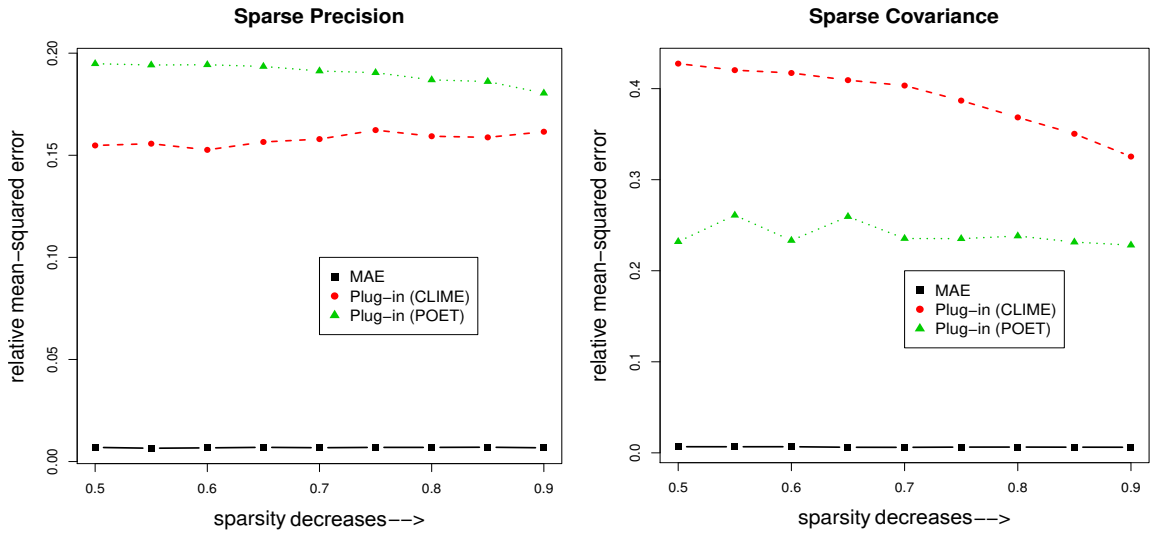

Figure 1: Plug-in estimators and MAE ($p = 100$, $n = 20$, true distribution is multivariate Gaussian). For plug-in estimators, we use two estimators of $\mathbf{\Omega}$, CLIME (Cai et al., 2011) and POET (Fan et al., 2013). CLIME requires $\mathbf{\Omega}$ to be sparse and POET assumes $\mathbf{\Sigma}$ has a low-rank plus sparse structure. When these structural assumptions are violated, the plug-in estimator of θ_2 has a poor performance (y-axis is log of squared errors). In contrast, MAE always outperforms the plug-in estimators.

The assumption of CLIME is violated, so the associated plug-in estimator of θ_2 has a unsatisfactory performance.

In fact, the philosophy of plug-in estimators is problematic. Estimating large precision matrices is a well-known difficult problem (even for Gaussian data), as one needs to estimate a large number of parameters. On the other hand, our problem only involves estimating one single parameter θ_m. Intuitively, the latter should be much easier than the former. The plug-in estimators are realy using "a sledgehammer to crack an egg."

1.2 The marginal aggregation estimator (MAE)

Is it possible to avoid using the "sledgehammer" of precision matrix estimation? We show that this is possible by a new marginal aggregation estimator. In model (1), letting $\widetilde{U}_j$ be the j-th coordinate of $\widetilde{\boldsymbol{U}} \equiv \mathbf{\Sigma}^{1/2}\boldsymbol{U}$, we have

$$Y_j = \mu_j + \xi\, \widetilde{U}_j, \qquad j = 1, \ldots, p. \tag{6}$$

Our key observation is that each individual coordinate of $\boldsymbol{Y}$ contains information of ξ. It motivates us to construct an estimator of θ_m using only one coordinate of samples. Let σ_{jj} be the j-th diagonal of $\mathbf{\Sigma}$. We notice that (6) implies $\xi^{2m} = (Y_j - \mu_j)^{2m}/\widetilde{U}_j^{2m}$. The random variable $\widetilde{U}_j$ is unobserved,

but its distribution is known once σ_{jj} is given. It can be shown that (see Proposition 3.1)

$$\mathbb{E}(\widetilde{U}_j^{2m}) = p^{-m} c_m \, \sigma_{jj}^m, \qquad \text{where} \quad c_m = (2m-1)!! \, (p/2)^m \frac{\Gamma(p/2)}{\Gamma(p/2+m)}. \tag{7}$$

Inspired by (6)-(7), we introduce an estimator of θ_m using the marginal data $Y_{1j}, \ldots, Y_{nj}$:

$$\widehat{\theta}_{m,j}^{\mathrm{M}}(\mu_j, \sigma_{jj}) = \frac{1}{np^m} \sum_{i=1}^{n} \frac{(Y_{ij}-\mu_j)^{2m}}{\mathbb{E}(\widetilde{U}_j^{2m})} = \frac{1}{c_m \, \sigma_{jj}^m} \frac{1}{n} \sum_{i=1}^{n} (Y_{ij}-\mu_j)^{2m}. \tag{8}$$

We call $\widehat{\theta}_{m,j}^{\mathrm{M}}(\mu_j, \sigma_{jj})$ the *Marginal Estimator*. It only requires knowledge of (μ_j, σ_{jj}) and successfully avoids precision matrix estimation. For each $1 \le j \le p$, we can define a marginal estimator and we will show that all marginal estimator contains the same amount of information about θ_m (see Theorem 2.4). All these marginal estimators are unbiased, so taking their average gives rise to a new unbiased estimator:

$$\widehat{\theta}_m^{\mathrm{M}}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\Sigma})) = \frac{1}{p} \sum_{j=1}^{p} \widehat{\theta}_{m,j}^{\mathrm{M}}(\mu_j, \sigma_{jj}) = \frac{1}{c_m \, np} \sum_{j=1}^{p} \left\{ \frac{1}{\sigma_{jj}^m} \sum_{i=1}^{n} (Y_{ij}-\mu_j)^{2m} \right\}. \tag{9}$$

We call $\widehat{\theta}_m^{\mathrm{M}}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\Sigma}))$ the *Marginal Aggregation Estimator (MAE)*. The "aggregation" of marginal estimators helps reduce the asymptotic variance. Our proposed estimator is a natural plug-in version of (9) given by

$$\widehat{\theta}_m^{\mathrm{M}}(\widehat{\boldsymbol{\mu}}, \mathrm{diag}(\widehat{\boldsymbol{\Sigma}})) = \frac{1}{c_m \, np} \sum_{j=1}^{p} \left\{ \frac{1}{\widehat{\sigma}_{jj}^m} \sum_{i=1}^{n} (Y_{ij}-\widehat{\mu}_j)^{2m} \right\}, \tag{10}$$

where c_m is as in (7), $\widehat{\boldsymbol{\mu}}$ is an estimator of $\boldsymbol{\mu}$, and $\{\widehat{\sigma}_{jj}\}_{j=1}^p$ are the estimators of $\{\sigma_{jj}\}_{j=1}^p$.

Compared with the plug-in estimator (5), MAE is numerically more appealing, as it only needs to estimate the diagonal entries of $\boldsymbol{\Sigma}$. Back to the example in Figure 1, we implement MAE using sample mean as $\widehat{\boldsymbol{\mu}}$ and sample covariance matrix as $\widehat{\boldsymbol{\Sigma}}$. MAE significantly outperforms the plug-in estimators, even when the structural assumptions of the plug-in estimators are satisfied.

1.3 Organization of the paper

In Section 2, we study the theoretical properties of MAE. Under mild regularity conditions, we show that MAE is unbiased and root-n consistent, regardless of the structure of $\boldsymbol{\Sigma}$. We also show that MAE is asymptotically efficient, with an asymptotic variance matching that of the ideal estimator when $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are given. We also discuss how to construct a confidence interval of θ_m.

In Section 3, we generalize the idea of MAE to develop estimators of θ_m that use a small subset of the coordinates. We introduce the block-wise estimator and the blockwise aggregation estimator (BAE), analogous to the marginal estimator and MAE. These ideas help further reduce the estimation errors in the second order.

Section 4 validates the theoretical insight by extensive simulations. Section 5 gives an application of MAE to time series data. We consider an extension of model (1) to multivariate time series:

$$\boldsymbol{Y}_t = \boldsymbol{\mu}_t + \mathbf{B}\boldsymbol{f}_t + \xi_t \boldsymbol{\Sigma}_t^{1/2} \boldsymbol{U}_t, \qquad t = 1, \cdots, T,$$

where $\boldsymbol{\mu}_t$ is the time-varying mean, $\boldsymbol{f}_t \in \mathbb{R}^K$ is a vector of K observed factors, and $\mathbf{B}$ is a $p \times K$ matrix of factor loadings. We extend MAE to a method for estimating the realized ξ_t. Its application to stock returns provides a new index that captures information of *whole market*. Section 6 contains conclusions and discussions. All the proofs are relegated to the appendix.

NOTATION: Throughout this paper, for any vector $\mathbf{v}$ and matrix $\mathbf{M}$, we let $\|\mathbf{v}\|$ denote the Euclidean norm of $\mathbf{v}$ and let $\|\mathbf{M}\|$, $\|\mathbf{M}\|_F$ and $\|\mathbf{M}\|_{\max}$ denote its spectral norm, Frobenius norm and entry-wise maximum norm, respectively. We use $\widehat{\theta}^{\mathrm{M}}_{m,j}(\mu_j, \sigma_{jj})$, $\widehat{\theta}^{\mathrm{M}}_m(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\Sigma}))$, $\widehat{\theta}^{\mathrm{I}}_m(\boldsymbol{\mu}, \boldsymbol{\Omega})$ and $\widehat{\theta}^{\mathrm{B}}_m(\boldsymbol{\mu}, \mathrm{diag}_{\mathcal{A}}(\boldsymbol{\Sigma}))$ to denote the Marginal Estimator, MAE, Ideal Estimator, and BAE (to be introduced), respectively, with given $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; when $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are replaced by $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, it means we plug in estimators of the mean vector and covariance matrix. We frequently use notations $(\theta_m, c_m, \eta_m, r_m)$, where θ_m is defined in (3), c_m is defined in (7), η_m and r_m are defined in Definition 2.1. For all settings in this paper, η_m is a constant, (θ_m, c_m, r_m) depend on p but are at the constant scale.

2 Theoretical properties of MAE

We study the asymptotic properties of MAE defined in (10), assuming both (n, p) tend to infinity. First, we study the consistency of MAE. The following theorem shows that, when the distribution is marginally sub-Gaussian, if we plug in the sample mean and sample covariance matrix as $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, then MAE is always root-n consistent.

Theorem 2.1 (Root-n consistency)**.** *Under model* (1)*, suppose* $\log^2(p) = o(n)$ *and* $\max_{1 \le j \le p} \|Y_j - \mu_j\|_{\psi_2} \le C$*, where* $\|\cdot\|_{\psi_2}$ *denotes the sub-Gaussian norm.*[2] *Given iid samples* $\{\boldsymbol{Y}_i\}_{i=1}^n$*, consider the MAE in* (10)*, where* $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ *are the sample mean vector and sample covariance matrix. Then,*

$$|\widehat{\theta}^{\mathrm{M}}_m(\widehat{\boldsymbol{\mu}}, \mathrm{diag}(\widehat{\boldsymbol{\Sigma}})) - \theta_m| = O_{\mathbb{P}}(n^{-1/2}).$$

The root-n consistency of MAE requires *no* conditions on either $\boldsymbol{\Sigma}$ or $\boldsymbol{\Omega}$. It confirms our previous insight that estimating moment parameters is an "easier" statistical problem than estimating large matrices. On the other hand, the plug-in estimators only perform well when the assumed structural assumptions (e.g., sparsity) on $\boldsymbol{\Sigma}$ or $\boldsymbol{\Omega}$ are satisfied.

[2]For a random variable X, its sub-Gaussian norm is defined as $\|X\|_{\psi_2} = \sup_{k \ge 1} k^{-1} (\mathbb{E}|X|^k)^{1/k}$.

Many distributions in the elliptical family are heavy-tailed and don't satisfy the marginal sub-Gaussianity assumption. In these cases, we prefer to use robust estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Fan et al., 2017; Sun et al., 2018+). They are M-estimators with robust loss functions or rank-based estimators. Compared to the sample mean and sample covariance estimators, these robust estimators lead to sharper bounds of $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty$ and $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max}$ in the case of heavy-tailed data. The next theorem studies MAE with general mean/covariance estimators.

Theorem 2.2 (Consistency, with general mean/covariance estimators)**.** *Under model* (1)*, suppose* $\log^2(p) = o(n)$ *and* $\theta_{2m} \le C$*. Given iid samples* $\{\boldsymbol{Y}_i\}_{i=1}^n$*, consider the MAE in* (10)*. We assume the estimators* $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ *satisfy* $\max_{1\le j\le p} |\widehat{\mu}_j - \mu_j| \le \alpha_n$ *and* $\max_{1\le j\le p} |\widehat{\sigma}_{jj} - \sigma_{jj}| \le \beta_n$ *with probability* $1 - o(1)$*, where* $\alpha_n \to 0$ *and* $\beta_n \to 0$ *as* $n, p \to \infty$*. Then, for any* $\epsilon > 0$*, with probability* $1 - \epsilon$*, there is a constant* $C_\epsilon > 0$ *such that*

$$\big|\widehat{\theta}_m^{\mathrm{M}}(\widehat{\boldsymbol{\mu}}, \mathrm{diag}(\widehat{\boldsymbol{\Sigma}})) - \theta_m\big| \le C_\epsilon\big(n^{-1/2} + \max\{\alpha_n, \beta_n\}\big).$$

The typical error rate of robust estimators is $\alpha_n \asymp \sqrt{\log(p)/n}$ and $\beta_n \asymp \sqrt{\log(p)/n}$ (Fan et al., 2017; Sun et al., 2018+), so the associated MAE satisfies $|\widehat{\theta}_m^{\mathrm{M}} - \theta_m| = O_{\mathbb{P}}(\sqrt{\log(p)/n})$. Compared with the rate in Theorem 2.1, the extra $\sqrt{\log(p)}$ factor here is a price paid for heavy tails.

Next, we study the asymptotic variance of MAE. By Theorem 2.1, MAE is already rate-optimal. We would like to see whether it also achieves the optimal "constant". We shall compare its asymptotic variance with that of the Ideal Estimator (4). Since the Ideal Estimator knows the true $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for a fair comparison, we consider MAE with true $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Definition 2.1. *For any* $k \ge 1$*, let* $\eta_k = \mathbb{E}[N(0,1)^{2k}]$ *and* $r_k = (\mathbb{E}\xi^{2k})/(\mathbb{E}\chi_p^{2k})$*, where* χ_p^2 *denotes the chi-square distribution with* p *degrees of freedom.*

The quantities r_k capture the difference between moments of an elliptical distribution and moments of a multivariate Gaussian distribution with matching mean and covariance matrix. It depends on p but is at the constant scale under our settings.

Theorem 2.3 (Variance)**.** *Under model* (1)*, suppose* $\log^2(p) = o(n)$ *and* $\theta_{2m} \le C$*. Given iid samples* $\{\boldsymbol{Y}_i\}_{i=1}^n$*, consider the MAE in* (9) *where* $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *are given. Let* $\boldsymbol{\Lambda} = [\mathrm{diag}(\boldsymbol{\Sigma})]^{-1/2}\boldsymbol{\Sigma}[\mathrm{diag}(\boldsymbol{\Sigma})]^{-1/2}$ *be the correlation matrix. There is a constant* $C_m > 0$*, independent of the distribution of* ξ*, such that*

$$\frac{\mathrm{var}\big(\widehat{\theta}_m^{\mathrm{M}}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\Sigma}))\big)}{\theta_m^2} \le \frac{1}{n}\frac{r_{2m} - r_m^2}{r_m^2} + \frac{1}{np}\frac{r_{2m}}{r_m^2}\frac{\eta_{2m} - \eta_m^2}{\eta_m^2} + \frac{C_m}{n}\frac{r_{2m}}{r_m^2\eta_m^2}\frac{\|\boldsymbol{\Lambda} - \mathbf{I}\|_F^2}{p^2}.$$

When $m = 2$*, the equality holds with* $C_m = 72$*.*

The upper bound for the variance has three terms: The first term is $O(n^{-1})$; as we shall see, this term matches with the variance of the benchmark estimator. The second term is $O(n^{-1}p^{-1})$ and is negligible for diverging p. The third term is caused by correlations among different marginal estimators $\widehat{\theta}^{\mathrm{M}}_{m,j}$. This term is negligible as long as $\|\mathbf{\Lambda}-\mathbf{I}\|_F^2 = o(p^2)$; consider a special case where $\|\mathbf{\Sigma}\|$ is bounded, then $\|\mathbf{\Lambda}-\mathbf{I}\|_F^2 = O(p)$; so the requirement of $\|\mathbf{\Lambda}-\mathbf{I}\|_F^2 = o(p^2)$ is mild. Indeed, if requires that the sparsity of correlation coefficients: $\sum_{i\neq j}\lambda_{ij} = o(p^2)$, where $\mathbf{\Lambda} = (\lambda_{ij})$. The next proposition confirms that the asymptotic variance of MAE is the same as the asymptotic variance of the Ideal Estimator:

Proposition 2.1 (Comparison with benchmark)**.** *Let $\{\boldsymbol{Y}_i\}_{i=1}^n$ be iid samples of model* (1)*. Suppose $\theta_{2m} \le C$. For the Ideal Estimator in* (4)*,*

$$\frac{\mathrm{var}\big(\widehat{\theta}^{\mathrm{I}}_m(\boldsymbol{\mu},\mathbf{\Omega})\big)}{\theta_m^2} = \frac{1}{n}\frac{r_{2m}-r_m^2}{r_m^2} + \frac{1}{np}\frac{r_{2m}}{r_m^2}2m^2\big[1+O(p^{-1})\big].$$

As a result, if $\|\mathbf{\Lambda}-\mathbf{I}\|_F^2 = o(p^2)$, where $\mathbf{\Lambda} = [\mathrm{diag}(\mathbf{\Sigma})]^{-1/2}\mathbf{\Sigma}[\mathrm{diag}(\mathbf{\Sigma})]^{-1/2}$ is the correlation matrix, then

$$\frac{\mathrm{var}\big(\widehat{\theta}^{\mathrm{M}}_m(\boldsymbol{\mu},\mathrm{diag}(\mathbf{\Sigma}))\big)}{\mathrm{var}\big(\widehat{\theta}^{\mathrm{I}}_m(\boldsymbol{\mu},\mathbf{\Omega})\big)} \to 1.$$

Last, we construct confidence intervals of θ_m. Since MAE is the average of p strongly dependent marginal estimators, its asymptotic normality is hard to approach. We instead use the marginal estimator in (8) to construct confidence intervals.

Theorem 2.4 (Asymptotic normality)**.** *Under model* (1)*, suppose* $\log^2(p) = o(n)$ *and* $\max_{1\le j\le p}\|Y_j-\mu_j\|_{\psi_2} \le C$*, where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm. Given iid samples $\{\boldsymbol{Y}_i\}_{i=1}^n$, consider the Marginal Estimator in* (8) *for an arbitrary $1\le j\le p$, where $(\widehat{\mu}_j,\widehat{\sigma}_{jj})$ are the sample mean and sample variance of $\{Y_{ij}\}_{i=1}^n$. Then,*

$$\frac{\sqrt{n}\Big(\widehat{\theta}^{\mathrm{M}}_{m,j}(\widehat{\mu}_j,\widehat{\sigma}_{jj})-\theta_m\Big)}{\sqrt{\frac{c_{2m}}{c_m^2}\widehat{\theta}_{2m}-\widehat{\theta}_m^2}} \to_d N(0,1),$$

where $c_k = (2k-1)!!\,(p/2)^k\frac{\Gamma(p/2)}{\Gamma(p/2+k)}$ for $k\ge 1$, and $(\widehat{\theta}_{2m},\widehat{\theta}_m)$ are consistent estimators of (θ_{2m},θ_m).

This theorem shows somewhat surprisingly that all marginal estimator contains the same amount of information about θ_m. Given consistent estimators $(\widehat{\theta}_{2m},\widehat{\theta}_m)$, the asymptotic level-$\alpha$ confidence interval of θ_m is

$$\widehat{\theta}^{\mathrm{M}}_{m,j} \pm \frac{q_{1-\alpha/2}}{\sqrt{n}}\sqrt{\frac{c_{2m}}{c_m^2}\widehat{\theta}_{2m}-\widehat{\theta}_m^2}, \tag{11}$$

where $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal. It doesn't matter which of $1 \le j \le p$ we use, as these marginal estimators have the same asymptotic variance. For the estimators $(\widehat{\theta}_{2m}, \widehat{\theta}_m)$, we suggest using MAE.

If we only need a point estimator but not a confidence interval, we prefer MAE to the Marginal Estimator, as MAE has a smaller variance in many scenarios. For example, when $\|\mathbf{\Lambda} - \mathbf{I}\|_F^2 = o(p^2)$, by plugging in the true $(\boldsymbol{\mu}, \mathbf{\Sigma})$,

$$\frac{\mathrm{var}(\widehat{\theta}_m^{\mathrm{M}})}{\theta_m^2} \sim \frac{\mathrm{var}(\widehat{\theta}_m^{\mathrm{I}})}{\theta_m^2} \sim \frac{1}{n}\frac{r_{2m} - r_m^2}{r_m^2}, \qquad \frac{\mathrm{var}(\widehat{\theta}_{m,j}^{\mathrm{M}})}{\theta_m^2} \sim \frac{1}{n}\frac{(\eta_{2m}/\eta_m^2)r_{2m} - r_m^2}{r_m^2}.$$ [3]

Since $\eta_{2m} > \eta_m^2$, the latter variance is strictly larger. In contrast, MAE is first-order efficient.

3 Extension to blockwise aggregation

In the construction of MAE, each marginal estimator only uses one coordinate of the samples. It is convenient to implement and gives rise to an estimator that is first-order efficient, provided that the third term in Theorem 2.3 is negligible. It turns out that, the second order term in the variance can be improved upon by using blockwise aggregation, and so is the third term, which is related to the correlation structure. Our simulation studies below show that the improvement is real. This motivates us to extend the marginal estimator to a blockwise estimator that uses a small number of coordinates of the samples and takes into account their correlation structures. We then generalize MAE to BAE — an aggregation of many blockwise estimators.

BAE can be applied to settings where the covariance matrix is approximately blockwise diagonal after row/column permutation. Figure 2 gives such an example, where the S&P 500 stocks divide into many small-size blocks according to sectors or industries of stocks and the stock returns within each block are correlated but admits block structure after taking out the market factor. BAE can take advantage of the within-block correlations and further improve MAE in the second order term.

3.1 A block-wise aggregation estimator (BAE)

We fix a block $J \subset \{1, 2, \ldots, p\}$ and let $K = |J|$. For any vector $\mathbf{v} \in \mathbb{R}^p$ and matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$, let $\mathbf{v}_J$ be the subvector of $\mathbf{v}$ containing the coordinates indexed by J and let $\mathbf{M}_{JJ}$ be the submatrix of $\mathbf{M}$ containing the entries indexed by $J \times J$. By Fang and Zhang (1990), when $\boldsymbol{Y}$ follows an elliptical distribution (1), the subvector $\boldsymbol{Y}_J$ satisfies that

$$\boldsymbol{Y}_J \overset{(d)}{=} \boldsymbol{\mu}_J + B^{1/2}\xi \cdot \mathbf{\Sigma}_{JJ}^{1/2}\boldsymbol{U}_K, \tag{12}$$

[3] This expression combines the asymptotic variance in Theorem 2.4 and the fact that $c_m\theta_m = r_m\eta_m$

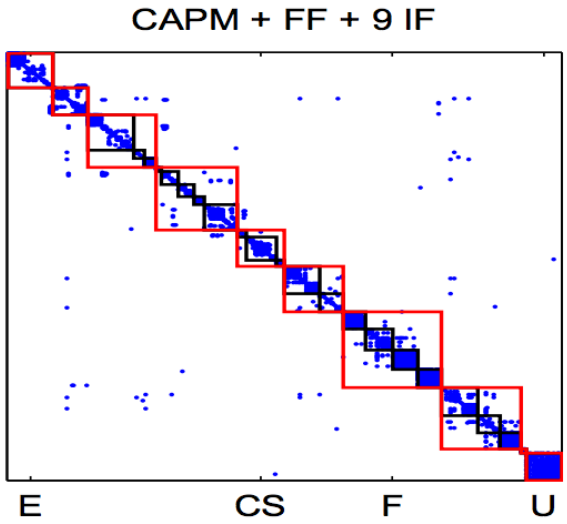

Figure 2: Estimated $\mathbf{\Sigma}$ after factor-removal from S&P500 returns in 2010—2012. Red squares: Sector blocks. Black squares: Industry groups. (From Fan et al. (2015a))

where B is a random variable that follows a beta distribution Beta $(\frac{K}{2}, \frac{p-K}{2})$, the random vector $\boldsymbol{U}_K$ follows a uniform distribution on the unit sphere $\mathbb{S}^{K-1}$, and $(\xi, B, \boldsymbol{U}_K)$ are mutually independent. Since $\|\boldsymbol{U}_K\| = 1$,

$$\xi^{2m} = \frac{\{(\boldsymbol{Y}_J - \boldsymbol{\mu}_J)^{\mathrm{T}} \mathbf{\Sigma}_{JJ}^{-1} (\boldsymbol{Y}_J - \boldsymbol{\mu}_J)\}^m}{B^m}.$$

The random variable B is not directly observable, but its expectation is known:

Proposition 3.1. *For each $m \geq 1$ and $1 \leq K \leq p$, define $c^*_{m,K} = p^m \mathbb{E}(B^m)$ with $B \sim$ Beta $(\frac{K}{2}, \frac{p-K}{2})$. Then,*

$$c^*_{1,K} = K, \qquad c^*_{m,K} = p \times \frac{K + 2m - 2}{p + 2m - 2} \times c^*_{m-1,K} \qquad \text{for } m \geq 2.$$

Replacing B^m by its expectation, we immediately have an estimator of θ_m based on $\{\boldsymbol{Y}_{i,J}\}_{i=1}^n$:

$$\begin{aligned} \widehat{\theta}^{\mathrm{B}}_{m,J}(\boldsymbol{\mu}_J, \mathbf{\Sigma}_{JJ}) &= \frac{1}{np^m} \sum_{i=1}^n \frac{\{(\boldsymbol{Y}_{i,J} - \boldsymbol{\mu}_J)^{\mathrm{T}} \mathbf{\Sigma}_{JJ}^{-1} (\boldsymbol{Y}_{i,J} - \boldsymbol{\mu}_J)\}^m}{\mathbb{E} B^m} \\ &= \frac{1}{nc^*_{m,K}} \sum_{i=1}^n \{(\boldsymbol{Y}_{i,J} - \boldsymbol{\mu}_J)^{\mathrm{T}} \mathbf{\Sigma}_{JJ}^{-1} (\boldsymbol{Y}_{i,J} - \boldsymbol{\mu}_J)\}^m. \end{aligned} \tag{13}$$

We call $\widehat{\theta}^{\mathrm{B}}_{m,J}(\boldsymbol{\mu}_J, \mathbf{\Sigma}_{JJ})$ the *Blockwise Estimator*. Now, given a collection of blocks $\mathcal{A} = \{J_1, J_2, \ldots, J_N\}$, we can define a blockwise estimator for each $J \in \mathcal{A}$ and then take their average:

$$\widehat{\theta}^{\mathrm{B}}_m\big(\boldsymbol{\mu},\ \mathrm{diag}_{\mathcal{A}}(\mathbf{\Sigma})\big) = \frac{1}{|\mathcal{A}|} \sum_{J \in \mathcal{A}} \widehat{\theta}^{\mathrm{B}}_{m,J}(\boldsymbol{\mu}_J, \mathbf{\Sigma}_{JJ}). \tag{14}$$

We call $\widehat{\theta}_m^{\mathrm{B}}\big(\boldsymbol{\mu},\ \mathrm{diag}_{\mathcal{A}}(\boldsymbol{\Sigma})\big)$ the *Blockwise Aggregation Estimator (BAE)*. Here $\mathrm{diag}_{\mathcal{A}}(\boldsymbol{\Sigma})$ denotes the collection of diagonal blocks $\boldsymbol{\Sigma}_{JJ}$ with $J \in \mathcal{A}$. Our final estimator is a plug-in version of BAE by plugging in an estimator $\widehat{\boldsymbol{\mu}}$ and estimators of those diagonal blocks of $\boldsymbol{\Sigma}$.

Since BAE only estimates the small-size diagonal blocks of $\boldsymbol{\Sigma}$ and does not need to estimate $\boldsymbol{\Omega}$, it inherits a nice property of MAE: root-n consistency is guaranteed with no conditions on $\boldsymbol{\Sigma}$ or $\boldsymbol{\Omega}$.

Theorem 3.1 (Root-n consistency)**.** *Fix $m \geq 2$ and $K \geq 1$. Under model (1), suppose $\log^{2m}(p) = o(n)$ and $\max_{1\leq j\leq p} \|Y_j - \mu_j\|_{\psi_2} \leq C$. We assume the minimum eigenvalue of any $K \times K$ diagonal block of $\boldsymbol{\Sigma}$ is lower bounded by C. Let $\mathcal{A}$ be a collection of nonrandom, non-overlapping blocks such that the size of each block is bounded by K. Given iid samples $\{\boldsymbol{Y}_i\}_{i=1}^n$, consider the BAE in (14), where $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are estimated by the sample mean vector and sample covariance matrix. Then,*

$$|\widehat{\theta}_m^{\mathrm{B}}(\widehat{\boldsymbol{\mu}}, \mathrm{diag}_{\mathcal{A}}(\widehat{\boldsymbol{\Sigma}})) - \theta_m| = O_{\mathbb{P}}(n^{-1/2}).$$

Theorem 3.2 (Consistency, with general mean/covariance estimators)**.** *Fix $m \geq 2$ and $K \geq 1$. Under model (1), we assume $\log^{2m}(p) = o(n)$, $\theta_{2m} \leq C$, and the minimum eigenvalue of any $K \times K$ diagonal block of $\boldsymbol{\Sigma}$ is lower bounded by C. Let $\mathcal{A}$ be a collection of nonrandom, non-overlapping blocks where the size of blocks is bounded by K. Given iid samples $\{\boldsymbol{Y}_i\}_{i=1}^n$, consider the BAE in (14), where $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ satisfy $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty \leq \alpha_n$ and $\max_{J\in\mathcal{A}} \|\widehat{\boldsymbol{\Sigma}}_{JJ} - \boldsymbol{\Sigma}_{JJ}\| \leq \beta_n$ with probability $1 - o(1)$, with $\alpha_n \to 0$ and $\beta_n \to 0$ as $n, p \to \infty$. Then, for any $\epsilon > 0$, with probability $1 - \epsilon$, there is a constant $C_\epsilon > 0$ such that*

$$\big|\widehat{\theta}_m^{\mathrm{B}}(\widehat{\boldsymbol{\mu}}, \mathrm{diag}_{\mathcal{A}}(\widehat{\boldsymbol{\Sigma}})) - \theta_m\big| \leq C_\epsilon\big(n^{-1/2} + \max\{\alpha_n, \beta_n\}\big).$$

We note that MAE is a special case of BAE, with all block size equal to 1. The motivation of generalizing MAE to BAE is to better take advantage of correlation structures, and this is revealed by comparing the asymptotic variances of two methods; see Section 3.2 below. To implement BAE, we need to determine the collection of blocks, and in Section 3.3 we discuss how to select blocks.

3.2 Variance comparison

We compute the asymptotic variance of BAE and compare it with the asymptotic variances of MAE and Ideal Estimator. Same as before, in the variance calculation we assume $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are given.

Definition 3.1. *For each $k \geq 1$, let $h_m(k) = \frac{k\cdot\mathrm{var}(\chi_k^{2m})}{(\mathbb{E}\chi_k^{2m})^2}$, where χ_k^2 denotes the chi-square distribution with k degrees of freedom. Given a collection of blocks $\mathcal{A}$, let $\bar{h}_m(\mathcal{A}) = \frac{p}{|\mathcal{A}|^2}\sum_{J\in\mathcal{A}} \frac{h_m(|J|)}{|J|}$.*

Theorem 3.3 (Variance of BAE). *Let $\{\boldsymbol{Y}_i\}_{i=1}^n$ be iid samples of model* (1). *Fix $m \geq 2$ and suppose $\theta_{2m} \leq C$. There exists a constant $\widetilde{C}_m > 0$, independent of the distribution of ξ, such that for any collection $\mathcal{A}$ of non-overlapping blocks,*

$$\frac{\operatorname{var}\big(\widehat{\theta}_m^{\mathrm{B}}\big(\boldsymbol{\mu}, \operatorname{diag}_{\mathcal{A}}(\boldsymbol{\Sigma})\big)\big)}{\theta_m^2} \leq \frac{1}{n}\frac{r_{2m} - r_m^2}{r_m^2} + \frac{1}{np}\frac{r_{2m}}{r_m^2}\bar{h}_m(\mathcal{A}) + \frac{C_m}{n}\frac{1}{|\mathcal{A}|^2}\sum_{\substack{I,J\in\mathcal{A}\\ I\neq J}} \|\boldsymbol{\Sigma}_{II}^{-1/2}\boldsymbol{\Sigma}_{IJ}\boldsymbol{\Sigma}_{JJ}^{-1/2}\|^2.$$

The upper bound of the variance has three terms:

- The first term is $O(n^{-1})$, which also appears in the variance of MAE and Ideal Estimator. It is the dominating term of the variance.

- The second term is $O(p^{-1}n^{-1})$, where the constant in front of it is related to a quantity $\bar{h}_m(\mathcal{A})$. We call $\bar{h}_m(\mathcal{A})$ the *block-division factor*, as it is only a function of $\mathcal{A}$. To see how this factor changes with block size, let's consider a special case where all blocks have an equal size k and p is a multiple of k. Then,

  $$\bar{h}_m(\mathcal{A}) = h_m(k) = \frac{k \cdot \operatorname{var}(\chi_k^{2m})}{(\mathbb{E}\chi_k^{2m})^2}.$$

  It is a monotone decreasing function of k (see Figure 3). Hence, increasing the block size leads to a reduction of this term, which indicates that the second order efficiency of MAE can be improved with $m > 1$.

- The last term comes from the correlations among estimators associated with different blocks. It doesn't exist for the Ideal Estimator, but both MAE and BAE have this extra term. For MAE, all off-diagonal entries of $\boldsymbol{\Sigma}$ contribute to this term. However, for BAE, only off-diagonal blocks contribute. Especially, when $\boldsymbol{\Sigma}$ is blockwise diagonal with respect to $\mathcal{A}$, this extra term becomes zero. Again, increasing the block size leads to a reduction of this term.

From MAE to BAE, we can see that the dominating term in the variance bound remains the same, but the other two terms are reduced and the performance still improves. However, we cannot use too large blocks, because BAE needs to invert an estimate of $\boldsymbol{\Sigma}_{J,J}$ and the error of estimating $\widehat{\boldsymbol{\Sigma}}_{J,J}$ increases as the block size increases.

We now give a more thorough comparison of four estimators, the Ideal Estimator (IE) $\widehat{\theta}_m^{\mathrm{I}}$, the Marginal Estimator (ME) $\widehat{\theta}_{m,j}^{\mathrm{M}}$, the MAE $\widehat{\theta}_m^{\mathrm{M}}$, and the BAE $\widehat{\theta}_m^{\mathrm{B}}$; see Table 1. We conclude that

- IE has the optimal variance, but it works unsatisfactorialy in the real case of unknown $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, as it requires estimating $\boldsymbol{\Omega}$.
- ME avoids estimating $\boldsymbol{\Omega}$ and works in the real case, but its asymptotic variance is non-optimal.

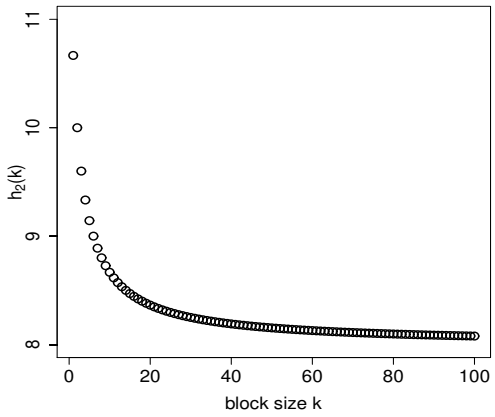

Figure 3: Plot of $\bar{h}_m(\mathcal{A})$ when all blocks have an equal size k ($m = 2$; the x-axis represents k). As the block size increases, $\bar{h}_m(\mathcal{A})$ decreases, suggesting a variance reduction.

- MAE aggregates a number of ME's and achieves the optimal variance when $\|\mathbf{\Lambda} - \mathbf{I}\|_F^2 = o(p^2)$.
- Compared with MAE, BAE relaxes the condition of $\|\mathbf{\Lambda} - \mathbf{I}\|_F^2 = o(p^2)$ and reduces the second-order term of the variance.

From ME to BAE, we have used two methodological ideas: to aggregate "local" estimators and to use a block of coordinates in each "local" estimator. Both help reduce the variance of the estimator, with the first idea playing a more significant role.

Table 1: Variance comparison of estimators (known $(\boldsymbol{\mu}, \mathbf{\Sigma})$; ** means the constant is optimal).

	IE	ME	MAE	BAE
dominating term	$\frac{1}{n}\big(\frac{r_{2m}}{r_m^2} - 1\big)^{**}$	$\frac{1}{n}\big(\frac{r_{2m}\eta_{2m}}{r_m^2\eta_m^2} - 1\big)$	$\frac{1}{n}\big(\frac{r_{2m}}{r_m^2} - 1\big)^{**}$	$\frac{1}{n}\big(\frac{r_{2m}}{r_m^2} - 1\big)^{**}$
2nd-order term	$\frac{1}{np}\frac{r_{2m}}{r_m^2}h_m(p)^{**}$	—	$\frac{1}{np}\frac{r_{2m}}{r_m^2}h_m(1)$	$\frac{1}{np}\frac{r_{2m}}{r_m^2}h_m(k)$
correlation term	0^{**}	0^{**}	$\frac{C}{np}\sum_{1\le i\ne j\le p}\lvert\Lambda_{jj}\rvert^2$	$\frac{C}{np}\sum_{I\ne J\in\mathcal{A}}\lVert\mathbf{\Lambda}_{I,J}\rVert_F^2$

Remark 1. IE and MAE are special cases of BAE with equal-size blocks of $k = 1$ and $k = p$, respectively. We note that $h_m(1) = \frac{\eta_{2m} - \eta_m^2}{\eta_m^2}$ and $h_m(p) = 2m^2[1 + O(p^{-1})]$, so Theorem 3.3 matches with the variance bounds of MAE (Theorem 2.3) and the IE (Proposition 2.1).

Remark 2 (multivariate Gaussian). Let's consider a special case where the data are multivariate Gaussian but the user doesn't know and still applies the estimators in this paper. For Gaussian distributions, the first term in the variance bound disappears, so the estimators considered here all have a faster rate of convergence as $O(p^{-1}n^{-1})$. This is the only case where a large p helps, i.e., "dimensionality is a blessing." Moreover, the difference between MAE and BAE is more prominent,

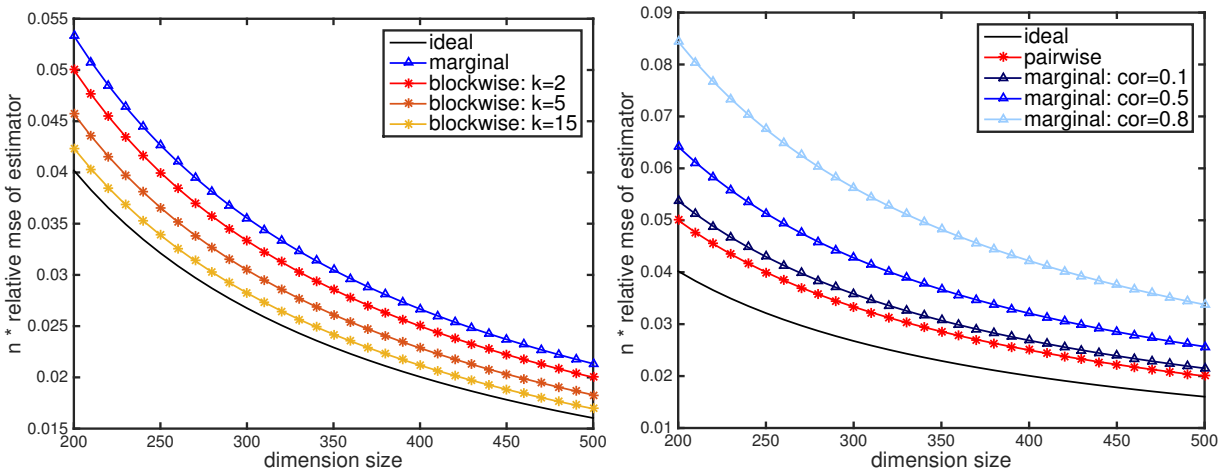

Figure 4: Comparison of IE, MAE and BAE for multivariate Gaussian distributions (y-axis is $\mathrm{var}(\widehat{\theta}_2^2/\theta_2^2)$. Left: $\mathbf{\Sigma} = \mathbf{I}_p$. Right: $\mathbf{\Sigma}$ is a blockwise diagonal matrix with 2×2 blocks whose diagonals are 1 and off-diagonals are ρ, where ρ takes values in $\{0.1, 0.5, 0.8\}$. The pairwise estimator refers to BAE with $k = 2$. Curves are from theoretical calculations (see Corollary C.1 in the appendix). The variance of IE and BAE is independent of ρ, so there is only one curve for all values of ρ.

as the second term in the variance bound is now dominating. Figure 4 displays the error bound according to Theorem 3.3 for the case of $\mathbf{\Sigma} = \mathbf{I}$ and $\mathbf{\Sigma}$ being a blockwise diagonal matrix with 2×2 blocks whose off-diagonal element is ρ. The results favor BAE, especially for the blockwise $\mathbf{\Sigma}$ with large within-block off-diagonals.

3.3 Construction of blocks

We provide two approaches of selecting the blocks. The first approach works well when the true $\mathbf{\Sigma}$ is approximately block-wise diagonal, such as example on the returns of the S&P 500 components(see Figure 2). The second approach is a random scheme and works for general settings.

BAE1: Constructing blocks from a raw estimate of $\mathbf{\Sigma}$. Let $\widetilde{\mathbf{\Sigma}}$ be a raw estimate of $\mathbf{\Sigma}$; for example, it can be the sample covariance matrix or the robust estimator of $\mathbf{\Sigma}$ in Section 4. Fixing a threshold $t \in (0, 1)$, we define a graph $\mathcal{G}_t$ with nodes $\{1, 2 \cdots, p\}$, where there is an undirected edge between nodes i and j if and only if the estimated absolute correlation exceeds t, namely,

$$|\widetilde{\Sigma}(i,j)|/\sqrt{\widetilde{\Sigma}(i,i)\widetilde{\Sigma}(j,j)} > t, \qquad \text{for } 1 \le i < j \le p.$$

The nodes of this graph uniquely partitions into components (a component of a graph is a maximal connected subgraph). We propose using

$$\mathcal{A} = \{\text{all components of } \mathcal{G}_t\}.$$

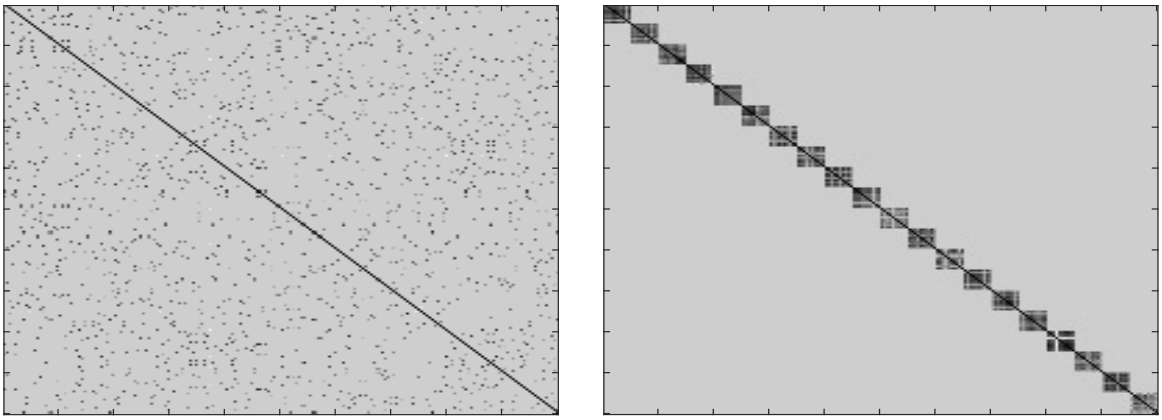

Figure 5: Construction of a blockwise correlation matrix by thresholding. Left panel: Graph of the original correlation matrix. Right panel: Transformation into a block diagonal correlation matrix.

See Figure 5 for an illustration of this procedure.

This approach guarantees that all blocks are non-overlapping. Numerical evidence suggests that it performs well with an appropriate choice of t, especially when the true $\mathbf{\Sigma}$ is blockwise diagonal. However, the threshold t is a tuning parameter, and it can be inconvenient to select t in a data-driven fashion. Below, we introduce a tuning-free approach.

BAE2: Randomly selecting pairs as blocks. In this approach, we let

$$\mathcal{A} = \big\{p \text{ pairs uniformly drawn from } \{(i,j) : 1 \le i < j \le p\} \text{ without replacement}\big\}.$$

This approach is designed for block size equal to 2, and the obtained blocks may overlap. Although it sounds ad-hoc, this approach has an appealing numerical performance. When the number of pairs are sampled sufficiently large, by the law of large numbers, it approaches the all pairwise aggregation estimator and this explains why the approach has an appealing numerical performance. This approach can easily be extended to blocks of any size that is smaller than n so long as the estimated covariance matrix for each block can be easily inverted and estimated well.

4 Simulations

We investigate the performance of estimators on extensive simulations. To have realistic simulation settings, we use a $\mathbf{\Sigma}$ calibrated from stock returns. The calibration procedure is the same as that in Fan et al. (2015c) and Fan et al. (2013). Fix p. We take the daily returns of p companies in S&P 500 index with the largest market capitalization from July 1st, 2013 to June 29th, 2018 (data were downloaded from the COMPUSTAT website). We fit the Fama-French three-factor model to the

excess returns $\{\boldsymbol{Y}_t\}_{t=1}^T$:

$$\mathbf{Y}_t = \mathbf{a} + \mathbf{B}\boldsymbol{f}_t + \mathbf{u}_t,$$

where $\mathbf{B} \in \mathbb{R}^{p\times 3}$ is the factor loading matrix, $\boldsymbol{f}_t \in \mathbb{R}^3$ denotes the Fama-French factors with covariance matrix $\mathrm{cov}(\boldsymbol{f}_t) \in \mathbb{R}^{3\times 3}$ and $\mathbf{u}_t$ is the idiosyncratic component. This factor model induces a covariance structure for $\mathbf{Y}_t$:

$$\boldsymbol{\Sigma}_Y = \mathrm{cov}(\boldsymbol{Y}) = \mathbf{B}\,\mathrm{cov}(\boldsymbol{f}_t)\mathbf{B}^{\mathrm{T}} + \boldsymbol{\Sigma}_u,$$

where $\boldsymbol{\Sigma}_u$ is the covariance matrix of idiosyncratic noise $\mathbf{u}_t$. We downloaded the factors $\{\boldsymbol{f}_t\}_{t=1}^T$ from the Kenneth French data library and used the method in Fan et al. (2013) with the recommended threshold (for estimating sparse $\boldsymbol{\Sigma}_u$) to get and estimate $\widehat{\boldsymbol{\Sigma}}_Y$. We then use $\widehat{\boldsymbol{\Sigma}}_Y$ as the true $\boldsymbol{\Sigma}$ to generate data from model (1).

When implementing the estimators, we plug in two different estimators of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The first choice is to use sample mean and sample covariance matrix. The second choice is to use robust M-estimators, called adaptive Huber estimator (Fan et al., 2017; Sun et al., 2018+), which are designed for heavy-tailed data. These estimators lead to better large-deviation bounds. In detail, for a tuning parameter $\tau > 0$ chosen by cross-validation, we estimate $\boldsymbol{\mu}$ by $(\widehat{\mu}_1, \ldots, \widehat{\mu}_p)^{\mathrm{T}}$, where

$$\widehat{\mu}_j = \operatorname*{argmin}_{\beta\in\mathbb{R}} \sum_{i=1}^n \ell_\tau(Y_{ij} - \beta), \qquad \text{with} \quad \ell_\tau(u) = \begin{cases} \frac{1}{2}u^2, & \text{if } |u| \le \tau, \\ \tau|u| - \frac{1}{2}\tau^2, & \text{if } |u| > \tau, \end{cases}$$

the Huber loss. We estimate $\boldsymbol{\Sigma}$ by $(\widehat{\sigma}_{jk})_{1\le j,k\le p}$, where

$$\widehat{\sigma}_{jj} = \widehat{\beta}_j - \widehat{\mu}_j^2 \cdot 1\{\widehat{\beta}_j > \widehat{\mu}_j^2\}, \qquad \text{with} \quad \hat{\beta}_j = \operatorname*{argmin}_{\beta>0} \sum_{i=1}^n \ell_{\tau_{jj}}\big(Y_{ij}^2 - \beta\big),$$

$$\widehat{\sigma}_{jk} = \widehat{\beta}_{jk} - \widehat{\mu}_j\widehat{\mu}_k, \qquad \text{with} \quad \widehat{\beta}_{jk} = \operatorname*{argmin}_{\beta\in\mathbb{R}} \sum_{i=1}^n \ell_{\tau_{jk}}\big(Y_{ij}Y_{ik} - \beta\big).$$

Here, each tuning parameter τ_{jk} is selected via cross-validation using the data $\{(Y_{ij}, Y_{ik})\}_{i=1}^n$.

Experiment 1: Performance of MAE. Fix $m = 2$. We consider four sub-experiments:

- *Experiments 1.1 and 1.3*: We fix $p = 500$ and let n vary in $\{50, 100, 150, 200, 250, 300\}$. The data follow multivariate Gaussian distributions (Experiment 1.1) or multivariate t-distributions with degrees of freedom equal to 4.5 (Experiment 1.3).
- *Experiments 1.2 and 1.4*: We fix $n = 100$ and let p vary in $\{250, 400, 550, 700, 850, 1000\}$. The data follow multivariate Gaussian distributions (Experiment 1.2) or multivariate t-distributions with degrees of freedom equal to 4.5 (Experiment 1.4).

In all settings, $p > n$, so we focus on the challenging case of high-dimensionality. For each setting, we compare four estimators:

- $\widehat{\theta}^{\mathrm{I}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: Ideal Estimator, which knowns $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- $\widehat{\theta}^{\mathrm{M}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: MAE with given $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- $\widehat{\theta}^{\mathrm{M}}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$: MAE, where $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are estimated using the sample mean/covariance matrix in Experiment 1.1&1.2 and using the aforementioned robust-M estimators for Experiment 1.3&1.4.
- $\widehat{\theta}^{\mathrm{I}}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}_P)$: Plug-in Ideal Estimator, with plugged-in estimators of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We use the sample mean to estimate $\widehat{\boldsymbol{\mu}}$ and use POET (Fan et al., 2013) (with a default threshold) to estimate $\boldsymbol{\Sigma}$.

The results are presented in Figure 6, where the y-axis is $\log\{(\widehat{\theta}_2/\theta_2 - 1)^2\}$, based on the average over 200 repetitions. As we have expected, the Ideal Estimator always gives the lowest error, however, such an estimator is not practically feasible. Instead, we plug estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ into the Ideal Estimator to make it practically feasible, then it has an unsatisfactory performance; this confirms our previous insight about the drawback of the plug-in estimator. Our proposed MAE works well, always significantly better than the plug-in estimator. The performance of MAE becomes better as the sample size n grows, and its performance stays relatively stable as the dimension p grows. This is desirable: our proposed estimator doesn't face any curse of dimensionality. The results are similar for the multivariate Gaussian data and the multivariate t-data, except that for Gaussian data, MAE with $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ even outperforms MAE with true $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. One possible reason is the self-normalization phenomenon: An estimator, when divided by its sample variance, gives better performance than that divided by the true variance.

Experiment 2: Confidence Interval. For each of the experiments above: *Experiments 1.1, 1.2, 1.3 and 1.4*, we calculate the probability that the true value of θ_2 lies in the confidence interval derived in Theorem 2.4 and presented in Equation (11). In Table 2, we see that for a 95% confidence interval, the empirical coverage probabilities are close to the confidence level.

Experiment 3: Performance of BAE. We study whether BAE, which uses a block of coordinates at a time and takes advantage of the correlation structure, can further improve the performance of MAE. The four sub-experiments, Experiments 3.1-3.4, have the same settings as those of Experiments 1.1-1.4. When implementing BAE, we use the second approach in Section 3.3 to choose the blocks; note that the blocks all have a size 2 and may overlap. We use the sample mean/covariance to estimate $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for multivariate Gaussian data and the robust M-estimators for multivariate t

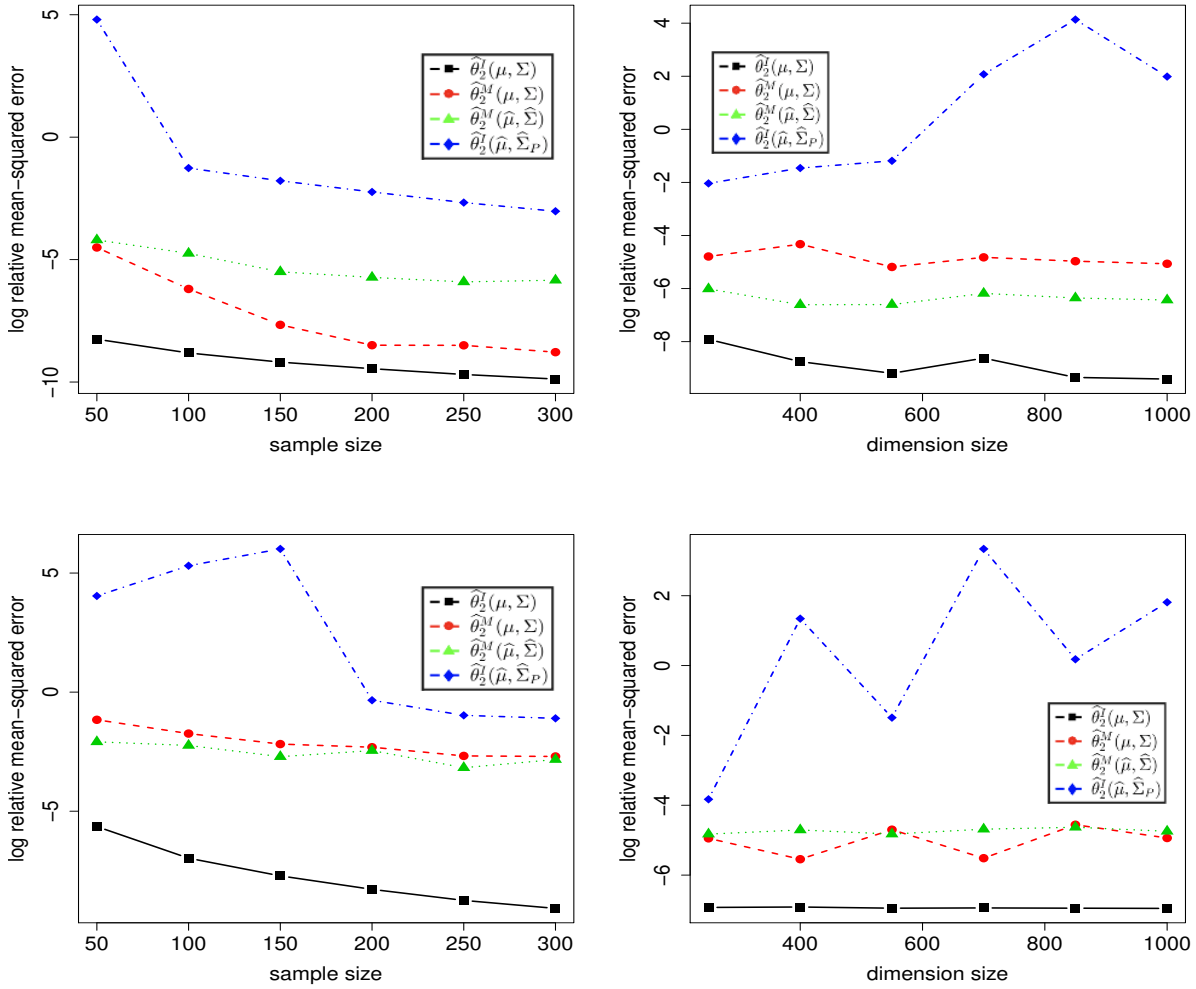

Figure 6: Experiment 1 (MAE). Top two panels: Experiment 1.1&1.2 (multivariate Gaussian data). Bottom panels: Experiment 1.3&1.4 (multivariate t data). Errors are the average of 200 repetitions. Black-squared for the ideal-estimator $\hat{\theta}_2^I(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; blue-diamond for the plug-in estimator $\hat{\theta}_2^I(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$; red-dot for the MAE $\hat{\theta}_2^M(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; green-triangle for the plug-in MAE $\hat{\theta}_2^M(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$

data. Since we focus on the comparison between MAE and BAE, we do not report the errors of the Ideal Estimator and plug-in estimator in this experiment.

The results are presented in Figure 7. First, we can see that BAE improves the performance of MAE, especially when p is large. Second, the self-normalization phenomenon is also observed: BAE with $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ even outperforms BAE with true $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, especially for Gaussian data.

Table 2: Empirical coverage probability that θ_2 lies in the 95% confidence interval by Equation (11) for data following multivariate Gaussian or multivariate t-distributions, across a variety of settings.

$n = 100$	$p = 250$	400	550	700	850	1000
Gaussian	92.0%	95.0%	93.5%	95.5%	95.5%	96.5%
Student's t	96.5%	98.0%	94.5%	97.0%	96.0%	96.5%
$p = 500$	$n = 50$	100	150	200	250	300
Gaussian	95.5%	94.2%	93.5%	93.0%	95.5%	94.0%
Student's t	98.0%	96.0%	95.5%	93.5%	94.5%	97.0%

5 Application: Estimating realized ξ_t in a time series

Given the returns of a panel of stocks, we are interested in extending the idea of MAE to provide a daily *risk index* for the whole panel of stocks. We cast it as the problem of estimating the realized ξ_t in a multivariate time series with elliptically-distributed noise. Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T \in \mathbb{R}^p$ be the returns of p stocks during a time period of T days. We extend model (1) to an elliptical model for multivariate time series

$$\boldsymbol{Y}_t = \boldsymbol{\mu}_t + \mathbf{B}\boldsymbol{f}_t + \xi_t \boldsymbol{\Sigma}_t^{1/2} \boldsymbol{U}_t, \qquad t = 1, \cdots, T, \tag{15}$$

where $\boldsymbol{\mu}_t$ is the time-varying mean, $\boldsymbol{f}_t \in \mathbb{R}^K$ is a vector of K factors, and $\mathbf{B}$ is a $p \times K$ matrix of factor loadings. We are interested in estimating the daily realized ξ_t.

Our method has four steps:

1. Estimate $\boldsymbol{\mu}_t$. For daily or higher frequency data, we set $\widehat{\boldsymbol{\mu}}_t \equiv \mathbf{0}$, since it is commonly believed that the short-time returns are not predictable. For weekly or monthly data, we estimate $\boldsymbol{\mu}_t$ by the weekly or monthly average.

2. Obtain the factor-adjusted returns $\widehat{\boldsymbol{Z}}_t$. Let $\widehat{\boldsymbol{f}}_t \in \mathbb{R}^K$ contain either observed factors or data-drive factors from PCA (Fan et al., 2013). We then follow the approach in Fan et al. (2013) to get $\widehat{\mathbf{B}}$, the estimated factor loading matrix. Let

$$\widehat{\boldsymbol{Z}}_t = \mathbf{Y}_t - \widehat{\boldsymbol{\mu}}_t - \widehat{\mathbf{B}}\widehat{\boldsymbol{f}}_t, \qquad t = 1, \cdots, T.$$

3. Estimate $\boldsymbol{\Sigma}_t$. We assume $\boldsymbol{\Sigma}_t$ is a diagonal matrix and estimate its diagonal elements by fitting an ARCH model on each coordinate of $\boldsymbol{Z}_t$. In detail, for each $1 \le j \le p$, let $Z_t(j)$ be the j-th coordinate of $\boldsymbol{Z}_t$. We assume there is idiosyncratic noise $\{\epsilon_t(j)\}_{t=1}^T$ such that

$$Z_t(j) = \lambda_t(j)\epsilon_t(j), \qquad \text{where} \quad \lambda_t^2(j) = a_0 + a_1 Z_{t-1}^2(j) + \ldots + a_k Z_{t-k}^2(j),$$

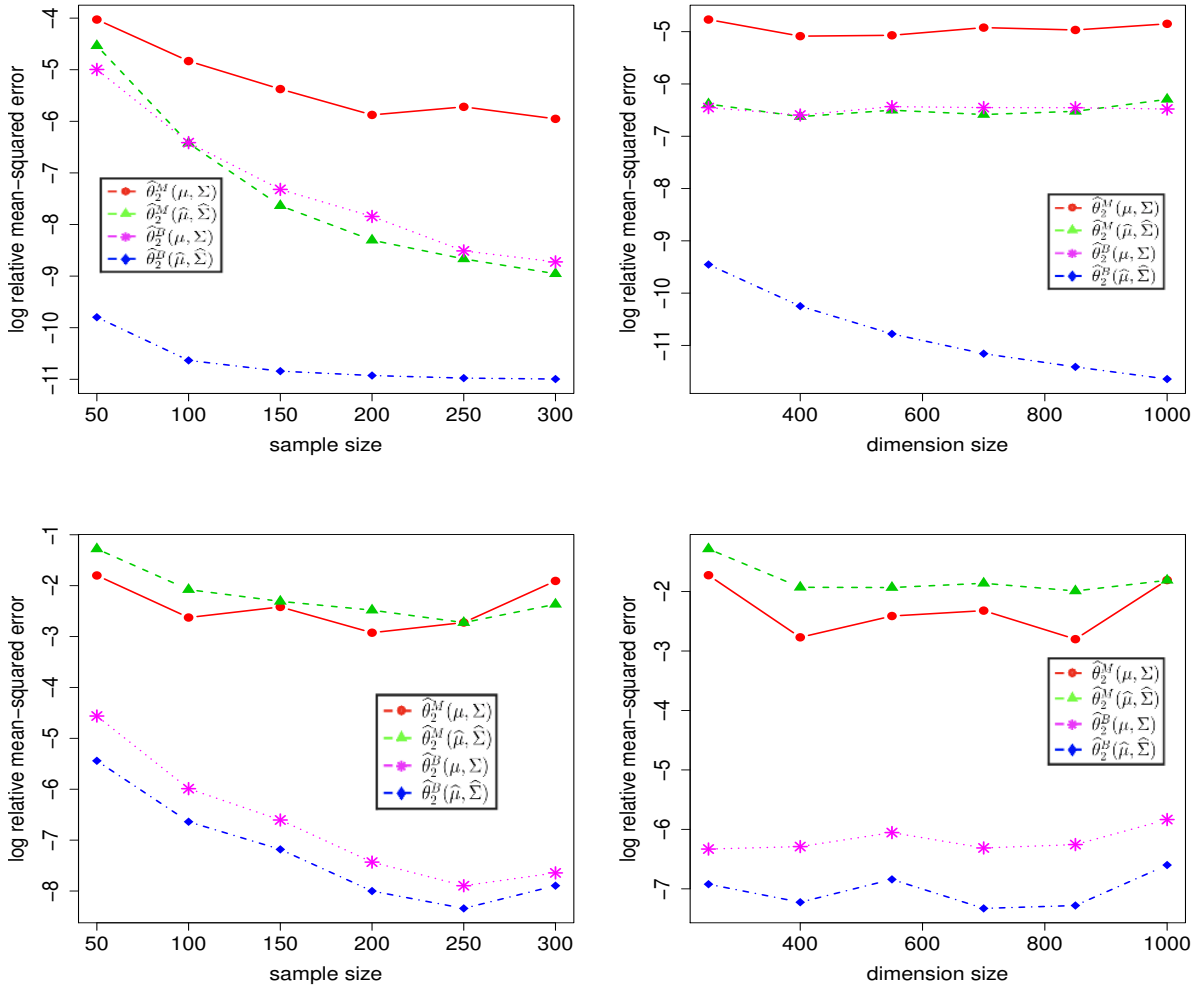

Figure 7: Experiment 3 (BAE). Top two panels: Experiment 3.1&3.2 (multivariate Gaussian data). Bottom panels: Experiment 3.3&3.4 (multivariate t data). Errors are the average of 200 repetitions. Magenta-star for the BAE $\hat{\theta}_2^{\mathrm{B}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; blue-diamond for the plug-in BAE $\hat{\theta}_2^{\mathrm{B}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$; red-dot for the MAE $\hat{\theta}_2^{M}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; green-triangle for the plug-in MAE $\hat{\theta}_2^{M}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$

where k is the order of ARCH model and $(a_0, \ldots, a_k)$ are parameters. We estimate $(a_0, \ldots, a_k)$ using the conditional maximum likelihood estimator and then construct $\{\widehat{\lambda}_t(j)\}_{t=1}^T$. Let

$$\widehat{\boldsymbol{\Sigma}}_t = \mathrm{diag}\big(\widehat{\lambda}_t(1), \ldots, \widehat{\lambda}_t(p)\big).$$

4. Estimate ξ_t. We adapt the idea of MAE to the current setting. Let $\boldsymbol{Z}_t = \boldsymbol{Y}_t - \mathbf{B}\boldsymbol{f}_t$. Our model

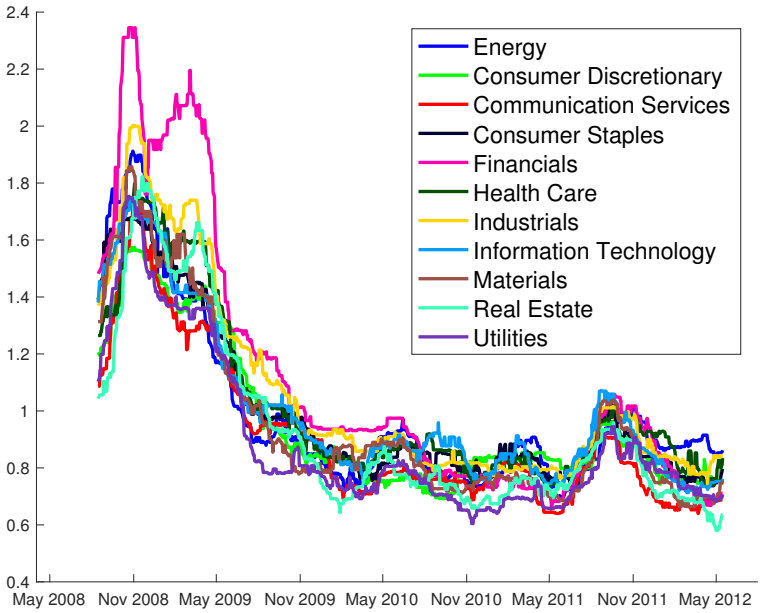

Figure 8: The estimated $\widehat{\xi}_t$ for 11 GCIS sectors. For a better representation, we have smoothed the curves by taking a moving average on a 65-day window.

becomes $\boldsymbol{Z}_t = \xi_t \boldsymbol{\Sigma}_t^{1/2} \boldsymbol{U}_t$, i.e., the j-th component of $\boldsymbol{Z}_t$ is $Z_t(j) = \xi_t (\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}_t)_j$. It follows that

$$\xi_t^2 = \frac{Z_t^2(j)}{(\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}_t)_j^2} \approx \frac{Z_t^2(j)}{\mathbb{E}[(\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}_t)_j^2]} = \frac{pZ_t^2(j)}{\Sigma_t(j,j)},$$

where $\Sigma_t(j,j)$ is the j-th diagonal of $\boldsymbol{\Sigma}_t$. Here, the last equality is due to $c_1 = 1$ in Equation (7). We approximate $(\boldsymbol{Z}_t, \boldsymbol{\Sigma}_t)$ by $(\widehat{\boldsymbol{Z}}_t, \widehat{\boldsymbol{\Sigma}}_t)$ and get a marginal estimator of ξ_t^2: $\widehat{\xi}_{t,j}^2 = p\widehat{Z}_t(j)/\widehat{\Sigma}_t(j,j)$. We then aggregate them:

$$\widehat{\xi}_t^2 = \sum_{j=1}^{p} \frac{\widehat{Z}_t^2(j)}{\widehat{\Sigma}_t(j,j)}, \qquad t = 1, 2, \ldots, T. \tag{16}$$

In Section D of the appendix, we investigate the performance of our estimator in simulations. Under a variety of settings, our estimated curve of $\widehat{\xi}_t$ fits the true curve of ξ_t very well. See details therein.

We applied our estimator to the S&P 500 stock returns. We took the daily returns of 300 stocks from the S&P 500 index with the largest market capitalization, from July 1, 2008 to June 29, 2012. Each stock is assigned a Global Industry Classification Standard (GCIS) code. The GCIS code divides 300 stocks into eleven sectors: Energy, Consumer Discretionary, Communication Services, Consumer Staples, Financials, Health Care, Industrials, Information Technology, Materials, Real Estate, and Utilities. We applied our estimator to stocks in each sector. When implementing our

Table 3: Pairwise correlations of $\widehat{\xi}_t$ across GCIS sectors. Numbers $\geq .45$ are marked in circles.

	E	CD	CO	CS	F	HC	IN	IT	M	R	U
Energy (E)	–	.36	.43	.42	(.53)	.37	.44	.35	(.48)	(.47)	(.46)
Consumer Discretionary (CD)	.36	–	.30	.35	.43	.34	.37	.33	.33	.36	.35
Communication Services (CO)	.43	.30	–	.33	.43	.33	.40	.32	.39	.44	.41
Consumer Staples (CS)	.42	.35	.33	–	.42	.36	.37	.35	.38	.38	.43
Financials (F)	(.53)	.43	.43	.42	–	.42	(.50)	.40	(.49)	(.52)	(.52)
Health Care (HC)	.37	.34	.33	.36	.42	–	.40	.39	.37	.36	.33
Industrials (IN)	.44	.37	.40	.37	(.50)	.40	–	.39	(.49)	.44	(.45)
Information Technology (IT)	.35	.33	.32	.35	.40	.39	.39	–	.36	.33	.34
Materials (M)	(.48)	.33	.39	.38	(.49)	.37	.49	.36	–	.42	.43
Real Estate (R)	(.47)	.36	.44	.38	(.52)	.36	.44	.33	.42	–	(.46)
Utilities (U)	(.46)	.34	.41	.43	(.52)	.33	.45	.34	.43	.46	–

method, we set $\widehat{\boldsymbol{\mu}}_t \equiv \mathbf{0}$ in Step 1, used three observed Fama-French factors as $\widehat{\boldsymbol{f}}_t$ in Step 2, and set the order of ARCH model to $k = 2$ in Step 3.

The curves of estimated $\widehat{\xi}_t$ for 11 sectors are displayed in Figure 8 (the curves are smoothed by taking a moving average on a 65-day window). The estimated $\widehat{\xi}_t$ for all sectors largely synchronize, reaching their peaks during the 2008 financial crisis. In the crisis, the estimated $\widehat{\xi}_t$ for the Financials sector is significantly larger than that of other sectors. The large value of $\widehat{\xi}_t$ for the Financials sector remains in the post-crisis period until May, 2009. We also computed the pairwise correlations among $\widehat{\xi}_t$ of 11 sectors, as shown in Table 3. It suggests that the $\widehat{\xi}_t$ for the Energy sector and the Financials sector are highly correlated with each other. These two sectors are also highly correlated with sectors of Materials, Real Estate, and Utilities. In comparison, for the Consumer Discretionary sector and Information Technology sector, their $\widehat{\xi}_t$ are less correlated with those of other sectors.

6 Discussion

In this paper, we consider the problem of estimating the even moments of ξ in an elliptical distribution $\boldsymbol{Y} = \boldsymbol{\mu} + \xi \boldsymbol{\Sigma}^{1/2} \boldsymbol{U}$. A natural idea is the plug-in estimator (Maruyama and Seo, 2003; Fan et al., 2015b), which requires an estimator $\widehat{\boldsymbol{\Omega}}$ of the precision matrix and whose performance crucially relies on structural assumptions on $\boldsymbol{\Omega}$ or $\boldsymbol{\Sigma}$. Instead, we propose a marginal aggregation estimator (MAE) that only needs to estimate the diagonal of $\boldsymbol{\Sigma}$. Our approach validates the insight that estimating a large precision matrix is statistically more challenging than estimating a moment parameter—it is unnecessary to use the sledge hammer to crack an egg. We prove that MAE is root-n consistent, under *no* conditions on $\boldsymbol{\Sigma}$ or $\boldsymbol{\Omega}$. We also show that MAE achieves the first-order efficiency, with an asymptotic variance matching with the variance of an ideal estimator when $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

are given. We further generalize MAE to a block-wise aggregation estimator (BAE) that needs to estimate small-size diagonal blocks of $\mathbf{\Sigma}$. BAE takes advantage of correlations among coordinates and improves MAE on the second-order efficiency. Our proposed estimators are conceptually simple and easy to implement.

Estimating the moment parameters of an elliptical distribution is useful in quadratic discriminant analysis (Fan et al., 2015b) and estimating tail behavior of financial returns (Fama, 1965; Bollerslev and Wooldridge, 1992; Eberlein and Keller, 1995; Frahm et al., 2003; Cizek et al., 2005). In an application on the stock returns, we propose a multivariate time series model with factor structures and elliptically distributed idiosyncratic noise. We extend MAE to an estimator for estimating the day-to-day value of ξ_t. We apply the method to stocks of each industry sector. It produces an "tail index" for each industry sector. These tail indices reveal interesting difference among industry sectors, especially during the financial crisis.

The study leaves a few open questions for future work. The first is how to improve the estimators for heavy tailed data. Our current approach plugs into MAE the robust estimators of mean and covariance matrix. Instead, we may construct a robust M-estimator for simultaneously estimating $(\theta_m, \mu_j, \sigma_{jj})$ with marginal data and then aggregate these marginal estimators of θ_m in a similar way. We hope such an approach helps remove the $\sqrt{\log(p)}$-factor in the error rate of Theorem 2.2. The second is the optimal strategy of constructing blocks in BAE. There is a trade-off in choosing the blocks: With larger blocks, it reduces the variance of the estimator when true $(\boldsymbol{\mu}, \mathbf{\Sigma})$ are plugged in, but at the same time, the errors of estimating diagonal blocks of $\mathbf{\Sigma}$ increase. How to construct the blocks in a data-driven way is an interesting question. Third, the current theory for BAE assumes non-overlapping blocks. The results can be extended to overlapping blocks, with nontrivial efforts. We leave it for future work. The last problem is to extend our estimators to time dependent data, where the distribution of ξ have change-points. For financial data, such change-points may relate to financial boom or crisis. We propose a kernel-smoothed version of MAE: Given data $\{\boldsymbol{Y}_t\}_{t=1}^n$, for a kernel function $K_h(\cdot)$ with bandwidth h, let

$$\widehat{\theta}_{m,t} = \frac{1}{\sum_{s=1}^n K_h(s-t)} \sum_{s=1}^n K_h(s-t) \Big[\sum_{j=1}^p \frac{p^{-1}(Y_{s,j} - \widehat{\mu}_j)^{2m}}{c_m \widehat{\sigma}_{jj}^{2m}} \Big].$$

We can similarly define the one-sided versions of the kernel estimator. We can combine these estimators with change-point detection methods, which we leave for future work.

References

Bollerslev, T. and Wooldridge, J. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews* **11** 143–172.

Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607.

Cizek, P., Härdle, W. K. and Weron, R. (2005). *Statistical Tools for Finance and Insurance*. Springer-Verlag Berlin Heidelberg.

Eberlein, E. and Keller, U. (1995). Hyperbolic distributions in finance. *Bernoulli* **1** 281–299.

Fama, E. (1965). The behavior of stock market prices. *Journal of Business* **38** 34–105.

Fan, J., Furger, A. and Xiu, D. (2015a). Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high frequency data. *Social Science Research Network, SSRN-id 2548613* .

Fan, J., Ke, Z., Liu, H. and Xia, L. (2015b). QUADRO: A supervised dimension reduction method via Rayleigh Quotient optimization. *The Annals of Statistics* **43** 1498–1534.

Fan, J., Li, Q. and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society,* Series B **79** 247–265.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B* **75** 603–680.

Fan, J., Liao, Y. and Shi, X. (2015c). Risks of large portfolios. *Journal of Econometrics* **186** 367–387.

Fan, J., Liu, H. and Wang, W. (2018). Large covariance estimation through elliptical factor models. *The Annals of Statistics* **46** 1383–1414.

Fang, K. and Zhang, Y. T. (1990). *Generalized Multivariate Analysis*. Science Press.

Frahm, G., Junker, M. and Szimayer, A. (2003). Elliptical copulas: applicability and limitations. *Statistics and Probability Letters* **63** 275–286.

Han, F. and Liu, H. (2012). Transelliptical component analysis. *Advances in Neural Information Processing Systems* **25**.

Hardin, J. and Wilson, J. (2009). A note on oligonucleotide expression values not being normally distributed,. *Biostatistics* **10** 446–450.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhya* **32** 419–430.

Liu, L., Hawkins, D. M., Ghosh, S. and Young, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences* **100** 13167–13172.

Maruyama, Y. and Seo, T. (2003). Estimation of moment parameters in elliptical distributions. *Journal of Japan Statistical Society* **33** 215–229.

Posekany, A., Felsenstein, K., and Sykacek, P. (2011). Biological assessment of robust noise models in microarray data analysis. *Bioinformatics* **27** 807–814.

Ruttimann, U. E., Unser, M., Rawlings, R. R., Rio, D., Ramsey, N. F., Mattay, V. S., Hommer, D. W., Frank, J. A. and Weinberger, D. R. (1998). Statistical analysis of functional MRI data in the wavelet domain. *IEEE Transactions on Medical Imaging* **17** 142–154.

Sun, Q., Zhou, W.-X. and Fan, J. (2018+). Adaptive huber regression: Nonasymptotic optimality and phase transition.

A Proof of main results

A.1 Proof of Theorem 2.1

Write for short $\widehat{\theta}_m^{\mathrm{M}} = \widehat{\theta}_m^{\mathrm{M}}(\widehat{\boldsymbol{\mu}}, \mathrm{diag}(\widehat{\boldsymbol{\Sigma}}))$ and $\widetilde{\theta}_m^{\mathrm{M}} = \widehat{\theta}_m^{\mathrm{M}}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\Sigma}))$. By Theorem 2.3, $\widetilde{\theta}_m^{\mathrm{M}}$ is unbiased and satisfies

$$\mathrm{var}\big(\widetilde{\theta}_m^{\mathrm{M}}\big) \le \theta_m^2 \cdot \bigg(\frac{1}{n} \frac{r_{2m} - r_m^2}{r_m^2} + \frac{1}{np} \frac{r_{2m}}{r_m^2} \frac{\eta_{2m} - \eta_m^2}{\eta_m^2} + \frac{C_m}{n} \frac{r_{2m}}{r_m^2 \eta_m^2} \frac{\|\boldsymbol{\Lambda} - \mathbf{I}\|_F^2}{p^2} \bigg).$$

We note that (η_m, C_m) are constants, (θ_m, r_m, r_{2m}) are bounded above/below by constants, and all entries of the correlation matrix $\boldsymbol{\Lambda}$ are bounded by 1. Hence, the right hand side is $O(n^{-1})$, and it implies

$$|\widetilde{\theta}_m^{\mathrm{M}} - \theta_m| = O_{\mathbb{P}}(n^{-1/2}).$$

To show the claim, it suffices to show that

$$|\widehat{\theta}_m^{\mathrm{M}} - \widetilde{\theta}_m^{\mathrm{M}}| = O_{\mathbb{P}}(n^{-1/2}). \tag{17}$$

Below, we show (17). Write for short $X_{ij} = (Y_{ij} - \mu_j)/\sqrt{\sigma_{jj}}$, for $1 \le i \le n, 1 \le j \le p$. For any $k \ge 0$, let $S_{kj} = \frac{1}{n}\sum_{i=1}^n X_{ij}^k$. Using these notations,

$$\widetilde{\theta}_m^{\mathrm{M}} = \frac{1}{npc_m} \sum_{j=1}^p \sum_{i=1}^n \bigg(\frac{Y_{ij} - \mu_j}{\sqrt{\sigma_{jj}}} \bigg)^{2m} = \frac{1}{c_m} \cdot \frac{1}{p} \sum_{j=1}^p S_{(2m)j}.$$

At the same time, noticing that $(\widehat{\mu}_j - \mu_j)/\sqrt{\sigma_{jj}} = S_{1j}$ and $(Y_{ij} - \widehat{\mu}_j)/\sqrt{\sigma_{jj}} = X_{ij} - S_{1j}$, we have

$$\begin{aligned}
\widehat{\theta}_m^{\mathrm{M}} &= \frac{1}{npc_m} \sum_{j=1}^p \sum_{i=1}^n \bigg(\frac{Y_{ij} - \widehat{\mu}_j}{\sqrt{\widehat{\sigma}_{jj}}} \bigg)^{2m} \\
&= \frac{1}{npc_m} \sum_{j=1}^p \bigg[\frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m} \sum_{i=1}^n \bigg(\frac{Y_{ij} - \widehat{\mu}_j}{\sqrt{\sigma_{jj}}} \bigg)^{2m} \bigg] \\
&= \frac{1}{npc_m} \sum_{j=1}^p \bigg[\frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m} \sum_{i=1}^n (X_{ij} - S_{1j})^{2m} \bigg] \\
&= \frac{1}{npc_m} \sum_{j=1}^p \bigg[\frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m} \sum_{i=1}^n \sum_{k=0}^{2m} \gamma_k S_{1j}^k X_{ij}^{2m-k} \bigg], \qquad \text{where } \gamma_k \equiv (-1)^k \binom{2m}{k} \\
&= \frac{1}{c_m} \sum_{k=0}^{2m} \gamma_k \bigg[\frac{1}{p} \sum_{j=1}^p \frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m} S_{1j}^k \Big(\frac{1}{n} \sum_{i=1}^n X_{ij}^{2m-k} \Big) \bigg] \\
&= \frac{1}{c_m} \sum_{k=0}^{2m} \gamma_k \bigg[\frac{1}{p} \sum_{j=1}^p \frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m} S_{1j}^k S_{(2m-k)j} \bigg].
\end{aligned}$$

Combining the above gives

$$\begin{aligned}
\widehat{\theta}_m^{\mathrm{M}} - \widetilde{\theta}_m^{\mathrm{M}} = \frac{2m}{c_m} \frac{1}{p} \sum_{j=1}^p \Big(\frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m} - 1 \Big) S_{(2m)j} + \frac{2m}{c_m} \frac{1}{p} \sum_{j=1}^p \frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m} S_{1j} S_{(2m-1)j} \\
+ \frac{1}{c_m} \sum_{k=2}^{2m} \gamma_k \bigg[\frac{1}{p} \sum_{j=1}^p \frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m} S_{1j}^k S_{(2m-k)j} \bigg]
\end{aligned}$$

$$= (I_1) + (I_2) + (I_3). \tag{18}$$

To bound the right hand side of (18), we define an event. By (6), $Y_{ij} = \mu_j + \xi_i(\mathbf{\Sigma}^{1/2}\boldsymbol{U}_i)_j$. Let $\mathbf{\Lambda} = [\mathrm{diag}(\mathbf{\Sigma})]^{-1/2}\mathbf{\Sigma}[\mathrm{diag}(\mathbf{\Sigma})]^{-1/2}$. Then, $X_{ij} = \frac{Y_{ij}-\mu_j}{\sqrt{\sigma_{jj}}} = \xi_i(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i)_j$. It follows that

$$S_{kj} = \frac{1}{n}\sum_{i=1}^{n}\xi_i^k(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i)_j^k. \tag{19}$$

Note that $\mathbb{E}X_{ij} = (\mathbb{E}\xi_i^k)\mathbb{E}[(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i)_j^k]$. At the same time, since $X_{ij} \sim N(0,1)$ when $\xi_i^2 \sim \chi_p^2$, it holds that $\mathbb{E}[N^k(0,1)] = (\mathbb{E}\chi_p^k)\mathbb{E}[(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i)_j^k]$. Together, we have $\mathbb{E}(X_{ij}^k) = \mathbb{E}[N^k(0,1)] \cdot [(\mathbb{E}\xi_i^k)/(\mathbb{E}\chi_p^k)]$. Our assumption of $\theta_{2m} \le C$ guarantees $(\mathbb{E}\xi_i^k)/(\mathbb{E}\chi_p^k) \le C$ for $1 \le k \le 4m$. It follows that $\mathbb{E}(X_{ij}^k) \le C$ and $\mathrm{var}(X_{ij}^k) \le C$ for $1 \le k \le 2m$. As a result,

$$\mathbb{E}(|S_{kj}|^2) \le C, \qquad \mathbb{E}\big(|S_{kj} - \mathbb{E}S_{kj}|^2\big) = O(n^{-1}), \qquad 1 \le k \le 2m. \tag{20}$$

Using the marginal sub-Gaussianity, for any $\epsilon > 0$, there exists a constant $C > 0$ such that, with probability $\ge 1 - \epsilon$,

$$\max_{\substack{1\le k\le 2m\\ 1\le j\le p}} |S_{kj} - \mathbb{E}S_{kj}| \le C\sqrt{(\log p)/n}. \tag{21}$$

Let B be the event that (21) holds. To show (17), it suffices to show that

$$|\widehat{\theta}_m^{\mathrm{M}} - \widetilde{\theta}_m^{\mathrm{M}}| \cdot I_B = O_{\mathbb{P}}(n^{-1/2}). \tag{22}$$

We now show (22). Consider (I_2) and (I_3). By (19) and using that $(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i)_j$ has a symmetric distribution, we have $\mathbb{E}S_{kj} = 0$ for any odd k. As a result, over the event B, $|S_{1j}| \le C\sqrt{(\log p)/n}$, $|S_{(2m-1)j}| \le C\sqrt{(\log p)/n}$ and $|S_{(2m-k)j}| \le C$, for all $1 \le j \le p$ and $1 \le k \le 2m$. Additionally, since $\widehat{\sigma}_{jj}/\sigma_{jj} = \frac{1}{n}\sum_{i=1}^n\big(\frac{Y_{ij}-\widehat{\mu}_j}{\sqrt{\sigma_{jj}}}\big)^2 = \frac{1}{n}\sum_{i=1}^n(X_{ij} - S_{1j})^2 = S_{2j} - S_{1j}^2$, where $\mathbb{E}S_{2j} = 1$, it holds that $\sigma_{jj}/\widehat{\sigma}_{jj} \le C$ over the event B. It follows that

$$\begin{aligned}
|(I_2)| &\le C\max_{1\le j\le p}\left\{\frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m}|S_{1j}||S_{(2m-1)j}|\right\} = O(n^{-1}\log(p)).\\
|(I_3)| &\le C\max_{1\le j\le p}\left\{\sum_{k=2}^{2m}\frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m}|S_{1j}|^k|S_{(2m-k)j}|\right\} = O(n^{-1}\log(p)).
\end{aligned} \tag{23}$$

Consider (I_1). Since $\widehat{\sigma}_{jj}/\sigma_{jj} = S_{2j} - S_{1j}^2$, we write

$$\widehat{\sigma}_{jj}/\sigma_{jj} - 1 = (S_{2j} - \mathbb{E}S_{2j}) - S_{1j}^2.$$

Over the event B, $\max_j |S_{1j}| \le C\sqrt{(\log p)/n}$ and $\max_{1\le j\le p}|\widehat{\sigma}_{jj}/\sigma_{jj} - 1| \le C\sqrt{(\log p)/n}$. This means, for all $1 \le j \le p$, $\widehat{\sigma}_{jj}/\sigma_{jj}$ is contained in a diminishing neighborhood of 1. We use Taylor expansion of the function $(1+x)^{-m} - 1$. It gives

$$\begin{aligned}
\frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m} - 1 &= -m\frac{\widehat{\sigma}_{jj} - \sigma_{jj}}{\sigma_{jj}^m} + O(n^{-1}\log(p))\\
&= -m[(S_{2j} - \mathbb{E}S_{2j}) - S_{1j}^2] + O(n^{-1}\log(p))
\end{aligned}$$

$$= -m(S_{2j} - \mathbb{E}S_{2j}) + O(n^{-1}\log(p))$$
$$= -\frac{m}{n}\sum_{i=1}^{n}\Big\{\xi_i^2(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i)_j^2 - (\mathbb{E}\xi_i^2)\mathbb{E}[(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i)_j^2]\Big\} + O(n^{-1}\log(p)). \tag{24}$$

where the third line is due to $\max_{1\le j\le p}|S_{1j}| \le O(\sqrt{(\log p)/n})$ over the event B and the fourth line is due to (19). By (19) and (24),

$$\begin{aligned}(I_1) &= -\frac{2m^2}{c_m p}\sum_{j=1}^{p}\Big[\frac{1}{n}\sum_{i=1}^{n}\big(\xi_i^2(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i)_j^2 - (\mathbb{E}\xi_i^2)\mathbb{E}[(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i)_j^2]\big)\Big]\Big[\frac{1}{n}\sum_{k=1}^{n}\xi_k^{2m}(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_k)_j^{2m}\Big] + o(n^{-1/2})\\ &= -\frac{2m^2}{c_m p n^2}\sum_{i,k=1}^{n}\Big\{\sum_{j=1}^{p}\Big[\xi_i^2(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i)_j^2 - (\mathbb{E}\xi_i^2)\mathbb{E}[(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i)_j^2]\Big]\Big[\xi_k^{2m}(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_k)_j^{2m}\Big]\Big\} + o(n^{-1/2})\\ &\equiv -\frac{2m^2}{c_m p n^2}\sum_{i,k=1}^{n}Q_{ik} + o(n^{-1/2}).\end{aligned} \tag{25}$$

Write $R_{ij} = (\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i)_j$ for short. Then,

$$Q_{ik} = \sum_{j=1}^{p}\big[\xi_i^2 R_{ij}^2 - (\mathbb{E}\xi_i^2)(\mathbb{E}R_{ij}^2)\big]\xi_k^{2m}R_{kj}^{2m}. \tag{26}$$

We introduce positive random variables $\{\omega_i\}_{i=1}^n$ such that $\omega_i^2 \overset{iid}{\sim} \chi_p^2$ and that $\{\omega_i\}_{i=1}^n$ are independent of $\{(\xi_i, \boldsymbol{U}_i) : 1 \le i \le n\}$. Then, $Z_i \equiv \omega_i(\mathbf{\Lambda}^{1/2}\boldsymbol{U}_i) \sim N(0, I_p)$. For even integers s, t and $1 \le j, j' \le p$,

$$\mathbb{E}[Z_i^s(j)Z_i^t(j')] = \mathbb{E}(\omega_i^{s+t})\mathbb{E}(R_{ij}^s R_{ij'}^t).$$

For all s, t such that $s + t \le 4m$, the left hand side is uniformly bounded by a constant. Additionally, by elementary probability, $\mathbb{E}(\omega_i^{s+t}) \asymp p^{(s+t)/2}$. It follows that

$$\max_{1\le j,j'\le p}\mathbb{E}(R_{ij}^s R_{ij'}^t) \le Cp^{-(s+t)/2}, \qquad \text{for even } s, t \text{ such that } s + t \le 4m. \tag{27}$$

In particular, by taking $s = 2\ell$ and $t = 0$ in the above, we have $\mathbb{E}R_{ij}^{2\ell} \le Cp^{-\ell}$ for all $0 \le \ell \le 2m$. Additionally, $\theta_s = p^{-s}\mathbb{E}\xi^{2s}$ by definition, so the assumption $\theta_{2m} \le C$ guarantees

$$\mathbb{E}(\xi_i^{2s}) \le Cp^s, \qquad 0 \le s \le 2m. \tag{28}$$

Using (27)-(28), we first bound $|\sum_{i=1}^n Q_{ii}|$. It is seen that

$$\mathbb{E}|Q_{ii}| \le \sum_{j=1}^{p}\mathbb{E}(\xi_i^{2m+2})\mathbb{E}(R_{ij}^{2m+2}) + (\mathbb{E}\xi_i^2)(\mathbb{E}R_{ij}^2)\mathbb{E}(\xi_i^{2m})\mathbb{E}(R_{ij}^{2m}) \le Cp.$$

As a result,

$$\mathbb{E}\Big(\frac{1}{pn^2}\Big|\sum_{i=1}^{n}Q_{ii}\Big|\Big) = O(n^{-1}) \qquad \Longrightarrow \qquad \frac{1}{pn^2}\Big|\sum_{i=1}^{n}Q_{ii}\Big| = o_{\mathbb{P}}(n^{-1/2}). \tag{29}$$

We then bound $|\sum_{i\ne k}Q_{ik}|$. Consider (i, k, i', k') such that $i \ne k$ and $i' \ne k'$. By (26), $\mathbb{E}Q_{ik} = 0$ for $i \ne k$. Therefore, if (i, k, i', k') are mutually distinct, $\mathbb{E}(Q_{ik}Q_{i'k'}) = 0$. It follows that

$$\mathbb{E}\Big[\Big(\sum_{i\ne k}Q_{ik}\Big)^2\Big] = 6\sum_{\text{distinct } i,k,k'}\mathbb{E}(Q_{ik}Q_{ik'}) + 2\sum_{\text{distinct } i,k}\mathbb{E}(Q_{ik}^2).$$

By (26) and (27)-(28),

$$\begin{aligned}
\mathbb{E}(Q_{ik}Q_{ik'}) &= \mathbb{E}\Big\{ \sum_{j,j'=1}^{p} \big[\xi_i^2 R_{ij}^2 - (\mathbb{E}\xi_i^2)(\mathbb{E}R_{ij}^2)\big]\big[\xi_i^2 R_{ij'}^2 - (\mathbb{E}\xi_i^2)(\mathbb{E}R_{ij'}^2)\big]\xi_k^{2m}\xi_{k'}^{2m} R_{kj}^{2m} R_{k'j'}^{2m} \Big\} \\
&\le \sum_{j,j'=1}^{p} \mathbb{E}(\xi_i^4)\mathbb{E}(R_{ij}^2 R_{ij'}^2)(\mathbb{E}\xi_k^{2m})(\mathbb{E}\xi_{k'}^{2m})\mathbb{E}(R_{kj}^{2m} R_{k'j'}^{2m}) \quad \le Cp^2, \\
\mathbb{E}(Q_{ik}^2) &= \mathbb{E}\Big\{ \sum_{j,j'=1}^{p} \big[\xi_i^2 R_{ij}^2 - (\mathbb{E}\xi_i^2)(\mathbb{E}R_{ij}^2)\big]\big[\xi_i^2 R_{ij'}^2 - (\mathbb{E}\xi_i^2)(\mathbb{E}R_{ij'}^2)\big]\xi_k^{4m} R_{kj}^{2m} R_{kj'}^{2m} \Big\} \\
&\le \sum_{j,j'=1}^{p} \mathbb{E}(\xi_i^4)\mathbb{E}(R_{ij}^2 R_{ij'}^2)(\mathbb{E}\xi_k^{4m})\mathbb{E}(R_{kj}^{2m} R_{kj'}^{2m}) \quad \le Cp^2.
\end{aligned}$$

Moreover, the total number of such distinct (i,k,k') is $O(n^3)$. It follows that

$$\mathbb{E}\Big[\Big(\frac{1}{pn^2}\sum_{i\neq k} Q_{ik}\Big)^2\Big] = O(n^{-1}) \qquad \Longrightarrow \qquad \frac{1}{pn^2}\Big|\sum_{i\neq k} Q_{ik}\Big| = O_{\mathbb{P}}(n^{-1/2}). \tag{30}$$

Pluging (29) and (30) into (25) gives

$$(I_1) = O_{\mathbb{P}}(n^{-1/2}). \tag{31}$$

We further plug (23) and (31) into (18). It gives (22). The proof is now complete. □

A.2 Proof of Theorem 2.2

Similar to the proof of Theorem 2.1, let $\widetilde{\theta}_m^{\mathrm{M}}$ and $\widehat{\theta}_m^{\mathrm{M}}$ denote the MAE with true $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and estimates $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$; here, $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ are not necessarily the sample mean and sample covariance matrix. It follows from Theorem 2.3 that $\mathbb{E}[(\widetilde{\theta}_m^{\mathrm{M}} - \theta_m)^2] \le Cn^{-1}$. By Markov's inequality, for any constant $C_1 > 0$,

$$\mathbb{P}\Big(|\widetilde{\theta}_m^{\mathrm{M}} - \theta_m| > C_1 n^{-1/2}\Big) \le \frac{\mathbb{E}[(\widehat{\theta}_m^{\mathrm{M}} - \theta_m)^2]}{C_1^2 n^{-1}} \le \frac{C}{C_1^2}. \tag{32}$$

Hence, given $\epsilon > 0$, we can choose an appropriate $C_1 > 0$ such that the above probability is bounded by $\epsilon/3$.

Below, we bound $|\widehat{\theta}_m^{\mathrm{M}} - \widetilde{\theta}_m^{\mathrm{M}}|$. Letting $X_{ij} = (Y_{ij} - \mu_j)/\sqrt{\sigma_{jj}}$ and $\widehat{X}_{ij} = (Y_{ij} - \widehat{\mu}_j)/\sqrt{\widehat{\sigma}_{jj}}$, we have

$$\widehat{\theta}_m^{\mathrm{M}} - \widetilde{\theta}_m^{\mathrm{M}} = \frac{1}{npc_m}\sum_{j=1}^{p}\sum_{i=1}^{n}(\widehat{X}_{ij}^{2m} - X_{ij}^{2m}),$$

where

$$\widehat{X}_{ij} = X_{ij} + X_{ij}\left(\frac{\sqrt{\sigma_{jj}}}{\sqrt{\widehat{\sigma}_{jj}}} - 1\right) + \frac{\mu_j - \widehat{\mu}_j}{\sqrt{\widehat{\sigma}_{jj}}} \equiv X_{ij} + \Delta_{ij}.$$

It follows that

$$\begin{aligned}
\widehat{\theta}_m^{\mathrm{M}} - \widetilde{\theta}_m^{\mathrm{M}} &= \frac{1}{npc_m}\sum_{j=1}^{p}\sum_{i=1}^{n}\sum_{k=1}^{2m}\binom{2m}{k} X_{ij}^{2m-k}\Delta_{ij}^k \\
&= \frac{1}{npc_m}\sum_{j=1}^{p}\sum_{i=1}^{n} 2m X_{ij}^{2m-1}\Delta_{ij} + \sum_{k=2}^{2m}\binom{2m}{k}\cdot\frac{1}{npc_m}\sum_{j=1}^{p}\sum_{i=1}^{n} X_{ij}^{2m-k}\Delta_{ij}^k
\end{aligned}$$

$$\equiv (J_1) + (J_2).$$

First, we consider (J_1). By direct calculations,

$$\begin{aligned}(J_1) &= \frac{2m}{npc_m}\sum_{j=1}^{p}\sum_{i=1}^{n} X_{ij}^{2m}\Big(\frac{\sqrt{\sigma_{jj}}}{\sqrt{\widehat{\sigma}_{jj}}}-1\Big) + \frac{2m}{npc_m}\sum_{j=1}^{p}\sum_{i=1}^{n} X_{ij}^{2m-1}\frac{\mu_j-\widehat{\mu}_j}{\sqrt{\widehat{\sigma}_{jj}}} \\ &= \frac{2m}{pc_m}\sum_{j=1}^{p} S_{(2m)j}\Big(\frac{\sqrt{\sigma_{jj}}}{\sqrt{\widehat{\sigma}_{jj}}}-1\Big) + \frac{2m}{pc_m}\sum_{j=1}^{p} S_{(2m-1)j}\frac{\sqrt{\sigma_{jj}}}{\sqrt{\widehat{\sigma}_{jj}}}\frac{\mu_j-\widehat{\mu}_j}{\sqrt{\sigma_{jj}}},\end{aligned}$$

where $S_{kj} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}^k$ for $k \geq 0$. Under our assumption, $\max_j |\sigma_{jj}/\widehat{\sigma}_{jj}| \lesssim 1$, and $|\sqrt{\widehat{\sigma}_{jj}/\sigma_{jj}}-1| \leq C|\widehat{\sigma}_{jj}-\sigma_{jj}|$. Moreover, by similar technique in the proof of Theorem 2.3, we can prove that, $\frac{1}{p}\sum_{j=1}^{p}\mathbb{E}|S_{kj}| \leq C$, for $1 \leq k \leq 2m$. As a result, for any $\epsilon > 0$, there exists $C_2 > 0$ such that, $\frac{1}{p}\sum_{j=1}^{p}|S_{(2m)j}| \leq C_2$ simultaneously for $1 \leq k \leq 2m$, with probability $1-\epsilon/3$. On this event,

$$|(J_1)| \leq C\Big(\frac{1}{p}\sum_{j=1}^{p}|S_{(2m)j}|\Big)|\widehat{\sigma}_{jj}-\sigma_{jj}| + C\Big(\frac{1}{p}\sum_{j=1}^{p}|S_{(2m-1)j}|\Big)|\widehat{\mu}_j-\mu_j| \leq C\max\{\alpha_n,\beta_n\}. \tag{33}$$

Next, we consider (J_2). By our assumption, $|\Delta_{ij}| \leq \alpha_n + \beta_n|X_{ij}|$. It follows that

$$|\Delta_{ij}|^k \leq C\alpha_n^k + C\beta_n^k|X_{ij}|^k.$$

Plugging it into the definition of (J_2), we have

$$\begin{aligned}|(J_2)| &\leq C\sum_{k=2}^{2m}\frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}|X_{ij}|^{2m-k}(\alpha_n^k+\beta_n^k|X_{ij}|^k) \\ &\leq C\sum_{k=2}^{2m}\alpha_n^k\Big(\frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}|X_{ij}|^{2m-k}\Big) + C\sum_{k=2}^{2m}\beta_n^k\Big(\frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}|X_{ij}|^{2m}\Big).\end{aligned}$$

Again, we can easily prove that $\frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}\mathbb{E}|X_{ij}|^k \leq C$ for all $1 \leq k \leq 2m$. It follows that, for any $\epsilon > 0$, there exists $C_3 > 0$, such that $\frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}|X_{ij}|^k \leq C_3$ simultaneously for all $1 \leq k \leq 2m$. On this event,

$$|(J_2)| \leq C\sum_{k=2}^{2m}(\alpha_n^k+\beta_n^k) \leq C\max\{\alpha_n^2,\beta_n^2\}. \tag{34}$$

Combining (33)-(34) gives $|\widehat{\theta}_m^{\mathrm{M}} - \widetilde{\theta}_m^{\mathrm{M}}| \leq C\max\{\alpha_n,\beta_n\}$. We further combine it with (32). It gives the claim.

A.3 Proof of Theorem 2.3

Write for short $\widehat{\theta}_m^{\mathrm{M}} = \widehat{\theta}_m^{\mathrm{M}}(\boldsymbol{\mu},\boldsymbol{\Omega})$ and $\widehat{\theta}_{m,j}^{\mathrm{M}} = \widehat{\theta}_{m,j}^{\mathrm{M}}(\mu_j,\sigma_{jj})$. First, we show that $\widehat{\theta}_m^{\mathrm{M}}$ is unbiased. Recall that $\widehat{\theta}_m^{\mathrm{M}} = p^{-1}\sum_{j=1}^{p}\widehat{\theta}_{m,j}^{\mathrm{M}}$. It suffices to show $\widehat{\theta}_{m,j}^{\mathrm{M}}$ is unbiased for each $1 \leq j \leq p$. Recall that

$$\widehat{\theta}_{m,j}^{\mathrm{M}} = \frac{1}{nc_m}\sum_{i=1}^{n}\frac{(Y_{ij}-\mu_j)^{2m}}{\sigma_{jj}^m}, \qquad \text{where} \quad c_m = (2m-1)!!\,(p/2)^m\frac{\Gamma(p/2)}{\Gamma(p/2+m)}. \tag{35}$$

By the form of elliptical distribution, $\boldsymbol{Y}_i - \boldsymbol{\mu} = \xi_i\widetilde{\boldsymbol{U}}_i$, where ξ_i and $\widetilde{\boldsymbol{U}}_i$ are independent of each other. We have seen in Section 1.2 that $\mathbb{E}\widetilde{U}_{ij}^{2m} = p^{-m}c_m\sigma_{jj}^m$. It follows that

$$\mathbb{E}[(Y_{ij}-\mu_j)^{2m}] = (\mathbb{E}\xi_i^{2m})(\mathbb{E}\widetilde{U}_i^{2m}) = (p^m\theta_m)(p^{-m}c_m\sigma_{jj}^m) = c_m\theta_m\sigma_{jj}^m.$$

Plugging it into (35) gives

$$\mathbb{E}\widehat{\theta}^{\mathrm{M}}_{m,j} = \frac{1}{nc_m}\sum_{i=1}^{n}\frac{\mathbb{E}[(Y_{ij}-\mu_j)^{2m}]}{\sigma_{jj}^m} = \frac{1}{nc_m}\sum_{i=1}^{n}\frac{c_m\theta_m\sigma_{jj}^m}{\sigma_{jj}^m} = \theta_m. \tag{36}$$

This proves that each $\widehat{\theta}^{\mathrm{M}}_{m,j}$ is unbiased. It follows that $\widehat{\theta}^{\mathrm{M}}_{m}$ is also unbiased.

Next, we calculate the variance of $\widehat{\theta}^{\mathrm{M}}_{m}$. For each $1 \le i \le n$, let $\boldsymbol{W}^{(i)} = (W_1^{(i)},\ldots,W_p^{(i)})^T$, where $W_j^{(i)} = (Y_{ij}-\mu_j)^{2m}/\sigma_{jj}^m$, $1 \le j \le p$. Noting that $\{\mathbf{W}^{(i)}\}_{i=1}^n$ are *iid* random vectors, we have

$$\mathrm{var}\big(\widehat{\theta}^{\mathrm{M}}_{m}\big) = \mathrm{var}\bigg(\frac{1}{npc_m}\sum_{j=1}^{p}\sum_{i=1}^{n}W_j^{(i)}\bigg) = \frac{1}{n}\mathrm{var}\bigg(\frac{1}{pc_m}\sum_{j=1}^{p}W_j^{(1)}\bigg). \tag{37}$$

It suffices to calculate the variance in the case of $n=1$. From now on, we fix $n=1$. Let $\boldsymbol{Y} = \boldsymbol{\mu}+\xi\boldsymbol{U}$ be the observed realization of the elliptical distribution. Write

$$\widehat{\theta}^{\mathrm{M}}_{m} = \frac{1}{c_m p}\sum_{j=1}^{p}W_j, \qquad \text{where} \quad W_j \equiv \frac{(Y_j-\mu_j)^{2m}}{\sigma_{jj}^m}.$$

We now calculate $\mathrm{var}(W_j)$ and $\mathrm{cov}(W_j,W_k)$. Recalling that $\widetilde{\boldsymbol{U}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{U}$, we define random vectors

$$\boldsymbol{Z} \equiv \chi_p^2\cdot\widetilde{\boldsymbol{U}} \qquad \text{and} \qquad \widetilde{\boldsymbol{Z}} \equiv \mathrm{diag}(\boldsymbol{\Sigma})^{-1/2}\boldsymbol{Z}, \tag{38}$$

where χ_p^2 is a chi-square random variable independent of $\widetilde{\boldsymbol{U}}$. Since the multivariate normal distribution is a special elliptical distribution with $\xi \sim \chi_p^2$, we immediately have $\boldsymbol{Z} \sim N(\mathbf{0},\boldsymbol{\Sigma})$. It follows that $\mathbb{E}\widetilde{Z}_j^{2m} = \sigma_{jj}^{-m}(\mathbb{E}Z_j^{2m}) = \sigma_{jj}^{-m}(\mathbb{E}\chi_p^{2m})(\mathbb{E}\widetilde{U}_j^{2m})$. At a result, for all $m \ge 1$,

$$\mathbb{E}[(Y_j-\mu_j)^{2m}] = (\mathbb{E}\xi^{2m})(\mathbb{E}\widetilde{U}_j^{2m}) = \mathbb{E}\xi^{2m}\cdot\frac{\sigma_{jj}^m(\mathbb{E}\widetilde{Z}_j^{2m})}{\mathbb{E}\chi_p^{2m}} = \sigma_{jj}^m\cdot r_m\mathbb{E}\widetilde{Z}_j^{2m}.$$

It follows that

$$\begin{aligned}\mathrm{var}(W_j) &= \frac{\mathbb{E}[(Y_j-\mu_j)^{4m}]}{\sigma_{jj}^{2m}} - \bigg(\frac{\mathbb{E}[(Y_j-\mu_j)^{2m}]}{\sigma_{jj}^m}\bigg)^2 \\ &= r_{2m}(\mathbb{E}\widetilde{Z}_j^{4m}) - r_m^2(\mathbb{E}\widetilde{Z}_j^{2m})^2 \\ &= r_{2m}\cdot\mathrm{var}(\widetilde{Z}_j^{2m}) + (r_{2m}-r_m^2)\cdot(\mathbb{E}\widetilde{Z}_j^{2m})^2. \end{aligned} \tag{39}$$

Similarly, since $Y_j-\mu_j = \xi\widetilde{U}_j$ and $Z_j = \chi_p^2\widetilde{U}_j$, we have

$$\begin{aligned}\mathbb{E}\big[(Y_j-\mu_j)^{2m}(Y_k-\mu_k)^{2m}\big] &= (\mathbb{E}\xi^{4m})\big(\mathbb{E}[\widetilde{U}_j^{2m}\widetilde{U}_k^{2m}]\big) \\ &= (r_{2m}\mathbb{E}\chi_p^{4m})\big(\mathbb{E}[\widetilde{U}_j^{2m}\widetilde{U}_k^{2m}]\big) \\ &= r_{2m}\mathbb{E}[Z_j^{2m}Z_k^{2m}] \\ &= \sigma_{jj}^m\sigma_{kk}^m\cdot r_{2m}\mathbb{E}[\widetilde{Z}_j^{2m}\widetilde{Z}_k^{2m}].\end{aligned}$$

Therefore,

$$\mathrm{cov}(W_j,W_k) = \frac{\mathbb{E}\big[(Y_j-\mu_j)^{2m}(Y_k-\mu_k)^{2m}\big]}{\sigma_{jj}^{2m}\sigma_{kk}^{2m}} - \frac{\mathbb{E}[(Y_j-\mu_j)^{2m}]}{\sigma_{jj}^m}\frac{\mathbb{E}[(Y_k-\mu_k)^{2m}]}{\sigma_{kk}^m}$$

$$= r_{2m}\mathbb{E}[\widetilde{Z}_j^{2m}\widetilde{Z}_k^{2m}] - r_m^2(\mathbb{E}\widetilde{Z}_j^{2m})(\mathbb{E}\widetilde{Z}_k^{2m})$$
$$= r_{2m}\cdot\mathrm{cov}(\widetilde{Z}_j^{2m},\widetilde{Z}_k^{2m}) + (r_{2m}-r_m^2)\cdot(\mathbb{E}\widetilde{Z}_j^{2m})(\mathbb{E}\widetilde{Z}_k^{2m}). \tag{40}$$

Combining (39) and (40) and noting that $\tilde{Z}_j \sim N(0,1)$ for all $1 \le j \le p$, we rewrite

$$\mathrm{cov}(W_j, W_k) = r_{2m}\,\mathrm{cov}(\widetilde{Z}_j^{2m},\widetilde{Z}_k^{2m}) + (r_{2m}-r_m^2)\eta_m^2, \qquad \text{where} \quad \eta_m = \mathbb{E}[N(0,1)^{2m}].$$

As a result,

$$\begin{aligned}\mathrm{var}(\widehat{\theta}_m^{\mathrm{M}}) &= \frac{1}{c_m^2 p^2}\sum_{1\le j,k\le p}\mathrm{cov}(W_j,W_k)\\ &= \frac{1}{c_m^2 p^2}\Big[r_{2m}\sum_{1\le j,k\le p}\mathrm{cov}(\widetilde{Z}_j^{2m},\widetilde{Z}_k^{2m}) + (r_{2m}-r_m^2)p^2\eta_m^2\Big]\\ &= \frac{1}{c_m^2}\bigg[r_{2m}\cdot\frac{1}{p^2}\,\mathrm{var}\Big(\sum_{j=1}^p \widetilde{Z}_j^{2m}\Big) + (r_{2m}-r_m^2)\cdot\eta_m^2\bigg].\end{aligned} \tag{41}$$

Moreover, since $\mathbb{E}\widetilde{U}_{ij}^{2m} = p^{-m}c_m\sigma_{jj}^m$ and $\mathbb{E}\widetilde{Z}_j^{2m} = \sigma_{jj}^{-m}(\mathbb{E}\chi_p^{2m})(\mathbb{E}\widetilde{U}_j^{2m})$, we have

$$c_m = \frac{p^m\mathbb{E}\widetilde{Z}_j^{2m}}{\mathbb{E}\chi_p^{2m}} = \frac{p^m\mathbb{E}[N(0,1)^{2m}]}{r_m^{-1}\mathbb{E}\xi^{2m}} = \frac{p^m\eta_m}{r_m^{-1}(p^m\theta_m)} = \frac{\eta_m r_m}{\theta_m}.$$

Plugging it into (41) gives

$$\mathrm{var}(\widehat{\theta}_m^{\mathrm{M}}) = \theta_m^2\bigg[\frac{r_{2m}}{r_m^2}\frac{\mathrm{var}\big(\sum_{j=1}^p\widetilde{Z}_j^{2m}\big)}{p^2\eta_m^2} + \frac{r_{2m}-r_m^2}{r_m^2}\bigg].$$

This is for the case of $n=1$. For a general n, we combine it with (37) to get

$$\frac{\mathrm{var}(\widehat{\theta}_m^{\mathrm{M}})}{\theta_m^2} = \frac{1}{n}\bigg[\frac{r_{2m}}{r_m^2}\frac{\mathrm{var}\big(\sum_{j=1}^p\widetilde{Z}_j^{2m}\big)}{p^2\eta_m^2} + \frac{r_{2m}-r_m^2}{r_m^2}\bigg]. \tag{42}$$

What remains is to calculate the variance of $\sum_{j=1}^p \widetilde{Z}_j^{2m}$. By definition,

$$\widetilde{\boldsymbol{Z}} \sim N(\boldsymbol{0},\boldsymbol{\Lambda}), \qquad \text{where} \quad \boldsymbol{\Lambda} = [\mathrm{diag}(\boldsymbol{\Sigma})]^{-1/2}\boldsymbol{\Sigma}[\mathrm{diag}(\boldsymbol{\Sigma})]^{-1/2}.$$

Here $\boldsymbol{\Lambda}$ coincides with the correlation matrix of the elliptical distribution. It is seen that

$$\begin{aligned}\mathrm{var}\Big(\sum_{j=1}^p\widetilde{Z}_j^{2m}\Big) &= \sum_{j=1}^p\mathrm{var}(\widetilde{Z}_j^{2m}) + 2\sum_{1\le j<k\le p}\mathrm{cov}(\widetilde{Z}_j^{2m},\widetilde{Z}_k^{2m})\\ &= p(\eta_{2m}-\eta_m^2) + 2\sum_{1\le j<k\le p}\beta_m(\Lambda_{jk}),\end{aligned}$$

where $\beta_m(\Lambda_{jk})$ denotes the covariance between X_1^{2m} and X_2^{2m} when $(X_1,X_2)^{\mathrm{T}}$ follows a bivariate normal distribution with covariances $\mathrm{var}(X_1)=\mathrm{var}(X_2)=1$ and $\mathrm{cov}(X_1,X_2)=\Lambda_{jk}$. The following lemma is proved in Section B.1:

Lemma A.1. *Let $\boldsymbol{X} = (X_1, X_2)^T$ be a bivariate normal random vector satisfying $\mathbb{E}(X_1^2) = \mathbb{E}(X_2^2) = 1$ and $\mathrm{cov}(X_1, X_2) = \rho$. Let $\eta_m = \mathbb{E}[N(0,1)^{2m}]$ and $\beta_m(\rho) = \mathrm{cov}(X_1^{2m}, X_2^{2m})$ for $m \geq 2$. Define*

$$B_m(s) = \sum_{\substack{1 \leq k_1, k_2 \leq m \\ k_1 + k_2 = s}} \binom{2m}{2k_1} \binom{2m}{2k_2} \cdot \eta_{m-k_1} \eta_{m-k_2} (\eta_s - \eta_{k_1} \eta_{k_2}), \quad s = 2, 3, \ldots, m$$

Then, for all $m \geq 2$,

$$\beta_m(\rho) = \sum_{s=2}^{m} B_m(s)(1 - |\rho|)^{m-s} |\rho|^s.$$

As a result, $\beta_m(\rho) = 72\rho^2$ for $m = 2$, and $\beta_m(\rho) \leq C_m \rho^2$ for $m \geq 3$, where $C_m > 0$ is a constant that only depends on m.

By Lemma A.1,

$$\mathrm{var}\Big(\sum_{j=1}^{p} \widetilde{Z}_j^{2m}\Big) \leq p(\eta_{2m} - \eta_m^2) + 2C_m \sum_{1 \leq j < k \leq p} \Lambda_{jk}^2 \leq p(\eta_{2m} - \eta_m^2) + C_m \|\boldsymbol{\Lambda} - \mathbf{I}\|_F^2. \tag{43}$$

Plugging it into (42) gives

$$\frac{\mathrm{var}(\widehat{\theta}_m^{\mathrm{M}})}{\theta_m^2} \leq \frac{1}{n} \frac{r_{2m} - r_m^2}{r_m^2} + \frac{1}{np} \frac{r_{2m}}{r_m^2} \bigg(\frac{\eta_{2m} - \eta_m^2}{\eta_m^2} + \frac{C_m}{\eta_m^2} \frac{\|\boldsymbol{\Lambda} - \mathbf{I}\|_F^2}{p} \bigg).$$

Moreover, for $m = 2$, the equality holds for $C_m = 72$. Since $\eta_m = 3$ and $\eta_{2m} = 105$, we have

$$\frac{\mathrm{var}(\widehat{\theta}_m^{\mathrm{M}})}{\theta_m^2} = \frac{1}{n} \frac{r_{2m} - r_m^2}{r_m^2} + \frac{1}{np} \frac{r_{2m}}{r_m^2} \Big(\frac{32}{3} + \frac{8\|\boldsymbol{\Lambda} - \mathbf{I}\|_F^2}{p} \Big), \qquad \text{for } m = 2.$$

A.4 Proof of Proposition 2.1

Write $\widehat{\theta}_m^{\mathrm{I}} = \widehat{\theta}_m^{\mathrm{I}}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ for short. By definition, $\widehat{\theta}_m^{\mathrm{I}} = \frac{1}{np^m} \sum_{i=1}^{n} \xi_i^{2m}$, and $\theta_m = p^{-m} \mathbb{E}(\xi^{2m})$. Therefore

$$\mathrm{var}(\widehat{\theta}_m^{\mathrm{I}}) = \frac{1}{np^{2m}} \mathrm{var}(\xi^{2m}) = \frac{1}{n}(\theta_{2m} - \theta_m^2).$$

We divide both sides by θ_m^2 and note that $\theta_m = p^{-m}(\mathbb{E}\xi^{2m}) = p^{-m} r_m (\mathbb{E}\chi_p^{2m})$. It follows that

$$\begin{aligned} \frac{\mathrm{var}(\widehat{\theta}_m^{\mathrm{I}})}{\theta_m^2} &= \frac{1}{n} \cdot \frac{r_{2m}(\mathbb{E}\chi_p^{4m}) - r_m^2 (\mathbb{E}\chi_p^{2m})^2}{r_m^2 (\mathbb{E}\chi_p^{2m})^2} \\ &= \frac{1}{n} \cdot \frac{r_{2m} \mathrm{var}(\chi_p^{2m}) + (r_{2m} - r_m^2)(\mathbb{E}\chi_p^{2m})^2}{r_m^2 (\mathbb{E}\chi_p^{2m})^2} \\ &= \frac{1}{n} \left[\frac{r_{2m}}{r_m^2} \frac{\mathrm{var}(\chi_p^{2m})}{(\mathbb{E}\chi_p^{2m})^2} + \frac{r_{2m} - r_m^2}{r_m^2} \right], \end{aligned} \tag{44}$$

By elementary statistics, $\mathbb{E}\chi_p^{2m} = \prod_{j=0}^{m-1}(p + 2j)$. As a result,

$$\frac{\mathrm{var}(\chi_p^{2m})}{(\mathbb{E}\chi_p^{2m})^2} = \frac{\prod_{j=0}^{2m-1}(p + 2j) - \prod_{j=0}^{m-1}(p + 2j)^2}{(\mathbb{E}\chi_p^{2m})^2}$$

$$
\begin{aligned}
&= \frac{\prod_{j=0}^{m-1}(p+2j)}{(\mathbb{E}\chi_p^{2m})^2}\Big\{\prod_{j=m}^{2m-1}(p+2j) - \prod_{j=0}^{m-1}(p+2j)\Big\} \\
&= \frac{1}{\mathbb{E}\chi_p^{2m}}\bigg\{\Big[p^m + \big(p^{m-1}\sum_{j=m}^{2m-1} 2j\big)\Big] - \Big[p^m + \big(p^{m-1}\sum_{j=0}^{m-1} 2j\big)\Big] + O(p^{m-2})\bigg\} \\
&= \frac{1}{\mathbb{E}\chi_p^{2m}} \cdot \big[2m^2 p^{m-1} + O(p^{m-2})\big] \\
&= \frac{2m^2}{p}[1+o(1)]. \qquad (45)
\end{aligned}
$$

Plugging (45) into (44) gives the claim.

A.5 Proof of Theorem 2.4

Fix $1 \le j \le p$. Using the Slutsky's lemma, we only need to prove

$$
\frac{\widehat{\theta}^{\mathrm{M}}_{m,j}(\widehat{\mu}_j, \widehat{\sigma}_{jj}) - \theta_m}{\sqrt{\frac{c_{2m}}{c_m^2}\theta_{2m} - \theta_m^2}} \to_d N(0,1). \qquad (46)
$$

Write for short $\widehat{\theta}^{\mathrm{M}}_{m,j} = \widehat{\theta}^{\mathrm{M}}_{m,j}(\widehat{\mu}_j, \widehat{\sigma}_{jj})$. Let $X_{ij} = (Y_{ij} - \mu_j)/\sqrt{\sigma_{jj}}$ and $S_{kj} = \frac{1}{n}\sum_{i=1}^n X_{ij}^k$, for $1 \le i \le n$ and $k \ge 0$. Then, $\widehat{\mu}_j = S_{1j}$, $\widehat{\sigma}_{jj} = S_{2j} - S_{1j}^2$, and

$$
\frac{Y_{ij} - \widehat{\mu}_j}{\sqrt{\widehat{\sigma}_{jj}}} = \frac{\sqrt{\sigma_{jj}}}{\sqrt{\widehat{\sigma}_{jj}}}\frac{Y_{ij} - \widehat{\mu}_j}{\sqrt{\sigma_{jj}}} = \frac{\sqrt{\sigma_{jj}}}{\sqrt{\widehat{\sigma}_{jj}}}(X_{ij} - S_{1j}).
$$

It follows that

$$
\begin{aligned}
\widehat{\theta}^{\mathrm{M}}_{m,j} &= \frac{1}{nc_m}\sum_{i=1}^n\bigg(\frac{Y_{ij} - \widehat{\mu}_j}{\sqrt{\widehat{\sigma}_{jj}}}\bigg)^{2m} = \frac{1}{nc_m}\frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m}\sum_{i=1}^n (X_{ij} - S_{1j})^{2m} \\
&= \frac{1}{nc_m}\frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m}\sum_{i=1}^n\sum_{k=0}^{2m}\gamma_k S_{1j}^k X_{ij}^{2m-k}, \qquad \text{where } \gamma_k \equiv (-1)^k\binom{2m}{k} \\
&= \frac{1}{c_m}\frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m}\sum_{k=0}^{2m}\gamma_k S_{1j}^k S_{(2m-k)j}. \qquad (47)
\end{aligned}
$$

Let $\boldsymbol{S} = (S_{1j}, S_{2j}, \ldots, S_{(2m)j})^{\mathrm{T}}$. Below, we first derive the asymptotic normality of $\boldsymbol{S}$, then we use the delta method to prove (46).

First, we study the random vector $\boldsymbol{S}$. It is not hard to see that $\mathbb{E}S_{kj} = \mathbb{E}X_{ij}^k$. By (6), $X_{ij} = \xi_i(\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}_i)_j$, where $\{(\xi_i, \boldsymbol{U}_i)\}_{i=1}^n$ are mutually indepependent and $\boldsymbol{\Lambda} = [\mathrm{diag}(\boldsymbol{\Sigma})]^{-1/2}\boldsymbol{\Sigma}[\mathrm{diag}(\boldsymbol{\Sigma})]^{-1/2}$ is the correlation matrix. Since $X_{ij} \sim N(0,1)$ when $\xi_i \sim \chi_p^2$, the symmetry of $N(0,1)$ implies that $(\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}_i)_j$ has a symmetric distribution. Hence, $\mathbb{E}X_{ij}^k = 0$ for an odd k. For an even $k = 2s$, by definition of c_m in (7), $\mathbb{E}[(\boldsymbol{\Lambda}^{1/2}\boldsymbol{U})_j^{2s}] = p^{-s}c_s$; also, $\mathbb{E}(\xi_i^{2s}) = p^s\theta_s$; combining them gives $\mathbb{E}X_{ij}^{2s} = \mathbb{E}(\xi_i^{2s})\mathbb{E}[(\boldsymbol{\Lambda}^{1/2}\boldsymbol{U})_j^{2s}] = c_s\theta_s$. It follows that

$$
\mathbb{E}(S_k) = \begin{cases} 0, & k \text{ is odd}, \\ c_{k/2}\theta_{k/2}, & k \text{ is even}. \end{cases} \qquad (48)
$$

Moreover, $\mathrm{cov}(S_{kj}, S_{\ell j}) = \frac{1}{n}\mathrm{cov}(X_{ij}^k, X_{ij}^\ell) = \frac{1}{n}[\mathbb{E}X_{ij}^{k+\ell} - (\mathbb{E}X_{ij}^k)(\mathbb{E}X_{ij}^\ell)]$. It follows that

$$\mathrm{Cov}(S_k, S_\ell) = \frac{1}{n}\begin{cases} 0, & k \text{ is odd, } \ell \text{ is even,} \\ c_{(k+\ell)/2}\theta_{(k+\ell)/2}, & k \text{ and } \ell \text{ are odd,} \\ c_{(k+\ell)/2}\theta_{(k+\ell)/2} - c_{k/2}\theta_{k/2}c_{\ell/2}\theta_{\ell/2}, & k \text{ and } \ell \text{ are even.} \end{cases} \tag{49}$$

By classical central limit theorem,

$$\sqrt{n}[\mathrm{cov}(\boldsymbol{S})]^{-1/2}(\boldsymbol{S} - \mathbb{E}\boldsymbol{S}) \to_d N\big(\mathbf{0}, \mathbf{I}_{2m}\big). \tag{50}$$

Next, we prove (46). Define a function $h : \mathbb{R}^{2m} \to \mathbb{R}$ by $h(\mathbf{x}) = \sum_{k=0}^{2m} \gamma_k x_1^k x_{2m-k}$. By (48),

$$h(\mathbb{E}\boldsymbol{S}) = \sum_{k=0}^{2m} \gamma_k (\mathbb{E}S_{1j})^k \mathbb{E}S_{(2m-k)j} = \mathbb{E}[S_{(2m)j}] = c_m\theta_m.$$

Note that $\frac{\partial}{\partial x_1}h(\mathbf{x}) = \sum_{k=1}^{2m} k\gamma_k x_1^{k-1}x_{2m-k}$, and $\frac{\partial}{\partial x_k}h(\mathbf{x}) = \gamma_{2m-k}x_1^{2m-k}$ for $k \neq 1$. Combining them with (48) and (49) gives

$$\bigtriangledown h(\mathbb{E}\boldsymbol{S}) = (0, 0, \ldots, 0, 1)^{\mathrm{T}}, \qquad [\bigtriangledown h(\mathbb{E}\boldsymbol{S})]^{\mathrm{T}}\,\mathrm{cov}(\boldsymbol{S})\,[\bigtriangledown h(\mathbb{E}\boldsymbol{S})] = c_{2m}\theta_{2m} - c_m^2\theta_m^2.$$

We then apply the delta method and obtain

$$\frac{\sqrt{n}[h(\boldsymbol{S}) - c_m\theta_m]}{\sqrt{c_{2m}\theta_{2m} - c_m^2\theta_m^2}} \to_d N(0, 1). \tag{51}$$

By (47), $\widehat{\theta}_{m,j}^{\mathrm{M}} = \frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m} \cdot c_m^{-1}h(\boldsymbol{S})$. Since $\frac{\sigma_{jj}^m}{\widehat{\sigma}_{jj}^m} \to 1$ in probability, using the Slutsky's lemma, we have

$$\frac{\sqrt{n}(\widehat{\theta}_{m,j}^{\mathrm{M}} - \theta_m)}{\frac{1}{c_m}\sqrt{c_{2m}\theta_{2m} - c_m^2\theta_m^2}} \to_d N(0, 1).$$

This proves (46).

A.6 Proof of Theorem 3.1

Write for short $\widehat{\theta}_m^{\mathrm{B}} = \widehat{\theta}_m^{\mathrm{B}}\big(\widehat{\boldsymbol{\mu}},\ \mathrm{diag}_{\mathcal{A}}(\widehat{\boldsymbol{\Sigma}})\big)$ and $\widetilde{\theta}_m^{\mathrm{B}} = \widehat{\theta}_m^{\mathrm{B}}\big(\boldsymbol{\mu},\ \mathrm{diag}_{\mathcal{A}}(\boldsymbol{\Sigma})\big)$. It follows from Theorem 3.3 that $\mathbb{E}[(\widetilde{\theta}_m^{\mathrm{B}} - \theta_m)^2] = O(n^{-1/2})$. This implies $|\widetilde{\theta}_m^{\mathrm{B}} - \theta_m| = O_{\mathbb{P}}(n^{-1/2})$. Hence, it suffices to show

$$|\widehat{\theta}_m^{\mathrm{B}} - \widetilde{\theta}_m^{\mathrm{B}}| = O_{\mathbb{P}}(n^{-1/2}). \tag{52}$$

First, we derive an expression of $\widehat{\theta}_m^{\mathrm{B}} - \widetilde{\theta}_m^{\mathrm{B}}$. Let $\boldsymbol{X}_{i,J} = \boldsymbol{\Sigma}_{J,J}^{-1/2}(\boldsymbol{Y}_{i,J} - \boldsymbol{\mu}_J)$ and $\widehat{\boldsymbol{X}}_{i,J} = \widehat{\boldsymbol{\Sigma}}_{J,J}^{-1/2}(\boldsymbol{Y}_{i,J} - \widehat{\boldsymbol{\mu}}_J)$ for all $1 \le i \le n$ and $J \in \mathcal{A}$. Then,

$$\widehat{\theta}_m^{\mathrm{B}} = \frac{1}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\sum_{i=1}^n \frac{\|\boldsymbol{X}_{i,J}\|^2}{c_{m,|J|}^*}, \qquad \widetilde{\theta}_m^{\mathrm{B}} = \frac{1}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\sum_{i=1}^n \frac{\|\widehat{\boldsymbol{X}}_{i,J}\|^2}{c_{m,|J|}^*}. \tag{53}$$

Let $\boldsymbol{S}_{1,J} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_{i,J}$ and $\boldsymbol{S}_{2,J} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_{i,J}\boldsymbol{X}_{i,J}^{\mathrm{T}}$. By direct calculations,

$$\boldsymbol{\Sigma}^{-1/2}(\widehat{\boldsymbol{\mu}}_J - \boldsymbol{\mu}_J) = \boldsymbol{S}_{1,J}, \qquad \boldsymbol{\Sigma}_{J,J}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{J,J}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{S}_{2,J} - \boldsymbol{S}_{1,J}\boldsymbol{S}_{1,J}^{\mathrm{T}}.$$

Define an event B such that

$$\begin{cases} \max_{1\le i\le n, J\in\mathcal{A}} \|\boldsymbol{X}_{i,J}\| \le C\sqrt{\log(n\vee p)}, \\ \max_{1\le i\le n, J\in\mathcal{A}, 1\le k\le 4m} \big|\|\boldsymbol{X}_{i,J}\|^k - \mathbb{E}\|\boldsymbol{X}_{i,J}\|^k\big| \le C\sqrt{\log(n\vee p)}, \\ \max_{J\in\mathcal{A}} \|\boldsymbol{S}_{1,J}\| \le C\sqrt{(\log p)/n}, \\ \max_{J\in\mathcal{A}} \|\boldsymbol{S}_{2,J} - \mathbb{E}\boldsymbol{S}_{2,J}\| \le C\sqrt{(\log p)/n}. \end{cases} \tag{54}$$

It is not hard to see that the event B holds with probability $1-o(1)$ (see the proof of Theorem 2.1 for similar arguments). On the event B, noting that $\mathbb{E}\boldsymbol{S}_{2,J} = \mathbf{I}_{|J|}$, we have

$$\begin{aligned} (\boldsymbol{\Sigma}_{J,J}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{J,J}\boldsymbol{\Sigma}^{-1/2})^{-1} &= \big[\mathbf{I}_{|J|} + (\boldsymbol{S}_{2,J} - \mathbb{E}\boldsymbol{S}_{2,J}) - \boldsymbol{S}_{1,J}\boldsymbol{S}_{1,J}^{\mathrm{T}}\big]^{-1} \\ &= \mathbf{I}_{|J|} - (\boldsymbol{S}_{2,J} - \mathbb{E}\boldsymbol{S}_{2,J}) + O(n^{-1}\log(p)). \end{aligned}$$

It follows that

$$\begin{aligned} \|\widehat{\boldsymbol{X}}_{i,J}\|^2 &= (\boldsymbol{Y}_{i,J} - \widehat{\boldsymbol{\mu}}_J)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{J,J}^{-1}(\boldsymbol{Y}_{i,J} - \widehat{\boldsymbol{\mu}}_J) \\ &= \big[\boldsymbol{\Sigma}_{J,J}^{-1/2}(\boldsymbol{Y}_{i,J} - \widehat{\boldsymbol{\mu}}_J)\big]^{\mathrm{T}}\big[\boldsymbol{\Sigma}_{J,J}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{J,J}\boldsymbol{\Sigma}^{-1/2}\big]^{-1}\big[\boldsymbol{\Sigma}_{J,J}^{-1/2}(\boldsymbol{Y}_{i,J} - \widehat{\boldsymbol{\mu}}_J)^{\mathrm{T}}\big] \\ &= (\boldsymbol{X}_{i,J} - \boldsymbol{S}_{1,J})^{\mathrm{T}}\big\{\mathbf{I}_{|J|} + (\boldsymbol{S}_{2,J} - \mathbb{E}\boldsymbol{S}_{2,J})\big\}(\boldsymbol{X}_{i,J} - \boldsymbol{S}_{1,J}) + O(n^{-1}\log^2(n\vee p)) \\ &= \|\boldsymbol{X}_{i,J}\|^2 \underbrace{-2\boldsymbol{S}_{1,J}^{\mathrm{T}}\boldsymbol{X}_{i,J} + \boldsymbol{X}_{i,J}^{\mathrm{T}}(\boldsymbol{S}_{2,J} - \mathbb{E}\boldsymbol{S}_{2,J})\boldsymbol{X}_{i,J}^{\mathrm{T}} + O(n^{-1}\log^2(n\vee p))}_{\equiv\Delta_{i,J}}. \end{aligned} \tag{55}$$

Over the event B, $|\Delta_{i,J}| \le Cn^{-1/2}\log(n\vee p)$. As a result,

$$\begin{aligned} \|\widehat{\boldsymbol{X}}_{i,J}\|^{2m} &= (\|\boldsymbol{X}_{i,J}\|^2 + \Delta_{i,J})^{2m} \\ &= \sum_{k=0}^{m}\binom{m}{k}\|\boldsymbol{X}_{i,J}\|^{2(m-k)}\Delta_{i,J}^k \\ &= \|\boldsymbol{X}_{i,J}\|^{2m} + m\|\boldsymbol{X}_{i,J}\|^{2m-2}\Delta_{i,k} + O(n^{-1}\log^m(n\vee p)). \end{aligned}$$

Plugging it into (53), we obtain

$$\begin{aligned} \widehat{\theta}_m^{\mathrm{B}} - \widetilde{\theta}_m^{\mathrm{B}} &= \frac{m}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\sum_{i=1}^{n}\frac{1}{c^*_{m,|J|}}\|\boldsymbol{X}_{i,J}\|^{2m-2}\Delta_{i,J} + O(n^{-1}\log^m(n\vee p)) \\ &= \frac{m}{n|\mathcal{A}|}\sum_{i=1}^{n}\sum_{J\in\mathcal{A}}\frac{1}{c^*_{m,|J|}}\|\boldsymbol{X}_{i,J}\|^{2m-2}\boldsymbol{X}_{i,J}^{\mathrm{T}}(\boldsymbol{S}_{2,J} - \mathbb{E}\boldsymbol{S}_{2,J})\boldsymbol{X}_{i,J}^{\mathrm{T}} \\ &\quad - \frac{2m}{n|\mathcal{A}|}\sum_{i=1}^{n}\sum_{J\in\mathcal{A}}\frac{1}{c^*_{m,|J|}}\|\boldsymbol{X}_{i,J}\|^{2m-2}\boldsymbol{S}_{1,J}^{\mathrm{T}}\boldsymbol{X}_{i,J} + O(n^{-1}\log^m(n\vee p)) \\ &= (K_1) + (K_2) + o(n^{-1/2}). \end{aligned} \tag{56}$$

Next, we bound (K_1) and (K_2). Note that $\boldsymbol{S}_{2,J} - \mathbb{E}\boldsymbol{S}_{2,J} = \frac{1}{n}\sum_{k=1}^{n}[\boldsymbol{X}_{k,J}\boldsymbol{X}_{k,J} - \mathbb{E}(\boldsymbol{X}_{k,J}\boldsymbol{X}_{k,J})]$. This allows us to re-write

$$(K_1) = \frac{m}{n^2|\mathcal{A}|}\sum_{i,k=1}^{n}\underbrace{\sum_{J\in\mathcal{A}}\frac{1}{c^*_{m,|J|}}\|\boldsymbol{X}_{i,J}\|^{2m-2}\boldsymbol{X}_{i,J}^{\mathrm{T}}\big[\boldsymbol{X}_{k,J}\boldsymbol{X}_{k,J} - \mathbb{E}(\boldsymbol{X}_{k,J}\boldsymbol{X}_{k,J})\big]\boldsymbol{X}_{i,J}^{\mathrm{T}}}_{\equiv Q_{ik}}.$$

It is not hard to see that $\mathbb{E}|Q_{ii}| \le C|\mathcal{A}|$ and that $\mathbb{E}|Q_{ik}Q_{i'k'}| \le C|\mathcal{A}|^2$ when $\{i,k,i',k'\}$ has at least two distinct values (see the proof of Theorem 2.1 for similar arguments). As a result,

$$\mathbb{E}\Big|\frac{m}{n^2|\mathcal{A}|}\sum_{i=1}^n Q_{ii}\Big| = O(n^{-1}) \qquad \Longrightarrow \qquad \Big|\frac{m}{n^2|\mathcal{A}|}\sum_{i=1}^n Q_{ii}\Big| = o_{\mathbb{P}}(n^{-1/2}).$$

Moreover, noting that $\mathbb{E}Q_{ik} = 0$ for $i \neq k$, we have $\mathbb{E}(Q_{ik}Q_{i'k'}) = 0$ for $\{i,k,i',k'\}$ that are mutually distinct. It follows that

$$\mathbb{E}\Big(\frac{m}{n^2|\mathcal{A}|}\sum_{1\le i\neq k\le n} Q_{ik}\Big)^2 = \frac{m^2}{n^4|\mathcal{A}|^2}\sum_{\substack{(i,k,i',k'):\text{at least}\\ \text{two are equal}}} \mathbb{E}(Q_{ik}Q_{i'k'}) \le \frac{m^2}{n^4|\mathcal{A}|^2}\cdot n^3\cdot C|\mathcal{A}|^2 = O(n^{-1})$$

$$\Longrightarrow \qquad \Big|\frac{m}{n^2|\mathcal{A}|}\sum_{1\le i\neq k\le n} Q_{ik}\Big| = O_{\mathbb{P}}(n^{-1/2}).$$

Combining the above gives

$$(K_1) = O_{\mathbb{P}}(n^{-1/2}). \tag{57}$$

Similarly, since $\boldsymbol{S}_{1,J} = \frac{1}{n}\sum_{k=1}^n \boldsymbol{X}_{i,J}$, we re-write

$$(K_2) = -\frac{2m}{n|\mathcal{A}|}\sum_{i,k=1}^n \underbrace{\sum_{J\in\mathcal{A}} \frac{1}{c^*_{m,|J|}}\|\boldsymbol{X}_{i,J}\|^{2m-2}\boldsymbol{X}_{k,J}^{\mathrm{T}}\boldsymbol{X}_{i,J}}_{R_{ik}}.$$

Then, $\mathbb{E}R_{ik} = 0$ for $i \neq k$, $\mathbb{E}|R_{ii}| \le C|\mathcal{A}|$, and $\mathbb{E}(R_{ik}R_{i'k'}) \le C|\mathcal{A}|^2$ when $\{i,k,i',k'\}$ has at least two distinct values. As a result,

$$\mathbb{E}\Big|\frac{m}{n^2|\mathcal{A}|}\sum_{i=1}^n R_{ii}\Big| = O(n^{-1}) \qquad \Longrightarrow \qquad \Big|\frac{m}{n^2|\mathcal{A}|}\sum_{i=1}^n R_{ii}\Big| = o_{\mathbb{P}}(n^{-1/2})$$

$$\mathbb{E}\Big(\frac{m}{n^2|\mathcal{A}|}\sum_{1\le i\neq k\le n} R_{ik}\Big)^2 = O\Big(\frac{n^3|\mathcal{A}|^2}{n^4|\mathcal{A}|^2}\Big) = O(n^{-1}) \quad \Longrightarrow \quad \Big|\frac{m}{n^2|\mathcal{A}|}\sum_{1\le i\neq k\le n} R_{ik}\Big| = O_{\mathbb{P}}(n^{-1/2}).$$

We immediately have

$$(K_2) = O_{\mathbb{P}}(n^{-1/2}). \tag{58}$$

Plugging (57)-(58) into (56) gives (52). The claim then follows.

A.7 Proof of Theorem 3.2

Similar to the proof of Theorem 2.1, let $\widetilde{\theta}_m^{\mathrm{B}}$ and $\widehat{\theta}_m^{\mathrm{B}}$ denote the BAE with true $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and estimates $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$; here, $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ may not be the sample mean and sample covariance matrix. By Theorem 3.3, $\mathbb{E}[(\widetilde{\theta}_m^{\mathrm{M}} - \theta_m)^2] \le Cn^{-1}$. It follows from the Markov's inequality that, for any $\epsilon > 0$, there is a constant $C_\epsilon > 0$ such that, with probability $1 - \epsilon/2$,

$$|\widetilde{\theta}_m^{\mathrm{M}} - \theta_m| \le C_\epsilon n^{-1/2}.$$

To show the claim, it suffices to show that, there is a constant $C'_\epsilon > 0$ such that with probability $1 - \epsilon/2$,

$$|\widehat{\theta}_m^{\mathrm{B}} - \widetilde{\theta}_m^{\mathrm{B}}| \le C_\epsilon \max\{\alpha_n, \beta_n\}. \tag{59}$$

We now show (59). Let $\boldsymbol{X}_{i,J} = \boldsymbol{\Sigma}_{J,J}^{-1/2}(\boldsymbol{Y}_{i,J} - \boldsymbol{\mu}_J)$ and $\widehat{\boldsymbol{X}}_{i,J} = \widehat{\boldsymbol{\Sigma}}_{J,J}^{-1/2}(\boldsymbol{Y}_{i,J} - \widehat{\boldsymbol{\mu}}_J)$. Then,

$$\widehat{\theta}_m^{\mathrm{B}} - \widetilde{\theta}_m^{\mathrm{B}} = \frac{1}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\sum_{i=1}^{n}\frac{\|\widehat{\boldsymbol{X}}_{i,J}\|^{2m} - \|\boldsymbol{X}_{i,J}\|^{2m}}{c^*_{m,|J|}}.$$

By direct calculations,

$$\begin{aligned}
\Delta_{i,J} &\equiv \|\widehat{\boldsymbol{X}}_{i,J}\|^2 - \|\boldsymbol{X}_{i,J}\|^2 \\
&= (\boldsymbol{Y}_{i,J} - \boldsymbol{\mu}_J)^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}_{J,J}^{-1} - \boldsymbol{\Sigma}_{J,J}^{-1})(\boldsymbol{Y}_{i,J} - \boldsymbol{\mu}_J) + 2(\boldsymbol{\mu}_J - \widehat{\boldsymbol{\mu}}_J)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{J,J}^{-1}(\boldsymbol{Y}_{i,J} - \boldsymbol{\mu}_J) \\
&\quad + (\boldsymbol{\mu}_J - \widehat{\boldsymbol{\mu}}_J)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{J,J}^{-1}(\boldsymbol{\mu}_J - \widehat{\boldsymbol{\mu}}_J) \\
&= \boldsymbol{X}_{i,J}^{\mathrm{T}}\big(\boldsymbol{\Sigma}_{J,J}^{1/2}\widehat{\boldsymbol{\Sigma}}_{J,J}^{-1}\boldsymbol{\Sigma}_{J,J}^{1/2} - \mathbf{I}_{|J|}\big)\boldsymbol{X}_{i,J} - 2\big[\boldsymbol{\Sigma}_{JJ}^{-1/2}(\widehat{\boldsymbol{\mu}}_J - \boldsymbol{\mu}_J)\big]^{\mathrm{T}}\big(\boldsymbol{\Sigma}_{J,J}^{1/2}\widehat{\boldsymbol{\Sigma}}_{J,J}^{-1}\boldsymbol{\Sigma}_{J,J}^{1/2}\big)\boldsymbol{X}_{i,J} \\
&\quad + \big[\boldsymbol{\Sigma}_{JJ}^{-1/2}(\widehat{\boldsymbol{\mu}}_J - \boldsymbol{\mu}_J)\big]^{\mathrm{T}}\big(\boldsymbol{\Sigma}_{J,J}^{1/2}\widehat{\boldsymbol{\Sigma}}_{J,J}^{-1}\boldsymbol{\Sigma}_{J,J}^{1/2}\big)\big[\boldsymbol{\Sigma}_{JJ}^{-1/2}(\widehat{\boldsymbol{\mu}}_J - \boldsymbol{\mu}_J)\big]. \qquad (60)
\end{aligned}$$

As a result,

$$\begin{aligned}
&\widehat{\theta}_m^{\mathrm{B}} - \widetilde{\theta}_m^{\mathrm{B}} \\
&= \frac{1}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\frac{1}{c^*_{m,|J|}}\sum_{i=1}^{n}\Big[\sum_{k=1}^{m}\binom{m}{k}\|\boldsymbol{X}_{i,J}\|^{2(m-k)}\Delta_{i,J}^k\Big] \\
&= \frac{m}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\frac{1}{c^*_{m,|J|}}\sum_{i=1}^{n}\|\boldsymbol{X}_{i,J}\|^{2(m-1)}\Delta_{i,J} + rem \\
&= \frac{m}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\frac{1}{c^*_{m,|J|}}\sum_{i=1}^{n}\|\boldsymbol{X}_{i,J}\|^{2m-2}\boldsymbol{X}_{i,J}^{\mathrm{T}}\big(\boldsymbol{\Sigma}_{J,J}^{1/2}\widehat{\boldsymbol{\Sigma}}_{J,J}^{-1}\boldsymbol{\Sigma}_{J,J}^{1/2} - \mathbf{I}_{|J|}\big)\boldsymbol{X}_{i,J} + rem \\
&\quad - \frac{2m}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\frac{1}{c^*_{m,|J|}}\sum_{i=1}^{n}\|\boldsymbol{X}_{i,J}\|^{2m-2}[\boldsymbol{\Sigma}_{JJ}^{-1/2}(\widehat{\boldsymbol{\mu}}_J - \boldsymbol{\mu}_J)]^{\mathrm{T}}(\boldsymbol{\Sigma}_{J,J}^{1/2}\widehat{\boldsymbol{\Sigma}}_{J,J}^{-1}\boldsymbol{\Sigma}_{J,J}^{1/2})\boldsymbol{X}_{i,J}. \qquad (61)
\end{aligned}$$

Introduce

$$\widetilde{\boldsymbol{S}}_{2,J}^{(m)} = \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{X}_{i,J}\|^{2m-2}\boldsymbol{X}_{i,J}\boldsymbol{X}_{i,J}^{\mathrm{T}}, \qquad \widetilde{\boldsymbol{S}}_{1,J}^{(m)} = \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{X}_{i,J}\|^{2m-2}\boldsymbol{X}_{i,J}.$$

Then, (61) can be rewritten as

$$\begin{aligned}
\widehat{\theta}_m^{\mathrm{B}} - \widetilde{\theta}_m^{\mathrm{B}} &= \frac{m}{|\mathcal{A}|}\sum_{J\in\mathcal{A}}\frac{1}{c^*_{m,|J|}}\operatorname{tr}\Big[\big(\boldsymbol{\Sigma}_{J,J}^{1/2}\widehat{\boldsymbol{\Sigma}}_{J,J}^{-1}\boldsymbol{\Sigma}_{J,J}^{1/2} - \mathbf{I}_{|J|}\big)\widetilde{\boldsymbol{S}}_{2,J}^{(m)}\Big] \\
&\quad - \frac{2m}{|\mathcal{A}|}\sum_{J\in\mathcal{A}}\frac{1}{c^*_{m,|J|}}[\boldsymbol{\Sigma}_{JJ}^{-1/2}(\widehat{\boldsymbol{\mu}}_J - \boldsymbol{\mu}_J)]^{\mathrm{T}}\widetilde{\boldsymbol{S}}_{1,J}^{(m)} + rem. \qquad (62)
\end{aligned}$$

First, we study the main terms in (62). Note that $\widetilde{\boldsymbol{S}}_{2,J}^{(m)}$ is the sample covariance matrix of $\{\|\boldsymbol{X}_{i,J}\|^{m-1}\boldsymbol{X}_{i,J} : 1 \le i \le n\}$, and $\widetilde{\boldsymbol{S}}_{1,J}^{(m)}$ is the sample mean of $\{\|\boldsymbol{X}_{i,J}\|^{2m-2}\boldsymbol{X}_{i,J} : 1 \le i \le n\}$. Using similar calculations as in the proof of Theorem 3.3, we can prove that

$$\Big\|\frac{1}{|\mathcal{A}|}\sum_{J\in\mathcal{A}}\mathbb{E}\widetilde{\boldsymbol{S}}_{2,J}^{(m)}\Big\| \le C, \qquad \Big\|\frac{1}{|\mathcal{A}|}\sum_{J\in\mathcal{A}}\mathbb{E}\widetilde{\boldsymbol{S}}_{1,J}^{(m)}\Big\| \le C.$$

Combining it with the Markov inequality, for any $\epsilon > 0$, there is $C > 0$ such that, with probability $1 - \epsilon/4$, $\big\|\frac{1}{|\mathcal{A}|}\sum_{J\in\mathcal{A}} \widetilde{\boldsymbol{S}}^{(m)}_{2,J}\big\| \le C$ and $\big\|\frac{1}{|\mathcal{A}|}\sum_{J\in\mathcal{A}} \widetilde{\boldsymbol{S}}^{(m)}_{1,J}\big\| \le C$. On this event, the sum of the first two terms in (62) is bounded in absolute value by

$$\begin{aligned}
&C\Big\|\frac{1}{|\mathcal{A}|}\sum_{J\in\mathcal{A}} \widetilde{\boldsymbol{S}}^{(m)}_{2,J}\Big\| \cdot \max_{J\in\mathcal{A}}\big\|\boldsymbol{\Sigma}^{1/2}_{J,J}\widehat{\boldsymbol{\Sigma}}^{-1}_{J,J}\boldsymbol{\Sigma}^{1/2}_{J,J} - \mathbf{I}_{|J|}\big\| + C\Big\|\frac{1}{|\mathcal{A}|}\sum_{J\in\mathcal{A}} \widetilde{\boldsymbol{S}}^{(m)}_{1,J}\Big\| \cdot \max_{J\in\mathcal{A}}\big\|\boldsymbol{\Sigma}^{-1/2}_{JJ}(\widehat{\boldsymbol{\mu}}_J - \boldsymbol{\mu}_J)\big\| \\
&\le C\max_{J\in\mathcal{A}}\big\|\boldsymbol{\Sigma}^{1/2}_{J,J}\widehat{\boldsymbol{\Sigma}}^{-1}_{J,J}\boldsymbol{\Sigma}^{1/2}_{J,J} - \mathbf{I}_{|J|}\big\| + C\max_{J\in\mathcal{A}}\big\|\boldsymbol{\Sigma}^{-1/2}_{JJ}(\widehat{\boldsymbol{\mu}}_J - \boldsymbol{\mu}_J)\big\| \le C\max\{\alpha_n, \beta_n\}. \qquad (63)
\end{aligned}$$

Next, we study the remainder terms in (62). By (60) and our assumption on $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, we have

$$\|\Delta_{i,J}\| \le C\beta_n\|\boldsymbol{X}_{i,J}\|^2 + C\alpha_n\|\boldsymbol{X}_{i,J}\|.$$

It follows that $\|\Delta_{i,J}\|^k \le C\beta_n^k\|\boldsymbol{X}_{i,J}\|^{2k} + C\alpha_n^k\|\boldsymbol{X}_{i,J}\|^k$. Then,

$$\begin{aligned}
|rem| &\le C\sum_{k=2}^m \frac{1}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\sum_{i=1}^n \beta_n^k\|\boldsymbol{X}_{i,J}\|^{2n} + C\sum_{k=2}^m \frac{1}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\sum_{i=1}^n \alpha_n^k\|\boldsymbol{X}_{i,J}\|^{2n-k} \\
&\le C\sum_{k=2}^m \beta_n^k\Big(\frac{1}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\sum_{i=1}^n \|\boldsymbol{X}_{i,J}\|^{2n}\Big) + C\sum_{k=2}^m \alpha_n^k\Big(\frac{1}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\sum_{i=1}^n \|\boldsymbol{X}_{i,J}\|^{2n-k}\Big).
\end{aligned}$$

Using similar calculations as in the proof of Theorem 3.3, we can prove that $\frac{1}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\sum_{i=1}^n \mathbb{E}\|\boldsymbol{X}_{i,J}\|^k \le C$, for all $1 \le k \le 4m$. It follows from the Markov inequality that, for a constant $C > 0$, with probability $1-\epsilon/4$, $\frac{1}{n|\mathcal{A}|}\sum_{J\in\mathcal{A}}\sum_{i=1}^n \|\boldsymbol{X}_{i,J}\|^k \le C$, for all $1 \le k \le 2m$. On this event,

$$|rem| \le C(\alpha_n^2 + \beta_n^2). \qquad (64)$$

Combining (63) and (64) gives $|\widehat{\theta}^{\mathrm{B}}_m - \widetilde{\theta}^{\mathrm{B}}_m| \le C\max\{\alpha_n, \beta_n\}$. This proves (59), and the claim follows immediately.

A.8 Proof of Theorem 3.3

Fix a collection $\mathcal{A}$ of blocks. Write for short $\widehat{\theta}^B_m = \widehat{\theta}^B_m(\boldsymbol{\mu}, \mathrm{diag}_{\mathcal{A}}(\boldsymbol{\Sigma}))$. For preparation, first, we verify that $\widehat{\theta}^B_m$ is an unbiased estimator. For any $J \in \mathcal{A}$, by (12) and the fact that $\|\boldsymbol{U}_{|J|}\| = 1$, we have

$$\big[(\boldsymbol{Y}_J - \boldsymbol{\mu}_J)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}_{J,J}(\boldsymbol{Y}_J - \boldsymbol{\mu}_J)\big]^m = \|\boldsymbol{\Sigma}^{-1/2}_{JJ}(\boldsymbol{Y}_J - \boldsymbol{\mu}_J)\|^{2m} = \|\xi B^{1/2}\boldsymbol{U}_{|J|}\|^{2m} = \xi^{2m}B^m.$$

As a result,

$$\mathbb{E}\big[(\boldsymbol{Y}_J - \boldsymbol{\mu}_J)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}_{J,J}(\boldsymbol{Y}_J - \boldsymbol{\mu}_J)\big]^m = (\mathbb{E}\xi^{2m})(\mathbb{E}B^m) = \theta_m \cdot c^*_{m,|J|}. \qquad (65)$$

In particular, it implies that

$$\theta_m = \frac{1}{|\mathcal{A}|n}\sum_{J\in\mathcal{A}}\bigg[\frac{1}{c^*_{m,|J|}}\sum_{i=1}^n \mathbb{E}\big\{(\boldsymbol{Y}_{i,J} - \boldsymbol{\mu}_J)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}_{JJ}(\boldsymbol{Y}_{i,J} - \boldsymbol{\mu}_J)\big\}^m\bigg].$$

Therefore, $\widehat{\theta}^B_m$ is unbiased. Additionally, we have

$$\frac{\mathrm{var}(\widehat{\theta}^B_m)}{\theta_m^2} = \frac{1}{n|\mathcal{A}|^2}\,\mathrm{var}\left(\sum_{J\in\mathcal{A}}\frac{\big[(\boldsymbol{Y}_J - \boldsymbol{\mu}_J)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}_{J,J}(\boldsymbol{Y}_J - \boldsymbol{\mu}_J)\big]^m}{\theta_m \cdot c^*_{m,|J|}}\right). \qquad (66)$$

Second, we introduce an alternative expression of $\theta_m \cdot c^*_{m,|J|}$. Consider the special case $\xi \sim \chi^2_p$. Since $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in this case, we then have $\boldsymbol{Y}_J \sim N(\boldsymbol{\mu}_J, \boldsymbol{\Sigma}_{J,J})$ and $(\boldsymbol{Y}_J - \boldsymbol{\mu}_J)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}_{J,J} (\boldsymbol{Y}_J - \boldsymbol{\mu}_J) \sim \chi^2_{|J|}$. Hence, in (65), the left hand side equals to $\mathbb{E}\chi^{2m}_{|J|}$. At the same time, the right hand side is equal to $\theta_m \cdot c^*_{m,|J|} = p^{-m}\mathbb{E}\xi^{2m} \cdot c^*_{m,|J|} = p^{-m}\mathbb{E}\chi^{2m}_p \cdot c^*_{m,|J|}$. Equating the left/right hand sides gives

$$c^*_{m,|J|} = \frac{p^m \mathbb{E}\chi^{2m}_{|J|}}{\mathbb{E}\chi^{2m}_p}.$$

We combine it with the definition of $\theta_m = p^{-m}\mathbb{E}\xi^{2m}$ and $r_m = \mathbb{E}\xi^{2m}/\mathbb{E}\chi^{2m}_p$. It implies that

$$\theta_m \cdot c^*_{m,|J|} = \mathbb{E}\xi^{2m} \cdot \frac{\mathbb{E}\chi^{2m}_{|J|}}{\mathbb{E}\chi^{2m}_p} = r_m \cdot \mathbb{E}\chi^{2m}_{|J|}. \tag{67}$$

We now show the claim. For $J \in \mathcal{A}$, let $W_J = [(\boldsymbol{Y}_J - \boldsymbol{\mu}_J)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}_{J,J} (\boldsymbol{Y}_J - \boldsymbol{\mu}_J)]^m$. By (66)-(67),

$$\begin{aligned}
\frac{\mathrm{var}(\widehat{\theta}^B_m)}{\theta^2_m} &= \frac{1}{n|\mathcal{A}|^2} \mathrm{var}\left(\sum_{J\in\mathcal{A}} \frac{W_J}{r_m \cdot \mathbb{E}\chi^{2m}_{|J|}}\right) \\
&= \frac{1}{n|\mathcal{A}|^2 r^2_m} \sum_{J\in\mathcal{A}} \frac{\mathrm{var}(W_J)}{(\mathbb{E}\chi^{2m}_{|J|})^2} + \frac{1}{n|\mathcal{A}|^2 r^2_m} \sum_{\substack{I,J\in\mathcal{A} \\ I\neq J}} \frac{\mathrm{cov}(W_I, W_J)}{(\mathbb{E}\chi^{2m}_{|I|})(\mathbb{E}\chi^{2m}_{|J|})} \\
&\equiv (I) + (II).
\end{aligned} \tag{68}$$

Consider (I). Combining (65) and (67), we have

$$\mathbb{E}W_J = r_m \cdot \mathbb{E}\chi^{2m}_{|J|}, \qquad \mathbb{E}W^2_J = r_{2m} \cdot \mathbb{E}\chi^{4m}_{|J|}. \tag{69}$$

Hence,

$$\begin{aligned}
(I) &= \frac{1}{n|\mathcal{A}|^2 r^2_m} \sum_{J\in\mathcal{A}} \frac{r_{2m}\mathbb{E}\chi^{4m}_{|J|} - r^2_m(\mathbb{E}\chi^{2m}_{|J|})^2}{(\mathbb{E}\chi^{2m}_{|J|})^2} \\
&= \frac{1}{n|\mathcal{A}|^2 r^2_m} \sum_{J\in\mathcal{A}} \frac{r_{2m}\,\mathrm{var}(\chi^{2m}_{|J|}) + (r_{2m} - r^2_m)(\mathbb{E}\chi^{2m}_{|J|})^2}{(\mathbb{E}\chi^{2m}_{|J|})^2} \\
&= \frac{1}{n|\mathcal{A}|^2} \sum_{J\in\mathcal{A}} \left[\frac{r_{2m}}{r^2_m} \frac{\mathrm{var}(\chi^{2m}_{|J|})}{(\mathbb{E}\chi^{2m}_{|J|})^2} + \frac{(r_{2m} - r^2_m)}{r^2_m}\right] \\
&= \frac{1}{np} \cdot \frac{r_{2m}}{r^2_m} \cdot \underbrace{\frac{p}{|\mathcal{A}|^2} \sum_{J\in\mathcal{A}} \frac{h_m(|J|)}{|J|}}_{\bar{h}_m(\mathcal{A})} + \frac{1}{n|\mathcal{A}|} \cdot \frac{(r_{2m} - r^2_m)}{r^2_m},
\end{aligned} \tag{70}$$

where the last two lines are from Definition 3.1.

Consider (II). Fix I and J. Note that

$$\mathrm{cov}(W_I, W_J) = \mathbb{E}(W_I W_J) - (\mathbb{E}W_I)(\mathbb{E}W_J).$$

We have had an expression of $\mathbb{E}W_I$ as in (69). We still need to get an expression of $\mathbb{E}(W_I W_J)$. For the set $I \cup J$, we apply (12) and find that

$$\begin{pmatrix} \boldsymbol{Y}_I \\ \boldsymbol{Y}_J \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_I \\ \boldsymbol{\mu}_J \end{pmatrix} + \xi \cdot B^{1/2} \cdot \boldsymbol{\Sigma}^{1/2}_{I\cup J, I\cup J} \boldsymbol{U}_{|I|+|J|},$$

where B is a Beta distribution with parameters $\frac{|I|+|J|}{2}$ and $\frac{p-(|I|+|J|)}{2}$. Let $\widetilde{\boldsymbol{U}}_I$ and $\widetilde{\boldsymbol{U}}_J$ be the vectors formed by the first $|I|$ coordinates and the last $|J|$ coordinates of $\boldsymbol{\Sigma}_{I\cup J,I\cup J}^{1/2}\boldsymbol{U}_{|I|+|J|}$, respectively. We then have $W_I = \xi^{2m}B^m\|\boldsymbol{\Sigma}_{II}^{-1/2}\widetilde{\boldsymbol{U}}_I\|^{2m}$ and $W_J = \xi^{2m}B^m\|\boldsymbol{\Sigma}_{JJ}^{-1/2}\widetilde{\boldsymbol{U}}_J\|^{2m}$. As a result,

$$\mathbb{E}(W_IW_J) = \mathbb{E}\xi^{4m}\cdot\mathbb{E}B^{2m}\cdot\mathbb{E}\big(\|\boldsymbol{\Sigma}_{II}^{-1/2}\widetilde{\boldsymbol{U}}_I\|^{2m}\|\boldsymbol{\Sigma}_{JJ}^{-1/2}\widetilde{\boldsymbol{U}}_J\|^{2m}\big). \tag{71}$$

We then use the cross-moments of multivariate normal distributions to get the last term above. Let $\xi_0^2\sim\chi_p^2$ be a random variable independent of B and $\boldsymbol{U}_{|I|+|J|}$. The random vector

$$\begin{pmatrix}\boldsymbol{Z}_I\\ \boldsymbol{Z}_J\end{pmatrix} \equiv \xi_0\cdot B^{1/2}\cdot\begin{pmatrix}\widetilde{\boldsymbol{U}}_I\\ \widetilde{\boldsymbol{U}}_J\end{pmatrix} \sim N\big(\boldsymbol{0},\ \boldsymbol{\Sigma}_{I\cup J,I\cup J}\big).$$

It follows that

$$\mathbb{E}\big(\|\boldsymbol{\Sigma}_{II}^{-1/2}\boldsymbol{Z}_I\|^{2m}\|\boldsymbol{\Sigma}_{JJ}^{-1/2}\boldsymbol{Z}_J\|^{2m}\big) = \mathbb{E}\chi_p^{4m}\cdot\mathbb{E}B^{2m}\cdot\mathbb{E}\big(\|\boldsymbol{\Sigma}_{II}^{-1/2}\widetilde{\boldsymbol{U}}_I\|^{2m}\|\boldsymbol{\Sigma}_{JJ}^{-1/2}\widetilde{\boldsymbol{U}}_J\|^{2m}\big). \tag{72}$$

Write $\widetilde{\boldsymbol{Z}}_1 = \boldsymbol{\Sigma}_{II}^{-1/2}\boldsymbol{Z}_I$ and $\widetilde{\boldsymbol{Z}}_2 = \boldsymbol{\Sigma}_{JJ}^{-1/2}\boldsymbol{Z}_J$. Note that

$$\begin{pmatrix}\widetilde{\boldsymbol{Z}}_1\\ \widetilde{\boldsymbol{Z}}_2\end{pmatrix} \sim N\left(\boldsymbol{0},\ \begin{bmatrix}\mathbf{I}_{|I|} & \boldsymbol{\Gamma}\\ \boldsymbol{\Gamma}^{\mathrm{T}} & \mathbf{I}_{|J|}\end{bmatrix}\right), \qquad \text{where}\quad \boldsymbol{\Gamma} = \boldsymbol{\Sigma}_{II}^{-1/2}\boldsymbol{\Sigma}_{IJ}\boldsymbol{\Sigma}_{JJ}^{-1/2}. \tag{73}$$

Combining (71) and (72) gives

$$\mathbb{E}(W_IW_J) = \mathbb{E}(\|\widetilde{\boldsymbol{Z}}_1\|^{2m}\|\widetilde{\boldsymbol{Z}}_2\|^{2m})\cdot\frac{\mathbb{E}\xi^{4m}}{\mathbb{E}\chi_p^{4m}} = \mathbb{E}(\|\widetilde{\boldsymbol{Z}}_1\|^{2m}\|\widetilde{\boldsymbol{Z}}_2\|^{2m})\cdot r_{2m}. \tag{74}$$

We now combine (69) and (74) and note that $\|\widetilde{\boldsymbol{Z}}_1\|^2\sim\chi_{|I|}^2$ and $\|\widetilde{\boldsymbol{Z}}_2\|^2\sim\chi_{|J|}^2$. It yields

$$\begin{aligned}
\frac{\mathrm{cov}(W_I,W_J)}{(\mathbb{E}\chi_{|I|}^{2m})(\mathbb{E}\chi_{|J|}^{2m})} &= \frac{r_{2m}\mathbb{E}(\|\widetilde{\boldsymbol{Z}}_1\|^{2m}\|\widetilde{\boldsymbol{Z}}_2\|^{2m}) - r_m^2(\mathbb{E}\chi_{|I|}^{2m})(\mathbb{E}\chi_{|J|}^{2m})}{(\mathbb{E}\chi_{|I|}^{2m})(\mathbb{E}\chi_{|J|}^{2m})}\\
&= \frac{r_{2m}\mathbb{E}(\|\widetilde{\boldsymbol{Z}}_1\|^{2m}\|\widetilde{\boldsymbol{Z}}_2\|^{2m}) - r_m^2(\mathbb{E}\|\widetilde{\boldsymbol{Z}}_1\|^{2m})(\mathbb{E}\|\widetilde{\boldsymbol{Z}}_1\|^{2m})}{(\mathbb{E}\chi_{|I|}^{2m})(\mathbb{E}\chi_{|J|}^{2m})}\\
&= \frac{r_{2m}\,\mathrm{cov}(\|\widetilde{\boldsymbol{Z}}_1\|^{2m},\|\widetilde{\boldsymbol{Z}}_2\|^{2m}) + (r_{2m}-r_m^2)(\mathbb{E}\|\widetilde{\boldsymbol{Z}}_1\|^{2m})(\mathbb{E}\|\widetilde{\boldsymbol{Z}}_1\|^{2m})}{(\mathbb{E}\chi_{|I|}^{2m})(\mathbb{E}\chi_{|J|}^{2m})}\\
&= r_{2m}\frac{\mathrm{cov}(\|\widetilde{\boldsymbol{Z}}_1\|^{2m},\|\widetilde{\boldsymbol{Z}}_2\|^{2m})}{(\mathbb{E}\|\widetilde{\boldsymbol{Z}}_1\|^{2m})(\mathbb{E}\|\widetilde{\boldsymbol{Z}}_1\|^{2m})} + (r_{2m}-r_m^2).
\end{aligned}$$

As a result,

$$\begin{aligned}
(II) &= \frac{1}{n|\mathcal{A}|^2r_m^2}\sum_{\substack{I,J\in\mathcal{A}\\ I\neq J}}\left[r_{2m}\frac{\mathrm{cov}(\|\widetilde{\boldsymbol{Z}}_1\|^{2m},\|\widetilde{\boldsymbol{Z}}_2\|^{2m})}{(\mathbb{E}\|\widetilde{\boldsymbol{Z}}_1\|^{2m})(\mathbb{E}\|\widetilde{\boldsymbol{Z}}_1\|^{2m})} + (r_{2m}-r_m^2)\right]\\
&= \frac{1}{n}\cdot\frac{r_{2m}}{r_m^2}\cdot\frac{1}{|\mathcal{A}|^2}\sum_{\substack{I,J\in\mathcal{A}\\ I\neq J}}\frac{\mathrm{cov}(\|\widetilde{\boldsymbol{Z}}_1\|^{2m},\|\widetilde{\boldsymbol{Z}}_2\|^{2m})}{(\mathbb{E}\|\widetilde{\boldsymbol{Z}}_1\|^{2m})(\mathbb{E}\|\widetilde{\boldsymbol{Z}}_1\|^{2m})} + \frac{1}{n}\cdot\frac{(r_{2m}-r_m^2)}{r_m^2}\Big(1-\frac{1}{|\mathcal{A}|}\Big).
\end{aligned} \tag{75}$$

We now plug (70) and (75) into (68). It gives

$$\frac{\mathrm{var}(\widehat{\theta}_m^B)}{\theta_m^2} \le \frac{1}{n}\cdot\frac{(r_{2m}-r_m^2)}{r_m^2} + \frac{1}{np}\cdot\frac{r_{2m}}{r_m^2}\bar{h}_m(\mathcal{A}) + \frac{1}{n}\cdot\frac{r_{2m}}{r_m^2}\cdot\frac{1}{|\mathcal{A}|^2}\sum_{\substack{I,J\in\mathcal{A}\\ I\neq J}}\frac{\mathrm{cov}(\|\widetilde{\boldsymbol{Z}}_1\|^{2m},\|\widetilde{\boldsymbol{Z}}_2\|^{2m})}{(\mathbb{E}\|\widetilde{\boldsymbol{Z}}_1\|^{2m})(\mathbb{E}\|\widetilde{\boldsymbol{Z}}_1\|^{2m})}. \tag{76}$$

What remains is to bound the last term. Since the random vectors $\widetilde{\boldsymbol{Z}}_1$ and $\widetilde{\boldsymbol{Z}}_2$ jointly follow a multivariate normal distribution as dictated in (73), we can apply the following lemma:

Lemma A.2. *Let $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ be two random vectors such that*

$$\begin{pmatrix}\boldsymbol{Z}_1\\ \boldsymbol{Z}_2\end{pmatrix} \sim N\left(\mathbf{0},\ \begin{bmatrix}\mathbf{I}_{k_1} & \boldsymbol{\Gamma}\\ \boldsymbol{\Gamma}' & \mathbf{I}_{k_2}\end{bmatrix}\right).$$

Then, for a constant $\widetilde{C}_m > 0$ that only depends on m but is independent of (k_1, k_2),

$$0 \le \frac{\mathrm{cov}(\|\boldsymbol{Z}_1\|^{2m},\|\boldsymbol{Z}_2\|^{2m})}{(\mathbb{E}\|\boldsymbol{Z}_1\|^{2m})(\mathbb{E}\|\boldsymbol{Z}_1\|^{2m})} \le \widetilde{C}_m\|\boldsymbol{\Gamma}\|^2.$$

We combine Lemma A.2 with (73) and them plug it into (76). It follows that

$$\frac{\mathrm{var}(\widehat{\theta}_m^B)}{\theta_m^2} \le \frac{1}{n}\,\frac{(r_{2m}-r_m^2)}{r_m^2} + \frac{1}{np}\,\frac{r_{2m}}{r_m^2}\bar{h}_m(\mathcal{A}) + \frac{1}{n}\frac{r_{2m}}{r_m^2}\,\frac{\widetilde{C}_m}{|\mathcal{A}|^2}\sum_{\substack{I,J\in\mathcal{A}\\ I\neq J}}\|\boldsymbol{\Sigma}_{II}^{-1/2}\boldsymbol{\Sigma}_{IJ}\boldsymbol{\Sigma}_{JJ}^{-1/2}\|^2.$$

This proves the claim.

B Supplementary proofs

B.1 Proof of Lemma A.1

Let $\theta = \arcsin\big(\mathrm{sign}(\rho)\cdot\sqrt{|\rho|}\big) \in [-\frac{\pi}{2},\frac{\pi}{2}]$. We then have $\sin\theta = \mathrm{sign}(\rho)\cdot\sqrt{|\rho|}$ and $\cos\theta = \sqrt{1-|\rho|}$. Let U_1, U_2, V be iid $N(0,1)$ random variables. It is easy to see that

$$(Z_1, Z_2) \overset{(d)}{=} \Big((\cos\theta)U_1 + (\sin\theta)V,\ \ (\cos\theta)U_2 + (\sin\theta)V\Big).$$

For notation simplicity, we omit the superscript (d) in all equations. It follows that

$$Z_1^{2m} = \sum_{k_1=0}^{2m}\binom{2m}{k_1}(\cos\theta)^{2m-k_1}(\sin\theta)^{k_1}U_1^{2m-k_1}V^{k_1},$$
$$Z_2^{2m} = \sum_{k_2=0}^{2m}\binom{2m}{k_2}(\cos\theta)^{2m-k_1}(\sin\theta)^{k_2}U_2^{2m-k_1}V^{k_2}.$$

Then,

$$\mathrm{cov}(Z_1^{2m}, Z_2^{2m}) = \sum_{k_1,k_2=0}^{2m}\binom{2m}{k_1}\binom{2m}{k_2}(\cos\theta)^{4m-k_1-k_2}(\sin\theta)^{k_1+k_2}\big[\mathrm{cov}(U_1^{2m-k_1}V^{k_1},\ U_2^{2m-k_2}V^{k_2})\big].$$

Note that for random variables (X, Y, W_1, W_2), when X, Y and (W_1, W_2) are mutually independent, $\mathrm{cov}(XW_1, YW_2) = \mathbb{E}X \cdot \mathbb{E}Y \cdot \mathrm{cov}(W_1, W_2)$. Plugging it into the above expression, we obtain

$$\begin{aligned}
&\mathrm{cov}(Z_1^{2m}, Z_2^{2m}) \\
=& \sum_{\substack{2\le k_1,k_2\le 2m \\ k_1,k_2 \text{ even}}} \binom{2m}{k_1}\binom{2m}{k_2} (\cos\theta)^{4m-k_1-k_2}(\sin\theta)^{k_1+k_2}(\mathbb{E}U_1^{2m-k_1})(\mathbb{E}U_2^{2m-k_2})\,\mathrm{cov}(V^{k_1}, V^{k_2}) \\
=& \sum_{s=2}^{m} (\cos\theta)^{2m-2s}(\sin\theta)^{2s} \sum_{\substack{1\le k_1,k_2\le m \\ k_1+k_2=s}} \binom{2m}{2k_1}\binom{2m}{2k_2} \big[\mathbb{E}U_1^{2(m-k_1)}\big]\big[\mathbb{E}U_2^{2(m-k_2)}\big]\big(\mathbb{E}V^{2s} - \mathbb{E}V^{2k_1}\mathbb{E}V^{2k_2}\big).
\end{aligned}$$

Using our previous notations, η_m is the $2m$-th moment of $N(0,1)$. By elementary statistics, $\eta_m = (2m-1)!! = \prod_{j=0}^{m-1}(1+2j)$. Using this formula, we can prove $\mathbb{E}V^{2s} - \mathbb{E}V^{2k_1}\mathbb{E}V^{2k_2} \ge 0$. Hence,

$$\mathrm{cov}(Z_1^{2m}, Z_2^{2m}) \ge 0.$$

At the same time, we note that $\cos^2\theta = 1 - |\rho|$ and $\sin^2\theta = |\rho|$. It follows that

$$\begin{aligned}
\mathrm{cov}(Z_1^{2m}, Z_2^{2m}) &\le \sum_{s=2}^{m} (1-|\rho|)^{m-s}|\rho|^s \cdot \underbrace{\sum_{\substack{1\le k_1,k_2\le m \\ k_1+k_2=s}} \binom{2m}{2k_1}\binom{2m}{2k_2} \cdot \eta_{m-k_1}\eta_{m-k_2}\eta_s}_{B_m(s)} \\
&\le \big[\max_s B_m(s)\big] \cdot \sum_{s=2}^{m} (1-|\rho|)^{m-s}|\rho|^s \le \big[\max_s B_m(s)\big] \cdot |\rho|^2.
\end{aligned}$$

The claim then follows.

B.2 Proof of Lemma A.2

Suppose the rank of $\mathbf{\Gamma}$ is $k \le \min\{k_1, k_2\}$. Let $\mathbf{\Gamma} = \mathbf{H}_1\mathbf{\Lambda}\mathbf{H}_2^{\mathrm{T}}$ be the singular value decomposition of $\mathbf{\Gamma}$. We note that all singular values have an absolute value no larger than 1. For $\ell = 1, 2$, let $\widetilde{\mathbf{H}}_\ell \in \mathbb{R}^{k_\ell, k_\ell - k}$ be such that $[\mathbf{H}_\ell, \widetilde{\mathbf{H}}_\ell]$ form an orthogonal basis of $\mathbb{R}^{k_\ell}$. Define

$$\mathbf{A}_\ell = \big[\mathbf{H}_\ell(\mathbf{I} - \mathbf{\Lambda})^{1/2},\ \ \widetilde{\mathbf{H}}_\ell\big], \qquad \ell = 1, 2.$$

It is easy to see that $\mathbf{A}_\ell\mathbf{A}_\ell' = \mathbf{I} - \mathbf{H}_\ell\mathbf{\Lambda}\mathbf{H}_\ell'$. Let $\boldsymbol{X}_1 \sim N(\mathbf{0}, \mathbf{I}_{k_1})$, $\boldsymbol{X}_2 \sim N(\mathbf{0}, \mathbf{I}_{k_2})$, and $\boldsymbol{W} \sim N(\mathbf{0}, \mathbf{I}_k)$ be mutually independent random variables. We claim that

$$\begin{pmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \end{pmatrix} \overset{(d)}{=} \begin{pmatrix} \mathbf{A}_1\boldsymbol{X}_1 + \mathbf{H}_1\mathbf{\Lambda}^{1/2}\boldsymbol{W} \\ \mathbf{A}_2\boldsymbol{X}_2 + \mathbf{H}_2\mathbf{\Lambda}^{1/2}\boldsymbol{W} \end{pmatrix}.$$

This can be verified by computing the covariance matrix of the right hand side. We shall omit the superscript (d) in all equations for notation simplicity. Write $\boldsymbol{X}_\ell = (\boldsymbol{X}_{\ell 1}^{\mathrm{T}}, \boldsymbol{X}_{\ell 2}^{\mathrm{T}})^{\mathrm{T}}$, corresponding to the first k_ℓ and the last $(k_\ell - k)$ coordinates, respectively, $\ell = 1, 2$. It follows that

$$\|\boldsymbol{Z}_\ell\|^2 = \|\mathbf{A}_\ell\boldsymbol{X}_\ell + \mathbf{H}_\ell\mathbf{\Lambda}^{1/2}\boldsymbol{W}\|^2$$

$$
\begin{aligned}
&= \|\mathbf{H}_\ell(\mathbf{I}-\mathbf{\Lambda})^{1/2}\boldsymbol{X}_{\ell 1} + \widetilde{\mathbf{H}}_\ell \boldsymbol{X}_{\ell 2} + \mathbf{H}_\ell \mathbf{\Lambda}^{1/2}\boldsymbol{W}\|^2 \\
&= \|\mathbf{H}_\ell(\mathbf{I}-\mathbf{\Lambda})^{1/2}\boldsymbol{X}_{\ell 1}\|^2 + \|\widetilde{\mathbf{H}}_\ell \boldsymbol{X}_{\ell 2}\|^2 + \|\mathbf{H}_\ell \mathbf{\Lambda}^{1/2}\boldsymbol{W}\|^2, \\
&= \underbrace{\|(\mathbf{I}-\mathbf{\Lambda})^{1/2}\boldsymbol{X}_{\ell 1}\|^2 + \|\boldsymbol{X}_{\ell 2}\|^2}_{\equiv U_\ell} + \underbrace{\|\mathbf{\Lambda}^{1/2}\boldsymbol{W}\|^2}_{\equiv V}, \qquad (77)
\end{aligned}
$$

where the third line is from the zero mean and mutual independence of $(\boldsymbol{X}_{\ell 1}, \boldsymbol{X}_{\ell 2}, \boldsymbol{W})$ and the last line is due to that $\mathbf{H}_\ell'\mathbf{H}_\ell = \mathbf{I}_k$ and $\widetilde{\mathbf{H}}_\ell\widetilde{\mathbf{H}}_\ell' = \mathbf{I}_{k_\ell - k}$. Since (U_1, U_2, V) are mutually independent, it follows that

$$
\begin{aligned}
\operatorname{cov}\big(\|\boldsymbol{Z}_1\|^{2m}, \|\boldsymbol{Z}_2\|^{2m}\big) &= \operatorname{cov}\left(\sum_{j_1=1}^m \binom{m}{j_1} U_1^{m-j_1}V^{j_1}, \quad \sum_{j_2=1}^m \binom{m}{j_2} U_2^{m-j_2}V^{j_1}\right) \\
&= \sum_{j_1,j_2=1}^m \binom{m}{j_1}\binom{m}{j_2} \operatorname{cov}(U_1^{m-j_1}V^{j_1}, U_2^{m-j_2}V^{j_2}) \\
&= \sum_{j_1,j_2=1}^m \binom{m}{j_1}\binom{m}{j_2} (\mathbb{E}U_1^{m-j_1})(\mathbb{E}U_2^{m-j_2}) \operatorname{cov}(V^{j_1}, V^{j_2}). \qquad (78)
\end{aligned}
$$

It is not hard to see that $\operatorname{cov}(V^{j_1}, V^{j_2}) \geq 0$. Hence, $\operatorname{cov}\big(\|\boldsymbol{Z}_1\|^{2m}, \|\boldsymbol{Z}_2\|^{2m}\big) \geq 0$. Furthermore, since all entries of the diagonal matrix $\mathbf{\Lambda}$ are between 0 and 1, we have

$$
U_\ell \leq \sum_{j=1}^{k_\ell} X_\ell^2(j), \qquad V \leq \|\mathbf{\Lambda}\| \sum_{j=1}^{k} W^2(j),
$$

where $X_\ell(j)$'s and $W(j)$'s are all *iid* standard normal variables. In particular,

$$
0 \leq \mathbb{E}U_\ell^{m-j_\ell} \leq \mathbb{E}\chi_{k_\ell}^{2(m-j_\ell)}, \qquad \operatorname{cov}(V^{j_1}, V^{j_2}) \leq \mathbb{E}V^{j_1+j_2} \leq \|\mathbf{\Lambda}\|^{j_1+j_2}\mathbb{E}\chi_k^{2(j_1+j_2)}.
$$

Plugging these results into (78) gives

$$
\begin{aligned}
\frac{\operatorname{cov}(\|\boldsymbol{Z}_1\|^{2m}, \|\boldsymbol{Z}_2\|^{2m})}{(\mathbb{E}\|\boldsymbol{Z}_1\|^{2m})(\mathbb{E}\|\boldsymbol{Z}_1\|^{2m})} &= \frac{\operatorname{cov}(\|\boldsymbol{Z}_1\|^{2m}, \|\boldsymbol{Z}_2\|^{2m})}{(\mathbb{E}\chi_{k_1}^{2m})(\mathbb{E}\chi_{k_2}^{2m})} \\
&\leq \sum_{j_1,j_2=1}^m \|\mathbf{\Lambda}\|^{j_1+j_2}\binom{m}{j_1}\binom{m}{j_2} \frac{(\mathbb{E}\chi_{k_1}^{2m-2j_1})(\mathbb{E}\chi_{k_2}^{2m-2j_2})(\mathbb{E}\chi_k^{2(j_1+j_2)})}{(\mathbb{E}\chi_{k_1}^{2m})(\mathbb{E}\chi_{k_2}^{2m})}
\end{aligned}
$$

We note that m is bounded, but (k_1, k_2, k) can grow with (n, p). Note that $\mathbb{E}\chi_k^{2m} = \prod_{j=0}^{m-1}(k+2j)$ for all $k, m \geq 1$. As a result,

$$
\begin{aligned}
\frac{(\mathbb{E}\chi_{k_1}^{2m-2j_1})(\mathbb{E}\chi_{k_2}^{2m-2j_2})(\mathbb{E}\chi_k^{2(j_1+j_2)})}{(\mathbb{E}\chi_{k_1}^{2m})(\mathbb{E}\chi_{k_2}^{2m})} &= \frac{\prod_{j=0}^{m-j_1-1}(k_1+2j)\prod_{j=0}^{m-j_2-1}(k_2+2j)\prod_{j=0}^{j_1+j_2-1}(k+2j)}{\prod_{j=0}^{m-1}(k_1+2j)\prod_{j=0}^{m-1}(k_2+2j)} \\
&= \frac{\prod_{j=0}^{j_1+j_2-1}(k+2j)}{\prod_{j=m-j_1}^{m-1}(k_1+2j)\prod_{j=m-j_2}^{m-1}(k_2+2j)} \leq 1.
\end{aligned}
$$

Therefore,

$$
\frac{\operatorname{cov}(\|\boldsymbol{Z}_1\|^{2m}, \|\boldsymbol{Z}_2\|^{2m})}{(\mathbb{E}\|\boldsymbol{Z}_1\|^{2m})(\mathbb{E}\|\boldsymbol{Z}_1\|^{2m})} \leq \sum_{j_1,j_2=1}^m \|\mathbf{\Lambda}\|^{j_1+j_2}\binom{m}{j_1}\binom{m}{j_2} = O(\|\mathbf{\Lambda}\|^2). \qquad (79)
$$

Noticing that $\mathbf{\Lambda}$ is a diagonal matrix containing the singular values of $\mathbf{\Gamma}$, we have proved the claim.

C The case of multivariate Gaussian distributions

We present a corollary about the errors of MAE and BAE for the special case of multivariate Gaussian distributions. Here $R_n(\widehat{\theta}_2) = \mathbb{E}[(\widehat{\theta}_2 - \theta_2)^2/\theta_2^2]$. The proof is elementary and omitted.

Corollary C.1. *Let $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ be i.i.d. samples of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For a constant integer $k \geq 2$, we assume the blocks in BAE are $J_i = \{(i-1)k+1, (i-1)k+2, \cdots, \min\{ik, p\}\}$, $1 \leq i \leq \lceil p/k \rceil$.*

- *Suppose $\boldsymbol{\Sigma} = \mathbf{I}_p$. Then, $R_n(\widehat{\theta}_2^{\mathrm{I}}) \sim \frac{8}{np}$, $R_n(\widehat{\theta}_2^{\mathrm{M}}) \sim \frac{32}{3np}$, and $R_n(\widehat{\theta}_2^{\mathrm{B}}) \sim \frac{8(k+3)}{(k+2)np}$.*
- *Suppose $\boldsymbol{\Sigma}$ is a block-wise diagonal matrix with 2×2 blocks, where each block has diagonals 1 and off-diagonals $\rho \in (-1, 1)$. Let $k = 2$ in BAE. Then, $R_n(\widehat{\theta}_2^{\mathrm{I}}) \sim \frac{8}{np}$, $R_n(\widehat{\theta}_2^{\mathrm{M}}) \sim \frac{8(4+3\rho^2+\rho^4)}{3np}$, and $R_n(\widehat{\theta}_2^{\mathrm{B}}) \sim \frac{10}{np}$.*

D Simulations for the estimator in Section 5

We conducted simulations to investigate the performance of the estimator of realized ξ_t in Section 5.

In the first experiment, we generate $\{\boldsymbol{Y}_t\}_{t=1}^T$ iid from model (1) with a constant covariance matrix $\boldsymbol{\Sigma}$. The covariance is set to be $\Sigma_{ij} = 0.3^{|i-j|}$, which is approximately banded. We fix $T = 100$ and let p varies. The results are displayed in Figure 9, where we study both cases of multivariate Gaussian data and multivariate $t_{4.5}$ data. We see that the estimated values are very close to the true values in all the cases.

In the second experiment, we generate data using the calibrated covariance matrix from S&P500 stock returns as in Section 4. In this case, the covariance matrix is heavily non-sparse, however, our estimator still works very well, no matter for Gaussian data or heavy-tailed data with multivariate t-distributions.

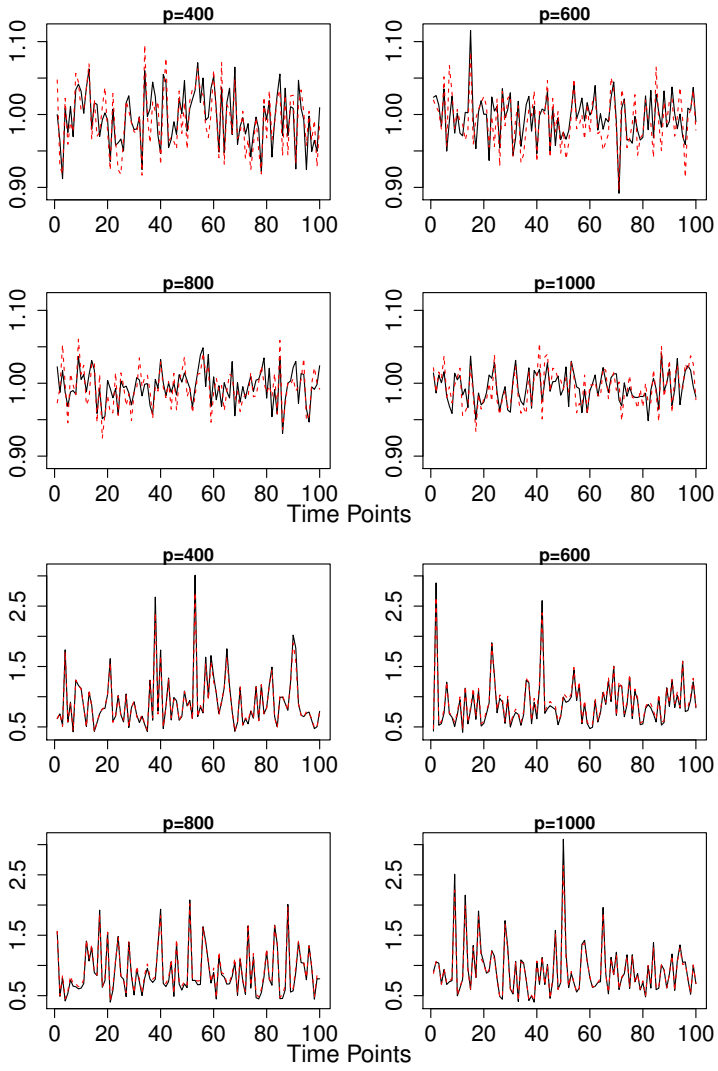

Figure 9: Estimated ξ_t (red broken line) versus true ξ_t (solid black line). The covariance matrix $\mathbf{\Sigma}$ is sparse. Top four panels: multivariate Gaussian data. Bottom four panels: multivariate t data.

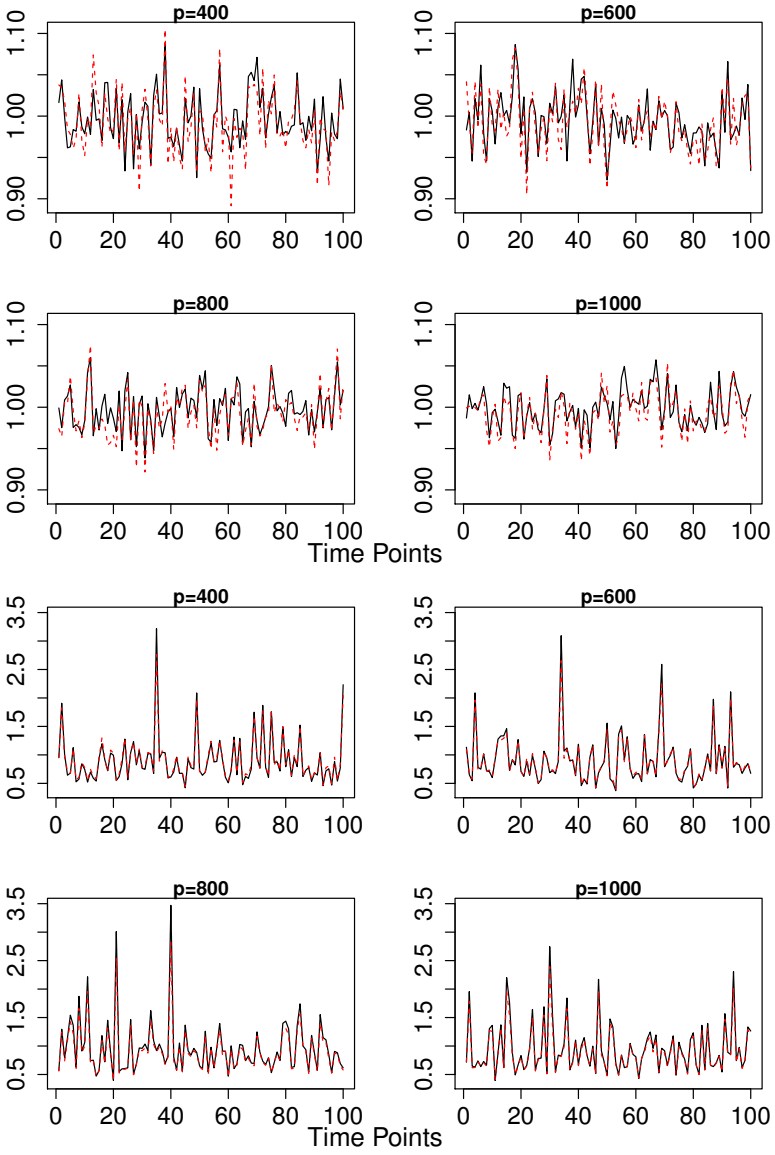

Figure 10: Estimated ξ_t (red broken line) versus true ξ_t (solid black line). The covariance matrix $\mathbf{\Sigma}$ is calibrated from S&P stock returns and is dense. Top four panels: multivariate Gaussian data. Bottom four panels: multivariate t data.