2016/44

Estimation of a Multiplicative Covariance Structure in the Large Dimensional Case

CHRISTIAN M. HAFNER, OLIVER B. LINTON AND HAIHAN TANG

CORE
Voie du Roman Pays 34, L1.03.01
Tel (32 10) 47 43 04
Fax (32 10) 47 43 01
Email: immaq-library@uclouvain.be
http://www.uclouvain.be/en-44508.html

Estimation of a Multiplicative Covariance Structure in the Large Dimensional Case*

Christian M. Hafner†
Université catholique de Louvain

Oliver B. Linton‡
University of Cambridge

Haihan Tang§
University of Cambridge

November 26, 2016

Abstract

We propose a Kronecker product structure for large covariance or correlation matrices. One feature of this model is that it scales logarithmically with dimension in the sense that the number of free parameters increases logarithmically with the dimension of the matrix. We propose an estimation method of the parameters based on a log-linear property of the structure, and also a quasi-maximum likelihood estimation (QMLE) method. We establish the rate of convergence of the estimated parameters when the size of the matrix diverges. We also establish a central limit theorem (CLT) for our method. We derive the asymptotic distributions of the estimators of the parameters of the spectral distribution of the Kronecker product correlation matrix, of the extreme logarithmic eigenvalues of this matrix, and of the variance of the minimum variance portfolio formed using this matrix. We also develop tools of inference including a test for over-identification. We apply our methods to portfolio choice for S&P500 daily returns and compare with sample covariance-based methods and with the recent Fan, Liao, and Mincheva (2013) method.

Some key words: Correlation matrix; Kronecker product; Matrix logarithm; Multiarray data; Portfolio choice; Sparsity

JEL subject classification: C55, C58, G11

*We would like to thank Liudas Giraitis, Anders Bredahl Kock, Alexei Onatskiy, Chen Wang, Tom Wansbeek, and Jianbin Wu for helpful comments.

†Corresponding author. CORE and Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Voie du Roman Pays 20, B-1348. Louvain-la-Neuve, Belgium, christian.hafner@uclouvain.be. Financial support of the *Académie universitaire Louvain* is gratefully acknowledged.

‡Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD. Email: obl20@cam.ac.uk.Thanks to the ERC for financial support.

§Faculty of Economics, University of Cambridge. Email: hht23@cam.ac.uk.

1 Introduction

Covariance matrices are of great importance in many fields including finance and psychology. They are a key element in portfolio choice and risk management. In psychology there is a long history of modelling unobserved traits through factor models that imply specific structures on the covariance matrix of observed variables. Anderson (1984) is a classic reference on multivariate analysis that treats estimation of covariance matrices and testing hypotheses on them. More recently, theoretical and empirical work has considered the case where the covariance matrix is large, because in the era of big data, many datasets now used are large. For example, in financial applications there are many securities that one may consider in selecting a portfolio, and indeed finance theory argues one should choose a well diversified portfolio that perforce includes a large number of assets with non-zero weights. Although in practice the portfolios of many investors concentrate on a small number of assets, there are many exceptions to this. For example, the listed company Knight Capital Group claim to make markets in thousands of securities worldwide, and are constantly updating their inventories/portfolio weights to optimize their positions. In the large dimensional case, standard statistical methods based on eigenvalues and eigenvectors such as Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) break down, and applications to for example portfolio choice face considerable difficulties, see Wang and Fan (2016). There are many new methodological approaches for the large dimensional case, see for example Ledoit and Wolf (2003), Bickel and Levina (2008), Onatski (2009), Fan, Fan, and Lv (2008), and Fan et al. (2013). The general approach is to impose some sparsity on the model, meaning that many elements of the covariance matrix are assumed to be zero or small, thereby reducing the effective number of parameters that have to be estimated, or to use a shrinkage method that achieves effectively the same dimensionality reduction. Yao, Zheng, and Bai (2015) give an excellent account of the recent developments in the theory and practice of estimating large dimensional covariance matrices.

We consider a parametric model for the covariance or correlation matrix, the Kronecker product structure. This has been previously considered in Swain (1975) and Verhees and Wansbeek (1990) under the title of multimode analysis. Verhees and Wansbeek (1990) defined several estimation methods based on least squares and maximum likelihood principles, and provided asymptotic variances under assumptions that the data are normal and that the covariance matrix dimension is fixed. There is also a growing Bayesian and Frequentist literature on multiway array or tensor datasets, where this structure is commonly employed. See for example Akdemir and Gupta (2011), Allen (2012), Browne, MacCallum, Kim, Andersen, and Glaser (2002), Cohen, Usevich, and Comon (2016), Constantinou, Kokoszka, and Reimherr (2015), Dobra (2014), Fosdick and Hoff (2014), Gerard and Hoff (2015), Hoff (2011), Hoff (2015), Hoff (2016), Krijnen (2004), Leiva and Roy (2014), Leng and Tang (2012), Li and Zhang (2016), Manceura and Dutilleul (2013), Ning and Liu (2013), Ohlson, Ahmada, and von Rosen (2013), Singull, Ahmad, and von Rosen (2012), Volfovsky and Hoff (2014), Volfovsky and Hoff (2015), and Yin and Li (2012). In both these (apparently separate) literatures the dimension n is fixed and typically there are a small number of products each of whose dimension is of fixed but perhaps moderate size.

We consider the Kronecker product model in the setting where the matrix dimension n is large, i.e., increases with the sample size T. We allow the number of lower dimensional matrices of a Kronecker product to increase with n according to the prime factorization of n. In this setting, the model effectively imposes sparsity on the covariance/correlation

matrix, since the number of parameters in a Kronecker product covariance/correlation matrix grows *logarithmically* with n. In fact we show that the logarithm of a Kronecker product covariance/correlation matrix has many zero elements, so that sparsity is explicitly placed inside the logarithm of the covariance/correlation matrix. We do not impose a multi-array structure on the data a priori and our methods are applicable in cases where this structure is not present.

The Kronecker product structure has a number of intrinsic advantages for applications. First, the eigenvalues of a Kronecker product are products of the eigenvalues of its lower dimensional matrices, and the inverse, determinant, and other key quantities of it are easily obtained from the corresponding quantities of the lower dimensional matrices, which facilitates computation and analysis. Second, it can generate a very flexible eigenstructure. It is easy to establish limit laws for the population eigenvalues of the Kronecker product and to establish properties of the corresponding sample eigenvalues. In particular, under some conditions the eigenvalues of a large Kronecker product covariance/correlation matrix are log normally distributed. Empirically, this seems to be not a bad approximation for daily stock returns. Third, even when a Kronecker product structure is not true for a covariance/correlation matrix, we show that there always exists a Kronecker product matrix closest to the covariance/correlation matrix in the sense of minimising some norm in the logarithmic matrix space.

We show that the logarithm of the Kronecker product covariance/correlation matrix (closest to the covariance/correlation matrix) is linear in the unknown parameters, denoted θ^0, and use this as the basis for a closed-form minimum distance estimator $\hat{\theta}_T$ of θ^0. This allows some direct theoretical analysis, although this method is likely to be computationally intensive. We also propose a quasi-maximum likelihood estimator (QMLE) and an approximate QMLE (one-step estimator), the latter of which achieves the Cramer-Rao lower bound in the finite n case. We establish the rate of convergence and asymptotic normality of the estimated parameters when both n and T diverge under restrictions on the relative rate of growth of these quantities. In particular, we show that $\|\hat{\theta}_T - \theta^0\|_2 = O_p((n\kappa(W)/T)^{1/2})$, which improves on the crude rate implied by the unrestricted correlation matrix estimator, $O_p((n^2/T)^{1/2})$. Our QMLE procedure works much better numerically than the sample correlation matrix, consistent with the faster rate of convergence we expect.

There is a large literature on the optimal rates of estimation of high-dimensional covariance and its inverse (i.e., *precision*) matrices (see Cai, Zhang, and Zhou (2010) and Cai and Zhou (2012)). Cai, Ren, and Zhou (2014) gave a nice review on those recent results. However their optimal rates are not applicable to our setting because here sparsity is not imposed on the covariance matrix, but on its logarithm.

We provide a feasible central limit theorem (CLT) for inference regarding θ^0 and certain non-linear functions thereof. For example, we derive the CLT for the mean and variance of the spectral distribution of the logarithmic Kronecker product correlation matrix as well as for its extreme eigenvalues. The extreme eigenvalues of the sample correlation matrix are known to behave poorly when the dimension of the matrix increases, but in our case because of the tight structure we impose we obtain consistency and a CLT under general conditions. We also apply our methods to the question of estimating the variance of the minimum variance portfolio formed using the Kronecker product correlation matrix. Last, we give an over-identification test which allows us to test whether a correlation matrix has a Kronecker product structure or not.

We provide some evidence that the proposed procedures work well numerically both

when the Kronecker product structure is true for the covariance/correlation matrix and when it is not true. We also apply the method to portfolio selection and compare our method with the sample covariance matrix, a strict factor model, and the Fan et al. (2013) method. Our performance is close to that of Fan et al. (2013) and beats the other two methods.

The rest of the paper is structured as follows. In Section 2 we discuss our Kronecker product model in detail while in Section 3 we give three motivations of our model. We address identification and estimation in Sections 4 and 5, respectively. Section 6 gives the asymptotic properties of the minimum distance estimator, of a one-step approximation of the QMLE, of the estimators of the parameters of the spectral distribution, of the estimators of the extreme logarithmic eigenvalues, and of the estimator of the variance of the minimum variance portfolio. We also provide an over-identification test. In Section 7 we address some model selection issues. Sections 8 and 9 provide numerical evidence for the performance of the model in a simulation study and an empirical application, respectively. Section 10 concludes. All the proofs are deferred to Appendix A; further auxiliary lemmas needed in Appendix A are provided in Appendix B.

2 The Model

2.1 Notation

For $x \in \mathbb{R}^n$, let $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ denote the Euclidean norm. For any real matrix A, let maxeval(A) and mineval(A) denote its maximum and minimum eigenvalues, respectively. Let $\|A\|_F := [\mathrm{tr}(A^\intercal A)]^{1/2} \equiv [\mathrm{tr}(AA^\intercal)]^{1/2} \equiv \|\mathrm{vec}A\|_2$ and $\|A\|_{\ell_2} := \max_{\|x\|_2=1} \|Ax\|_2 \equiv \sqrt{\mathrm{maxeval}(A^\intercal A)}$ denote the Frobenius norm and spectral norm of A, respectively.

Let A be a $m \times n$ matrix. vec$A$ is a vector obtained by stacking the columns of the matrix A one underneath the other. The *commutation matrix* $K_{m,n}$ is a $mn \times mn$ *orthogonal* matrix which translates vec$A$ to vec$(A^\intercal)$, i.e., vec$(A^\intercal) = K_{m,n}$vec(A). If A is a symmetric $n \times n$ matrix, its $n(n-1)/2$ supradiagonal elements are redundant in the sense that they can be deduced from the symmetry. If we eliminate these redundant elements from vec$A$, this defines a new $n(n+1)/2 \times 1$ vector, denoted vech$A$. They are related by the full-column-rank, $n^2 \times n(n+1)/2$ *duplication matrix* D_n: vec$A = D_n$vech$A$. Conversely, vech$A = D_n^+$vec$A$, where D_n^+ is the Moore-Penrose generalised inverse of D_n. In particular, $D_n^+ = (D_n^\intercal D_n)^{-1} D_n^\intercal$ because D_n is full-column rank.

Consider two sequences of real random matrices X_T and Y_T. $X_T = O_p(\|Y_T\|)$, where $\|\cdot\|$ is some matrix norm, means that for every real $\varepsilon > 0$, there exist $M_\varepsilon > 0$ and $T_\varepsilon > 0$ such that for all $T > T_\varepsilon$, $\mathbb{P}(\|X_T\|/\|Y_T\| > M_\varepsilon) < \varepsilon$. $X_T = o_p(\|Y_T\|)$, where $\|\cdot\|$ is some matrix norm, means that $\|X_T\|/\|Y_T\| \xrightarrow{p} 0$ as $T \to \infty$.

Let $a \vee b$ and $a \wedge b$ denote $\max(a, b)$ and $\min(a, b)$, respectively. For two real sequences a_T and b_T, $a_T \lesssim b_T$ means that $a_T \leq C b_T$ for some positive real number C for all $T \geq 1$. $a_T \sim b_T$ means that a_T and b_T are asymptotically equivalent, i.e., $a_T/b_T \to 1$ as $T \to \infty$. For $x \in \mathbb{R}$, let $\lfloor x \rfloor$ denote the greatest integer *strictly less* than x and $\lceil x \rceil$ denote the smallest integer greater than or equal to x.

For matrix calculus, what we adopt is called the *numerator layout* or *Jacobian formulation*; that is, the derivative of a scalar with respect to a column vector is a row vector. As the result, our chain rule is never backward.

2.2 The Covariance Matrix

Suppose that the i.i.d. series $x_t \in \mathbb{R}^n$ $(t = 1, \ldots, T)$ with mean μ has the covariance matrix

$$\Sigma := \mathbb{E}(x_t - \mu)(x_t - \mu)^\intercal,$$

where Σ is positive definite. Suppose that n is composite and has a factorization $n = n_1 n_2 \cdots n_v$ (n_j may not be distinct).[1] Then consider the $n \times n$ matrix

$$\Sigma^* = \Sigma_1^* \otimes \Sigma_2^* \otimes \cdots \otimes \Sigma_v^*, \tag{2.1}$$

where Σ_j^* are $n_j \times n_j$ matrices. When each submatrix Σ_j^* is positive definite, then so is Σ^*. The total number of free parameters in Σ^* is $\sum_{j=1}^v n_j(n_j+1)/2$, which is much less than $n(n+1)/2$. When $n = 256$, the eightfold factorization with 2×2 matrices has 24 parameters, while the unconstrained covariance matrix has 32,896 parameters. In many cases it is possible to consider intermediate factorizations with different numbers of parameters (see Section 7). We note that the Kronecker product structure is invariant under the Lie group of transformations $\mathcal{G}$ generated by $A_1 \otimes A_2 \otimes \cdots \otimes A_v$, where A_j are $n_j \times n_j$ nonsingular matrices (see Browne and Shapiro (1991)). This structure can be used to characterise the tangent space $\mathcal{T}$ of $\mathcal{G}$ and to define a relevant equivariance concept for restricting the class of estimators for optimality considerations.

This Kronecker product structure does arise naturally in various contexts. For example, suppose that $u_{i,t}$ are errors terms in a panel regression model with $i = 1, \ldots, n$ and $t = 1, \ldots, T$, The interactive effects model of Bai (2009) is that $u_{i,t} = \gamma_i f_t$, which implies that $u = \gamma \otimes f$, where u is the $nT \times 1$ vector containing all the elements of $u_{i,t}$, $\gamma = (\gamma_1, \ldots, \gamma_n)^\intercal$, and $f = (f_1, \ldots, f_T)^\intercal$. If we assume that γ, f are random, γ is independent of f, and both vectors have mean zero, this implies that

$$\operatorname{var}(u) = \mathbb{E}[uu^\intercal] = \mathbb{E}\gamma\gamma^\intercal \otimes \mathbb{E}ff^\intercal.$$

We can think of our more general structure (2.1) arising from a multi-index setting with v multiplicative factors. One interpretation here is that there are v different indices that define an observation, as arises naturally in multiarray data (see Hoff (2015)). One might suppose that

$$u_{i_1,i_2,\ldots,i_v} = \varepsilon_{1,i_1}\varepsilon_{2,i_2}\cdots\varepsilon_{v,i_v}, \qquad i_j = 1, \ldots, n_j, \quad j = 1, \ldots, v,$$

where the random variables $\varepsilon_1, \ldots, \varepsilon_v$ are mutually independent and mean zero; in vector form

$$u = (u_{1,1,\ldots,1}, \ldots, u_{n_1,n_2,\ldots,n_v})^\intercal = \epsilon_1 \otimes \epsilon_2 \otimes \cdots \otimes \epsilon_v, \tag{2.2}$$

where $\epsilon_j = (\varepsilon_{j,1}, \ldots, \varepsilon_{j,n_j})^\intercal$ is a mean zero random vector of length n_j with covariance matrix Σ_j for $j = 1, \ldots, v$. Then

$$\Sigma = \mathbb{E}[uu^\intercal] = \Sigma_1 \otimes \Sigma_2 \otimes \cdots \otimes \Sigma_v.$$

[1] Note that if n is not composite one can add a vector of additional pseudo variables to the system until the full system is composite. It is recommended to add a vector of independent variables $u_t \sim N(0, I_k)$, where $n + k = 2^v$, say. Let $z_t = (x_t^\intercal, u_t^\intercal)^\intercal$ denote the $2^v \times 1$ vector with covariance matrix

$$B = \begin{bmatrix} \Sigma & 0 \\ 0 & I_k \end{bmatrix} = B_1 \otimes B_2 \otimes \cdots \otimes B_v,$$

where B_j are 2×2 positive definite matrices for $j = 1, \ldots, v$.

One motivation for considering this structure is that in a number of contexts multiplicative effects may be a valid description of relationships, especially in the multi-trait multi-method (MTMM) context in psychometrics (see e.g., Campbell and O'Connell (1967) and Cudeck (1988)). For a financial application one might consider different often employed sorting characteristics such as industry, size, and value, by which each stock is labelled. For example, we may have 10 industries, 3 sizes and 3 different value buckets, which yields 90 buckets. If one has precisely one firm in each industry ε_1, of each size ε_2 and of each value category ε_3 then the multi-array model is directly applicable.

This structure has been considered before by Swain (1975) and Verhees and Wansbeek (1990), and in the multi-array literature, where they emphasize the case where v is small and n_j is fixed and where the structure is known and correct (up to the unknown parameters of $\Sigma_1^*, \Sigma_2^*, \ldots, \Sigma_v^*$). Our framework emphasizes the case where v is large and n_j is fixed; in addition we do not explicitly require the multi-array structure and consider Σ^* in (2.1) as an approximation device to a general large covariance matrix Σ.

For comparison, consider the multi-way additive random effect model

$$u_{i_1,i_2,\ldots,i_v} = \varepsilon_{1,i_1} + \varepsilon_{2,i_2} + \cdots + \varepsilon_{v,i_v},$$

where the errors $\varepsilon_1, \ldots, \varepsilon_v$ are mutually uncorrelated. We can write the full $n \times 1$ vector $u = (u_{1,1,\ldots,1}, \ldots, u_{n_1,n_2,\ldots,n_v})^\intercal$ as

$$u = \sum_{j=1}^{v} D_j \epsilon_j,$$

where D_j are known $n \times n_j$ matrices of zeros and ones, so that

$$\Sigma = \mathbb{E}[uu^\intercal] = \sum_{j=1}^{v} D_j \Sigma_j D_j^\intercal,$$

(see for example Rao (1997)). In some sense as we shall see in Section 4 the Kronecker product structure corresponds to a kind of additive structure on the logarithm of the covariance matrix, and from a mathematical point of view log-linear models have some advantages over linear models for covariance, Shephard (1996).

There are two issues with the model (2.1). First, there is an identification problem even though the number of parameters in (2.1) is strictly less than $n(n+1)/2$. For example, if we multiply every element of Σ_1^* by a constant C and divide every element of Σ_2^* by C, then Σ^* is the same. A solution to the identification problem is to normalize $\Sigma_1^*, \Sigma_2^*, \cdots, \Sigma_{v-1}^*$ by setting the first diagonal element to be 1. Second, if the matrices Σ_j^*s are permuted one obtains a different Σ^*. Although the eigenvalues of this permuted matrix are the same, the eigenvectors are not. It is also the case that if the data are permuted then the Kronecker structure may be lost. This may be an issue in some applications, and begs the question of how one chooses the correct permutation; we discuss this briefly in Section 7.

2.3 The Transformed Covariance Matrix

In this paper, we will approximate a transformation of the covariance matrix with a Kronecker product structure. For example, we will model the *correlation matrix* instead of the covariance matrix. This will allow a more flexible approach to approximating a general covariance matrix, since we can estimate the diagonal elements by standard well

understood (even in the large dimensional case) methods; this will be useful in some applications.

Suppose again that we observe a sample of n-dimensional random vectors x_t, $t = 1, 2, \ldots, T$, which are i.i.d. distributed with mean $\mu := \mathbb{E}x_t$ and a positive definite $n \times n$ covariance matrix $\Sigma := \mathbb{E}(x_t - \mu)(x_t - \mu)^\intercal$. Let D be an $n \times n$ known diagonal matrix. For example, $D := \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, where $\sigma_i^2 := \mathbb{E}(x_{t,i} - \mu_i)^2$. Then define

$$y_t := D^{-1/2}(x_t - \mu)$$

such that $\mathbb{E}y_t = 0$ and $\mathrm{var}[y_t] = D^{-1/2}\Sigma D^{-1/2} =: \Theta$. In the case where $D :=\mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, Θ is the correlation matrix; that is, it has all its diagonal entries to be 1. In the case where $D = I_n$, Θ is the covariance matrix. We will assume that the matrix Θ possesses the Kronecker product structure. For simplicity, in the remainder of this paper we shall assume that Θ is the correlation matrix, the general case follows along similar lines.

Suppose $n = 2^v$. We show in Section 3.1 that there exists a *unique* matrix

$$\Theta^0 = \Theta_1^0 \otimes \Theta_2^0 \otimes \cdots \otimes \Theta_v^0 = \begin{bmatrix} 1 & \rho_1^0 \\ \rho_1^0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho_2^0 \\ \rho_2^0 & 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 & \rho_v^0 \\ \rho_v^0 & 1 \end{bmatrix} \tag{2.3}$$

that minimizes $\|\log\Theta - \log\Theta^*\|_W$ among all $\log\Theta^*$, where the norm $\|\cdot\|_W$ is defined in Section 3.1. Define

$$\begin{aligned}\Omega^0 &:= \log\Theta^0 \\ &= (\log\Theta_1^0 \otimes I_2 \otimes \cdots \otimes I_2) + (I_2 \otimes \log\Theta_2^0 \otimes \cdots \otimes I_2) + \cdots + (I_2 \otimes \cdots \otimes \log\Theta_v^0), \\ &=: (\Omega_1^0 \otimes I_2 \otimes \cdots \otimes I_2) + (I_2 \otimes \Omega_2^0 \otimes \cdots \otimes I_2) + \cdots + (I_2 \otimes \cdots \otimes \Omega_v^0),\end{aligned} \tag{2.4}$$

where Ω_i^0 is 2×2 for $i = 1, \ldots, v$. For the moment consider $\Omega_1^0 := \log\Theta_1^0$. The eigenvalues of Θ_1^0 are $1 + \rho_1$ and $1 - \rho_1$, respectively. The corresponding eigenvectors are $(1, 1)^\intercal/\sqrt{2}$ and $(1, -1)^\intercal/\sqrt{2}$, respectively. Therefore

$$\begin{aligned}\Omega_1^0 = \log\Theta_1^0 &= \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \log(1+\rho_1^0) & 0 \\ 0 & \log(1-\rho_1^0) \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \frac{1}{2} \\ &= \begin{pmatrix} \frac{1}{2}\log(1-[\rho_1^0]^2) & \frac{1}{2}\log\left(\frac{1+\rho_1^0}{1-\rho_1^0}\right) \\ \frac{1}{2}\log\left(\frac{1+\rho_1^0}{1-\rho_1^0}\right) & \frac{1}{2}\log(1-[\rho_1^0]^2) \end{pmatrix} =: \begin{pmatrix} a_1 & b_1 \\ b_1 & a_1 \end{pmatrix},\end{aligned} \tag{2.5}$$

whence we see that ρ_1^0 generates two distinct entries for Ω_1^0. The off-diagonal element $\frac{1}{2}\log\left(\frac{1+\rho_1^0}{1-\rho_1^0}\right)$ is the Fisher's z-transformation of ρ_1^0, which has a fine statistical pedigree. We also see that Ω_1^0 is not only symmetric about the diagonal, but also symmetric about the cross-diagonal (from the upper right to the lower left). We can use entries of Ω_1^0 to recover ρ_1^0 in some over-identified sense. The same reasoning applies to $\Omega_2^0, \ldots, \Omega_v^0$. We achieve dimension reduction because the original Θ has $n(n-1)/2$ parameters whereas Θ^0 has only $v = O(\log n)$ parameters. We shall discuss various aspects of estimation in detail in Section 5.

3 Motivation

In this section we give three motivational reasons for considering the Kronecker product model beyond the obvious case arising from multi-array data structures. First, we show

that for any given covariance/correlation matrix there is a uniquely defined member of the model that is closest to it in some sense. Second, we also discuss whether the model can approximate an arbitrarily large covariance matrix well. In particular, we show that the eigenstructure of large Kronecker product matrices can be easily described. Third, we argue that the structure is very convenient for a number of applications.

3.1 Best Approximation

For simplicity of notation, we suppose that $n = n_1 n_2$. Consider the set $\mathcal{C}_n$ of all $n \times n$ real positive definite matrices with the form

$$\Sigma^* = \Sigma_1^* \otimes \Sigma_2^*,$$

where Σ_j^* is a $n_j \times n_j$ matrix for $j = 1, 2$. We assume that both Σ_1^* and Σ_2^* are positive definite, which ensures that Σ^* is so. Regarding the identification issue we impose that the first diagonal of Σ_1^* is 1. Since Σ_1^* and Σ_2^* are symmetric, we can orthogonally diagonalize them:

$$\Sigma_1^* = U_1^\intercal \Lambda_1 U_1 \qquad \Sigma_2^* = U_2^\intercal \Lambda_2 U_2,$$

where U_1 and U_2 are orthogonal, and $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_{n_1})$ and $\Lambda_2 = \mathrm{diag}(u_1, \ldots, u_{n_2})$ are diagonal matrices containing eigenvalues. Positive definiteness of Σ_1^* and Σ_2^* ensures that these eigenvalues are real and positive. Then the (principal) logarithm of Σ^* is:

$$\begin{aligned} \log \Sigma^* &= \log(\Sigma_1^* \otimes \Sigma_2^*) = \log[(U_1 \otimes U_2)^\intercal (\Lambda_1 \otimes \Lambda_2)(U_1 \otimes U_2)] \\ &= (U_1 \otimes U_2)^\intercal \log(\Lambda_1 \otimes \Lambda_2)(U_1 \otimes U_2), \end{aligned} \tag{3.1}$$

where the second equality is due to the mixed product property of the Kronecker product, and the third equality is due to a property of matrix functions. Now

$$\begin{aligned} &\log(\Lambda_1 \otimes \Lambda_2) = \mathrm{diag}(\log(\lambda_1 \Lambda_2), \ldots, \log(\lambda_{n_1} \Lambda_2)) \\ &= \mathrm{diag}(\log(\lambda_1 I_{n_2} \Lambda_2), \ldots, \log(\lambda_{n_1} I_{n_2} \Lambda_2)) \\ &= \mathrm{diag}(\log(\lambda_1 I_{n_2}) + \log(\Lambda_2), \ldots, \log(\lambda_{n_1} I_{n_2}) + \log(\Lambda_2)) \\ &= \mathrm{diag}(\log(\lambda_1 I_{n_2}), \ldots, \log(\lambda_{n_1} I_{n_2})) + \mathrm{diag}(\log(\Lambda_2), \ldots, \log(\Lambda_2)) \\ &= \log(\Lambda_1) \otimes I_{n_2} + I_{n_1} \otimes \log(\Lambda_2), \end{aligned} \tag{3.2}$$

where the third equality holds only because $\lambda_j I_{n_2}$ and Λ_2 have real positive eigenvalues only and commute for all $j = 1, \ldots, n_1$ (Higham (2008) p270 Theorem 11.3). Substitute (3.2) into (3.1):

$$\begin{aligned} \log \Sigma^* &= (U_1 \otimes U_2)^\intercal (\log \Lambda_1 \otimes I_{n_2} + I_{n_1} \otimes \log \Lambda_2)(U_1 \otimes U_2) \\ &= (U_1 \otimes U_2)^\intercal (\log \Lambda_1 \otimes I_{n_2})(U_1 \otimes U_2) + (U_1 \otimes U_2)^\intercal (I_{n_1} \otimes \log \Lambda_2)(U_1 \otimes U_2) \\ &= \log \Sigma_1^* \otimes I_{n_2} + I_{n_1} \otimes \log \Sigma_2^*. \end{aligned}$$

Let $\mathcal{D}_n$ denote the set of all such matrices like $\log \Sigma^*$ as Σ_1^*, Σ_2^* varies.

Let $\mathcal{M}_n$ denote the set of all $n \times n$ real symmetric matrices. For any $n(n+1)/2 \times n(n+1)/2$ positive definite matrix W, define a map

$$\langle A, B \rangle_W := (\mathrm{vech} A)^\intercal W \, \mathrm{vech} B.$$

It is easy to show that $\langle \cdot, \cdot \rangle_W$ is an inner product. $\mathcal{M}_n$ with inner product $\langle \cdot, \cdot \rangle_W$ can be identified by $\mathbb{R}^{n(n+1)/2}$ with the Euclidean inner product. Since $\mathbb{R}^{n(n+1)/2}$ with the

Euclidean inner product is a Hilbert space (for finite n), so is $\mathcal{M}_n$. The inner product $\langle \cdot, \cdot \rangle_W$ induces the following norm

$$\|A\|_W := \sqrt{\langle A, A \rangle_W} = \sqrt{(\mathrm{vech} A)^\intercal W \, \mathrm{vech} A}.$$

The subset $\mathcal{C}_n \subset \mathcal{M}_n$ is not a subspace of $\mathcal{M}_n$. First, $\otimes$ and $+$ do not distribute in general. That is, there might not exist positive definite $\Sigma^*_{1,3}$ and $\Sigma^*_{2,3}$ such that

$$\Sigma^*_{1,1} \otimes \Sigma^*_{2,1} + \Sigma^*_{1,2} \otimes \Sigma^*_{2,2} = \Sigma^*_{1,3} \otimes \Sigma^*_{2,3}.$$

Second, $\mathcal{C}_n$ is a positive cone, hence not necessarily a subspace. In fact, the smallest subspace of $\mathcal{M}_n$ that contains $\mathcal{C}_n$ is $\mathcal{M}_n$ itself. On the other hand, $\mathcal{D}_n$ is a subspace of $\mathcal{M}_n$ as

$$\begin{aligned}
&(\log \Sigma^*_{1,1} \otimes I_{n_2} + I_{n_1} \otimes \log \Sigma^*_{2,1}) + (\log \Sigma^*_{1,2} \otimes I_{n_2} + I_{n_1} \otimes \log \Sigma^*_{2,2}) \\
&= \left(\log \Sigma^*_{1,1} + \log \Sigma^*_{1,2}\right) \otimes I_{n_2} + I_{n_1} \otimes \left(\log \Sigma^*_{2,1} + \log \Sigma^*_{2,2}\right) \in \mathcal{D}_n.
\end{aligned}$$

For finite n, $\mathcal{D}_n$ is also closed. Therefore, for any positive definite covariance matrix $\Sigma \in \mathcal{M}_n$, by the projection theorem of the Hilbert space, there exists a unique $\log \Sigma^0 \in \mathcal{D}_n$ such that

$$\|\log \Sigma - \log \Sigma^0\|_W = \min_{\log \Sigma^* \in \mathcal{D}_n} \|\log \Sigma - \log \Sigma^*\|_W.$$

Note also that $\log \Sigma^{-1} = -\log \Sigma$, so that this model simultaneously approximates the precision matrix in the same norm.

This says that any covariance matrix Σ has a closest approximating matrix Σ^0 (in the least squares sense) that is of the Kronecker product form, and that its precision matrix Σ^{-1} has a closest approximating matrix $(\Sigma^0)^{-1}$. This kind of best approximating property is found in linear regression (Best Linear Predictor) and provides a justification (i.e., interpretation) for using this approximation Σ^0 even when the model is not true.[2] The same reasoning applies to any correlation matrix Θ.

3.2 Eigenvalues and Large n Approximation Properties

In general, a covariance matrix can have a wide variety of eigenstructures, meaning that the behaviour of its eigenvalues can be quite diverse. The widely used factor models have a rather limited eigenstructure. Specifically, in a factor model the covariance matrix (normalized by diagonal values) has a spikedness property, namely, there are K eigenvalues $1 + \delta_1, \ldots, 1 + \delta_K$, where $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_K > 0$, and $n - K$ eigenvalues that take the value one.

We next consider the eigenvalues of the class of matrices formed from the Kronecker parameterization. Without loss of generality suppose $n = 2^v$. We consider the 2×2 matrices $\{\Sigma^*_j : j = 1, 2, \ldots, v\}$. Let $\overline{\lambda}_j$ and $\underline{\lambda}_j$ denote the larger and smaller eigenvalues of Σ^*_j, respectively. The eigenvalues of the Kronecker product matrix

$$\Sigma^* = \Sigma^*_1 \otimes \cdots \otimes \Sigma^*_v$$

[2] van Loan (2000) and Pitsianis (1997) also considered this nearest Kronecker product problem involving one Kronecker product only and in the original space (not in the logarithm space). In that simplified problem, they showed that the optimisation problem could be solved by the singular value decomposition.

are all of the products including $\overline{\lambda}_1 \times \cdots \times \overline{\lambda}_v, \ldots, \underline{\lambda}_1 \times \cdots \times \underline{\lambda}_v$ with a cardinality of $n = 2^v$. Define $\overline{\omega}_j := \log \overline{\lambda}_j$ and $\underline{\omega}_j := \log \underline{\lambda}_j$; let $U = \{\overline{\omega}_1, \ldots, \overline{\omega}_v\}$ and $L = \{\underline{\omega}_1, \ldots, \underline{\omega}_v\}$ denote the sets of larger and smaller values, respectively. We may write the set of eigenvalues of Σ^* in terms of the power sets of U and L. In particular, the logarithm of a generic eigenvalue of Σ^* is of the form

$$l_I = \sum_{j \in I} \overline{\omega}_j + \sum_{j \in I^c} \underline{\omega}_j,$$

for some $I \subset \{1, 2, \ldots, v\}$, and I varies over all such subsets of $\{1, \ldots, v\}$. The largest and smallest logarithmic eigenvalues of Σ^* are

$$\omega^*_{(1)} = \sum_{j=1}^{v} \overline{\omega}_j \quad ; \quad \omega^*_{(n)} = \sum_{j=1}^{v} \underline{\omega}_j, \tag{3.3}$$

respectively. Depending on the choice of U and L, one can have quite different outcomes for $\omega^*_{(1)}, \omega^*_{(n)}$, namely one can have a bounded or expanding range, at various rates.

In fact, we can think of the generic logarithmic eigenvalue as being the outcome of a random process whereby for each $j = 1, \ldots, v$ we choose either $\overline{\omega}_j$ or $\underline{\omega}_j$ with equal probability and then form the sum over j. That is, we may write the generic logarithmic eigenvalue as

$$\sum_{j=1}^{v} \zeta_j, \qquad \zeta_j := e_j \overline{\omega}_j + (1 - e_j) \underline{\omega}_j$$

where e_j are i.i.d. binary variables with probability 1/2. The support of $\{e_1, \ldots, e_v\}$ traces out the possible values the logarithmic eigenvalues can take. The random variables $\{\zeta_j\}_{j=1}^v$ are independent with $\mathbb{E}\zeta_j = (\overline{\omega}_j + \underline{\omega}_j)/2$ and $\text{var}(\zeta_j) = (\overline{\omega}_j - \underline{\omega}_j)^2/4$. Under some restrictions on U and L, we may apply the Lindeberg CLT for triangular arrays of independent random variables to obtain

$$\frac{\sum_{j=1}^{v_n} \zeta_j - \sum_{j=1}^{v_n} \mathbb{E}\zeta_j}{\sqrt{\sum_{j=1}^{v_n} \text{var}(\zeta_j)}} \xrightarrow{d} N(0,1),$$

as $n \to \infty$. This says that the spectral distribution of Σ^* can be represented by the cumulative distribution function of the log normal distribution whose mean parameter is $\sum_{j=1}^{v_n} \mathbb{E}\zeta_j$ and variance parameter $\sum_{j=1}^{v_n} \text{var}(\zeta_j)$.

A sufficient condition for the CLT is the following Lyapounov condition (Billingsley (1995) p362): for some $\delta > 0$

$$\frac{\sum_{j=1}^{v_n} \mathbb{E}|\zeta_j - \mathbb{E}\zeta_j|^{2+\delta}}{\left(\sum_{j=1}^{v_n} \text{var}(\zeta_j)\right)^{(2+\delta)/2}} = \frac{\sum_{j=1}^{v_n} |\overline{\omega}_j - \underline{\omega}_j|^{2+\delta}}{\left(\sum_{j=1}^{v_n} (\overline{\omega}_j - \underline{\omega}_j)^2\right)^{(2+\delta)/2}} \to 0,$$

as $n \to \infty$, provided $\mathbb{E}|\zeta_j - \mathbb{E}\zeta_j|^{2+\delta} < \infty$. (We shall suppress the subscript n of v.) This condition will be satisfied in many settings. We next give an example in which this condition is easily verified.

Example 1. *Suppose that* $\overline{\omega}_j = v^{-\alpha}\phi\left(j/v\right)$ *and* $\underline{\omega}_j = v^{-\alpha}\mu\left(j/v\right)$ *for some fixed smooth bounded functions* $\phi(\cdot), \mu(\cdot)$ *such that* $\int_0^1 |\phi(u) - \mu(u)|^{2+\delta} du < \infty$ *for some* $\delta > 0$ *and* $\int_0^1 |\phi(u) - \mu(u)|^2 du > 0$. *Then the Lyapounov condition is satisfied. To see this,*

$$\sum_{j=1}^{v} |\overline{\omega}_j - \underline{\omega}_j|^{2+\delta} = v^{-\alpha(2+\delta)} \sum_{j=1}^{v} \left|\phi(j/v) - \mu(j/v)\right|^{2+\delta} \sim v^{1-\alpha(2+\delta)} \int_0^1 \left|\phi(u) - \mu(u)\right|^{2+\delta} du.$$

$$\left(\sum_{j=1}^{v} (\overline{\omega}_j - \underline{\omega}_j)^2\right)^{(2+\delta)/2} = \left(v^{-2\alpha} \sum_{j=1}^{v} \left(\phi(j/v) - \mu(j/v)\right)^2\right)^{(2+\delta)/2}$$

$$\sim v^{1+\delta/2-\alpha(2+\delta)} \left(\int_0^1 \left(\phi(u) - \mu(u)\right)^2 du\right)^{(2+\delta)/2}.$$

Thus

$$\frac{\sum_{j=1}^{v_n} |\overline{\omega}_j - \underline{\omega}_j|^{2+\delta}}{\left(\sum_{j=1}^{v_n} (\overline{\omega}_j - \underline{\omega}_j)^2\right)^{(2+\delta)/2}} \sim C v^{-\delta/2} \to 0,$$

as $n \to \infty$.

We next turn to the extreme logarithmic eigenvalues (3.3). In this setting, the largest logarithmic eigenvalue of Σ^* *satisfies*

$$\omega^*_{(1)} = v^{-\alpha} \sum_{j=1}^{v} \phi(j/v) = v^{1-\alpha} \frac{1}{v} \sum_{j=1}^{v} \phi(j/v) \sim v^{1-\alpha} \int \phi(u) du,$$

which tends to infinity if $\int \phi(u) du > 0$ *and* $\alpha < 1$. *It follows that*

$$\overline{\lambda}_1 \times \cdots \times \overline{\lambda}_v = \exp \omega^*_{(1)} \sim \exp(C v^{1-\alpha}) \to \infty.$$

This says that the class of eigenstructures generated by the Kronecker parameterization can be quite general, and is determined by the two parameters $\sum_{j=1}^{v_n} \mathbb{E}\zeta_j$ and $\sum_{j=1}^{v_n} \mathrm{var}(\zeta_j)$. We discuss estimation and asymptotic properties of these two parameters based on our Kronecker product structures in Sections 5.3.1 and 6.3.1, respectively. We also examine estimation and asymptotic properties of the extreme logarithmic eigenvalues (as in (3.3)) in Sections 5.3.2 and 6.3.2, respectively.

In fact, the log-normal law appears to be quite a good approximation for financial data.[3] Figure 1 shows the kernel density estimate of the 441 log eigenvalues of the sample covariance matrix of daily stock return data calculated over a ten-year period in comparison with normal density with the same mean and variance. It seems that this approximation is quite good.

3.3 Portfolio Choices and Other Applications

In this section we consider some practical motivation for considering the Kronecker factorization. Many portfolio choice methods require the inverse of the covariance matrix, Σ^{-1}. For example, the weights of the minimum variance portfolio are given by

$$w_{MV} = \frac{\Sigma^{-1}\iota_n}{\iota_n^\intercal \Sigma^{-1} \iota_n},$$

[3]Log-normal laws are widely found in social sciences, following Gibrat (1931).

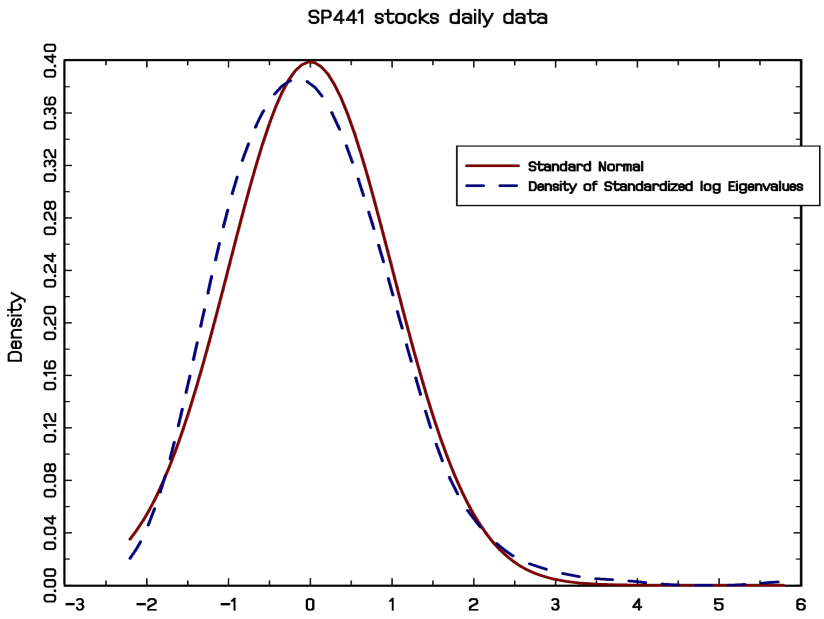

Figure 1: *The estimated density function (Silverman's kernel density estimate) of the 441 log eigenvalues of the sample covariance matrix of daily stock return data calculated over a ten-year period in comparison with normal density with the same mean and variance.*

where $\iota_n = (1, 1, \ldots, 1)^\intercal$, see e.g., Ledoit and Wolf (2003) and Chan, Karceski, and Lakonishok (1999). In the Kronecker structure case, the inverse of the covariance matrix is easily found by inverting the lower order submatrices Σ_j, which can be done analytically, since

$$\Sigma^{-1} = \Sigma_1^{-1} \otimes \Sigma_2^{-1} \otimes \cdots \otimes \Sigma_v^{-1}.$$

In fact, because $\iota_n = \iota_{n_1} \otimes \iota_{n_2} \otimes \cdots \otimes \iota_{n_v}$, we can write

$$w_{MV} = \frac{\left(\Sigma_1^{-1} \otimes \Sigma_2^{-1} \otimes \cdots \otimes \Sigma_v^{-1}\right) \iota_n}{\iota_n^\intercal \left(\Sigma_1^{-1} \otimes \Sigma_2^{-1} \otimes \cdots \otimes \Sigma_v^{-1}\right) \iota_n} = \frac{\Sigma_1^{-1}\iota_{n_1}}{\iota_{n_1}^\intercal \Sigma_1^{-1}\iota_{n_1}} \otimes \frac{\Sigma_2^{-1}\iota_{n_2}}{\iota_{n_2}^\intercal \Sigma_2^{-1}\iota_{n_2}} \otimes \cdots \otimes \frac{\Sigma_v^{-1}\iota_{n_v}}{\iota_{n_v}^\intercal \Sigma_v^{-1}\iota_{n_v}},$$

$$\operatorname{var}(w_{MV}^\intercal x_t) = \frac{1}{\iota_{n_1}^\intercal \Sigma_1^{-1}\iota_{n_1} \times \cdots \times \iota_{n_v}^\intercal \Sigma_v^{-1}\iota_{n_v}}.$$

In cases where n is large, this structure is very convenient computationally. We shall investigate this below in Sections 8 and 9. In Sections 5.3.3 and 6.3.3, we also briefly look at estimation and the asymptotic properties, respectively, of the following special case

$$\operatorname{var}(w_{MV}^\intercal x_t) = \frac{1}{\iota_2^\intercal [\Theta_1^0]^{-1}\iota_2 \times \cdots \times \iota_2^\intercal [\Theta_v^0]^{-1}\iota_2},$$

where we assume $\Theta = \Theta^0$.

Another context where the Kronecker product covariance model might be useful is in regression models. For example, suppose that

$$y = X\beta + \varepsilon,$$

where the error has covariance matrix Σ and interest centers on estimation of the parameter β. The GLS estimator in this case is

$$\begin{aligned}\hat{\beta} &= (X^\intercal \Sigma^{-1} X)^{-1} X^\intercal \Sigma^{-1} y \\ &= \left(X^\intercal \left(\Sigma_1^{-1} \otimes \Sigma_2^{-1} \otimes \cdots \otimes \Sigma_v^{-1}\right) X\right)^{-1} X^\intercal \left(\Sigma_1^{-1} \otimes \Sigma_2^{-1} \otimes \cdots \otimes \Sigma_v^{-1}\right) y,\end{aligned}$$

and our work below shows how one would obtain feasible versions of this procedure. Amemiya (1983) and other authors have shown how one can obtain efficiency gains by a feasible GLS procedure even in the case where the covariance matrix model is not correct.

4 Identification

In this section we derive a linear relationship between the logarithmic Kronecker product correlation matrix and the vector of free parameters, which delivers identification of these parameters. Let $\rho^0 := (\rho_1^0, \ldots, \rho_v^0)^\intercal \in \mathbb{R}^v$. Recall that Ω_1^0 in (2.5) has two distinct parameters a_1 and b_1. We use a similar notation for $\Omega_2^0, \ldots, \Omega_v^0$. Define $\theta^\dagger := (a_1, b_1, a_2, b_2, \ldots, a_v, b_v)^\intercal \in \mathbb{R}^{2v}$. Note that

$$\text{vech}\Omega_1^0 = \text{vech}\begin{pmatrix} a_1 & b_1 \\ b_1 & a_1 \end{pmatrix} = \begin{pmatrix} a_1 \\ b_1 \\ a_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a_1 \\ b_1 \end{pmatrix}.$$

The same principle applies to $\Omega_2^0, \ldots, \Omega_v^0$. By (2.4) and Proposition 5 in Appendix A, we have

$$\begin{aligned} \text{vech}(\Omega^0) &= \begin{bmatrix} E_1 & E_2 & \cdots & E_v \end{bmatrix} \begin{bmatrix} \text{vech}(\Omega_1^0) \\ \text{vech}(\Omega_2^0) \\ \vdots \\ \text{vech}(\Omega_v^0) \end{bmatrix} \\ &= \begin{bmatrix} E_1 & E_2 & \cdots & E_v \end{bmatrix} \left[I_v \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \right] \begin{bmatrix} a_1 \\ b_1 \\ a_2 \\ b_2 \\ \vdots \\ a_v \\ b_v \end{bmatrix} =: E_* \theta^\dagger, \end{aligned} \tag{4.1}$$

where E_i for $i = 1, \ldots, v$ are defined in (11.1). We next give an example to illustrate the form that E_* takes.

Example 2. *Suppose* $v = 3$.

$$\Omega_1^0 = \log \Theta_1^0 = \begin{pmatrix} a_1 & b_1 \\ b_1 & a_1 \end{pmatrix}, \quad \Omega_2^0 = \log \Theta_2^0 = \begin{pmatrix} a_2 & b_2 \\ b_2 & a_2 \end{pmatrix}, \quad \Omega_3^0 = \log \Theta_3^0 = \begin{pmatrix} a_3 & b_3 \\ b_3 & a_3 \end{pmatrix}.$$

Now

$$vech(\Omega^0) = vech(\Omega_1^0 \otimes I_2 \otimes I_2 + I_2 \otimes \Omega_2^0 \otimes I_2 + I_2 \otimes I_2 \otimes \Omega_3^0)$$

$$= vech \begin{pmatrix} \sum_{i=1}^3 a_i & b_3 & b_2 & 0 & b_1 & 0 & 0 & 0 \\ b_3 & \sum_{i=1}^3 a_i & 0 & b_2 & 0 & b_1 & 0 & 0 \\ b_2 & 0 & \sum_{i=1}^3 a_i & b_3 & 0 & 0 & b_1 & 0 \\ 0 & b_2 & b_3 & \sum_{i=1}^3 a_i & 0 & 0 & 0 & b_1 \\ b_1 & 0 & 0 & 0 & \sum_{i=1}^3 a_i & b_3 & b_2 & 0 \\ 0 & b_1 & 0 & 0 & b_3 & \sum_{i=1}^3 a_i & 0 & b_2 \\ 0 & 0 & b_1 & 0 & b_2 & 0 & \sum_{i=1}^3 a_i & b_3 \\ 0 & 0 & 0 & b_1 & 0 & b_2 & b_3 & \sum_{i=1}^3 a_i \end{pmatrix}$$

$$=: E_* \begin{pmatrix} a_1 \\ b_1 \\ a_2 \\ b_2 \\ a_3 \\ b_3 \end{pmatrix}$$

We can show that $E_*^\intercal E_*$ *is a* 6×6 *matrix*

$$E_*^\intercal E_* = \begin{pmatrix} 8 & 0 & 8 & 0 & 8 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 \\ 8 & 0 & 8 & 0 & 8 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 \\ 8 & 0 & 8 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{pmatrix}$$

Take Example 2 as an illustration. We can make the following observations:

(i) Each parameter in $\theta^\dagger$, e.g., $a_1, b_1, a_2, b_2, a_3, b_3$, appears exactly $n = 2^v = 8$ times in Ω^0. However in vech(Ω^0) because of the "diagonal truncation", each of a_1, a_2, a_3 appears $n = 2^v = 8$ times while each of b_1, b_2, b_3 only appears $n/2 = 4$ times.

(ii) In $E_*^\intercal E_*$, the diagonal entries summarize the information in (i). The off-diagonal entry of $E_*^\intercal E_*$ records how many times the pair to which the diagonal entry corresponds has appeared together as summands in an entry of vech(Ω^0).

(iii) The main diagonals of Ω^0 are of the form $\sum_{i=1}^3 a_i$. The rest of non-zero entries are b_1, b_2 and b_3, which are the Fisher's z-transformation of some ρ_i^0. The total number of zeros in Ω^0 is: $n(n-v-1) = 32$. Every column or row of Ω^0 has exactly $n - v - 1 = 4$ zeros.

(iv) The rank $E_*^\intercal E_*$ is $v + 1 = 4$. To see this, we left multiply $E_*^\intercal E_*$ by the $2v \times 2v$

permutation matrix

$$P := \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and right multiply $E_*^\intercal E_*$ by $P^\intercal$:

$$P(E_*^\intercal E_*)P^\intercal = \begin{pmatrix} 8 & 8 & 8 & 0 & 0 & 0 \\ 8 & 8 & 8 & 0 & 0 & 0 \\ 8 & 8 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

Note that rank is unchanged upon left or right multiplication by a nonsingular matrix. We hence deduce that $\text{rank}(E_*^\intercal E_*) = \text{rank}(E_*) = v + 1 = 4$.

(v) The eigenvalues of $E_*^\intercal E_*$ are

$$\left(0, 0, \frac{n}{2}, \frac{n}{2}, \frac{n}{2}, vn\right) = (0, 0, 4, 4, 4, 24).$$

To see this, we first recognize that $E_*^\intercal E_*$ and $P(E_*^\intercal E_*)P^\intercal$ have the same eigenvalues because P is orthogonal. The eigenvalues $P(E_*^\intercal E_*)P^\intercal$ are the eigenvalues of its blocks.

We summarize these observations in the following proposition

Proposition 1. *Recall that $n = 2^v$.*

(i) The $2v \times n(n+1)/2$ dimensional matrix $E_*^\intercal$ is sparse. $E_*^\intercal$ has $n = 2^v$ ones in odd rows and $n/2$ ones in even rows; the rest of entries are zeros.

(ii) In $E_*^\intercal E_*$, the ith diagonal entry records how many times the ith parameter of $\theta^\dagger$ has appeared in vech(Ω^0). The (i,j)th off-diagonal entry of $E_*^\intercal E_*$ records how many times the pair $(\theta_i^\dagger, \theta_j^\dagger)$ has appeared together as summands in an entry of vech(Ω^0).

(iii) The main diagonals of Ω^0 are of the form $\sum_{i=1}^v a_i$. The rest of non-zero entries are $\{b_i\}_{i=1}^v$, which are the Fisher's z-transformation of some ρ_i^0. The total number of zeros in Ω^0 is $n(n-v-1)$. Every column or row has exactly $n-v-1$ zeros.

(iv) rank$(E_*^\intercal E_*)$ = rank$(E_*^\intercal)$ = rank$(E_*) = v + 1$.

(v) The $2v$ eigenvalues of $E_*^\intercal E_*$ are

$$\left(\underbrace{0, \ldots, 0}_{v-1}, \underbrace{\frac{n}{2}, \ldots, \frac{n}{2}}_{v}, vn\right).$$

Proof. See Appendix A. □

Based on Example 2, we see that the number of *effective* parameters in $\theta^\dagger$ is actually $v+1$: $b_1, b_2, \ldots, b_v, \sum_{i=1}^v a_i$. That is, we cannot separately identify $a_1, a_2, \ldots, a_v$ as they always appear together. That is why the rank of E_* is only $v+1$ and $E_*^\intercal E_*$ has $v-1$ zero eigenvalues. It is possible to re-parametrise

$$\operatorname{vech}(\log \Theta^0) = \operatorname{vech}(\Omega^0) = E_* \theta^\dagger = E\theta^0, \tag{4.2}$$

where $\theta^0 := (\sum_{i=1}^v a_i, b_1, \ldots, b_v)^\intercal$ and E is the $n(n+1)/2 \times (v+1)$ submatrix of E_* after deleting the duplicate columns. (4.2) says that $\operatorname{vech}(\Omega^0)$ is linear in θ^0 and more generally vech of $\log \Theta^*$, not necessarily the one closest to $\log \Theta$, is linear in its parameters θ^*. We will use the relationship (4.2) in Section 5.2 to define a closed form estimator of the parameters θ^0. We also have the following proposition.

Proposition 2. *Recall that $n = 2^v$.*

(i) $\operatorname{rank}(E^\intercal E) = \operatorname{rank}(E^\intercal) = \operatorname{rank}(E)$ is $v+1$.

(ii) $E^\intercal E$ is a diagonal matrix

$$E^\intercal E = \begin{pmatrix} n & 0 \\ 0 & \frac{n}{2} I_v \end{pmatrix}.$$

(iii) The $v+1$ eigenvalues of $E^\intercal E$ are

$$\left(\underbrace{\frac{n}{2}, \ldots, \frac{n}{2}}_{v}, n \right).$$

Proof. Follows trivially from Proposition 1. □

Finally note that the dimension of θ^0 is $v+1$ whereas that of ρ^0 is v. Hence we have over-identification in the sense that any v parameters in θ^0 could be used to recover ρ^0. For instance, when $v = 2$ we have the following three equations:

$$\begin{aligned} \frac{1}{2}\log(1 - [\rho_1^0]^2) + \frac{1}{2}\log(1 - [\rho_2^0]^2) = \theta_1^0 =: a_1 + a_2 \\ \frac{1}{2}\log\left(\frac{1+\rho_1^0}{1-\rho_1^0}\right) = \theta_2^0 =: b_1 \\ \frac{1}{2}\log\left(\frac{1+\rho_2^0}{1-\rho_2^0}\right) = \theta_3^0 =: b_2. \end{aligned}$$

Any two of the preceding three equations allow us to recover ρ^0. In particular, ρ^0 and θ^0 are related by

$$\rho_j^0 = \frac{e^{2\theta_{j+1}^0} - 1}{e^{2\theta_{j+1}^0} + 1}, \qquad j = 1, 2. \tag{4.3}$$

However, it is advisable to keep all equations as they shed light on how to estimate $\sum_{j=1}^v \mathbb{E}\zeta_j$ and $\sum_{j=1}^v \operatorname{var}(\zeta_j)$.

5 Estimation

We now discuss estimation of the parameters of the Kronecker product correlation matrix Θ^0. Suppose that the setting in Section 2.3 holds. We observe a sample of n-dimensional random vectors x_t, $t = 1, 2, \ldots, T$, which are i.i.d. distributed with mean μ and a positive definite $n \times n$ covariance matrix

$$\Sigma = D^{1/2}\Theta D^{1/2}.$$

In this section, we want to estimate $\rho_1^0, \ldots, \rho_v^0$ in Θ^0 in (2.3) in the case where $n, T \to \infty$ simultaneously, i.e., *joint asymptotics* (see Phillips and Moon (1999)). We achieve dimension reduction because originally Θ has $n(n-1)/2$ free parameters whereas Θ^0 has only $v = O(\log n)$ free parameters.

To study the theoretical properties of our model, we assume that μ is *known*. We also assume that D is *known*. If $D = I_n$ this would impose no additional restriction, but in the case where $D :=\text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, this does impose a restriction. In that case, jointly estimating the n elements of D along with the parameters θ^0 of Θ^0 is not problematic computationally, but the theoretical analysis in this case is considerably more difficult. Not only will estimation of D affect the information bound for θ^0, but it also has a non-trivial impact on the derivation of the asymptotic distribution of the minimum distance estimator $\hat{\theta}_T$ due to its growing dimension n. On the other hand the properties of standard estimates of D are well known in the large n case. We therefore focus our analysis on the parameters θ^0 of the Kronecker product structure.

In Section 5.3.1 we discuss how to estimate $\sum_{j=1}^v \mathbb{E}\zeta_j$ and $\sum_{j=1}^v \text{var}(\zeta_j)$, the mean and variance parameters of the log normal distribution whose cumulative distribution function represents the spectral distribution of Θ^0. In Section 5.3.2, we try to estimate the extreme logarithmic eigenvalues of Θ^0. In Section 5.3.3, we look at the aspect of estimating $\text{var}(w_{MV}^\intercal x_t)$, the variance of the minimum variance portfolio formed using x_t, whose variance (correlation) does have a Kronecker product structure:

$$\text{var}[x_t] = \Theta^0 = \Theta_1^0 \otimes \Theta_2^0 \otimes \cdots \otimes \Theta_v^0.$$

5.1 The Quasi-Maximum Likelihood Estimator

The Gaussian QMLE is a natural starting point for estimation here. The log likelihood function for a sample $\{x_1, x_2, \ldots, x_T\}$ is given by

$$\ell_T(\rho) = -\frac{Tn}{2}\log(2\pi) - \frac{T}{2}\log\left|D^{1/2}\Theta(\rho)D^{1/2}\right| - \frac{1}{2}\sum_{t=1}^{T}(x_t-\mu)^\intercal D^{-1/2}[\Theta(\rho)]^{-1}D^{-1/2}(x_t-\mu).$$

Note that although Θ is an $n \times n$ correlation matrix, because of the Kronecker product structure, we can compute the likelihood itself very efficiently using

$$\begin{aligned}\Theta^{-1} &= \Theta_1^{-1} \otimes \Theta_2^{-1} \otimes \cdots \otimes \Theta_v^{-1} \\ |\Theta| &= |\Theta_1| \times |\Theta_2| \times \cdots \times |\Theta_v|.\end{aligned}$$

We let

$$\hat{\rho}_{QMLE} = \arg \max_{\rho \in [-1,1]^v} \ell_T(\rho).$$

Note that for a fixed v, the parameter space of ρ is compact. Writing $\Theta = \exp(\Omega)$ (as in (2.4)) and substituting this into the log likelihood function, we have

$$\ell_T(\theta) = \\ -\frac{Tn}{2}\log(2\pi) - \frac{T}{2}\log\left|D^{1/2}\exp(\Omega(\theta))D^{1/2}\right| - \frac{1}{2}\sum_{t=1}^{T}(x_t-\mu)^\intercal D^{-1/2}[\exp(\Omega(\theta))]^{-1}D^{-1/2}(x_t-\mu), \tag{5.1}$$

where the parametrization of Ω in terms of θ is similar to (4.2). We may define

$$\hat{\theta}_{QMLE} = \arg\max_{\theta}\ell_T(\theta),$$

and use the invariance principle of maximum likelihood to recover $\hat{\rho}_{QMLE}$ from $\hat{\theta}_{QMLE}$.

To compute the QMLE we use an iterative algorithm based on the derivatives of ℓ_T with respect to either ρ or θ. We give below formulas for the derivatives with respect to θ. The computations required are typically not too onerous, since for example the Hessian matrix is $(v+1)\times(v+1)$ (i.e., of order $\log n$ by $\log n$), but there is quite complicated non-linearity involved in the definition of the QMLE and so it is not so easy to analyse from a theoretical point of view. See Singull et al. (2012) and Ohlson et al. (2013) for discussion of estimation algorithms in the case where the data are multi-array and v is of low dimension.

In Section 5.2 we define a minimum distance estimator that can be analysed simply, i.e., we can obtain its large sample properties (as $n, T \to \infty$). In Section 6.2 we will consider a one-step estimator that uses the minimum distance estimator to provide a starting value and then takes a Newton-Raphson step towards the QMLE. In finite dimensional cases it is known that the one-step estimator is equivalent to the QMLE in the sense that it shares its large sample distribution (Bickel (1975)).

5.2 The Minimum Distance Estimator

Define the sample second moment matrix

$$M_T := D^{-1/2}\left[\frac{1}{T}\sum_{t=1}^{T}(x_t-\mu)(x_t-\mu)^\intercal\right]D^{-1/2} =: D^{-1/2}\tilde{\Sigma}D^{-1/2}. \tag{5.2}$$

Let W be a positive definite $n(n+1)/2 \times n(n+1)/2$ matrix and define the minimum distance (MD) estimator

$$\hat{\theta}_T(W) := \arg\min_{b\in\mathbb{R}^{v+1}}[\mathrm{vech}(\log M_T) - Eb]^\intercal W[\mathrm{vech}(\log M_T) - Eb], \tag{5.3}$$

where the matrix E is defined in (4.2). This has a closed form solution

$$\hat{\theta}_T = \hat{\theta}_T(W) = (E^\intercal W E)^{-1}E^\intercal W\,\mathrm{vech}(\log M_T). \tag{5.4}$$

Its corresponding population quantity, denoted $\theta^0(W)$, is defined

$$\theta^0(W) := \arg\min_{b\in\mathbb{R}^{v+1}}[\mathrm{vech}(\log\Theta) - Eb]^\intercal W[\mathrm{vech}(\log\Theta) - Eb],$$

whence we can solve

$$\theta^0 = \theta^0(W) = (E^\intercal W E)^{-1} E^\intercal W \operatorname{vech}(\log \Theta). \tag{5.5}$$

Note that θ^0 in (5.5) is indeed the θ^0 in (4.2) because Θ^0 is, by definition, the unique matrix minimising $\| \log \Theta - \log \Theta^* \|_W$ among all $\log \Theta^*$. To write this explicitly,

$$\| \log \Theta - \log \Theta^* \|_W = [\operatorname{vech}(\log \Theta) - Eb]^\intercal W [\operatorname{vech}(\log \Theta) - Eb]$$

which is exactly the population objective function. θ^0 is the quantity which one should expect $\hat{\theta}_T$ to converge to in some probabilistic sense regardless of whether the correlation matrix Θ has the Kronecker product structure Θ^0 or not. When Θ does have a Kronecker product structure, i.e., there exists a θ_0 such that $\operatorname{vech}(\log \Theta) = E\theta_0$, we have

$$\theta^0 = (E^\intercal W E)^{-1} E^\intercal W \operatorname{vech}(\log \Theta) = (E^\intercal W E)^{-1} E^\intercal W E \theta_0 = \theta_0.$$

In this case, $\hat{\theta}_T$ is indeed estimating the correlation matrix Θ. In Section 6.4, we also give an over-identification test based on the MD objective function in (5.3).

5.3 Estimation of Non-linear Functions of θ^0

5.3.1 Estimation of $\sum_{j=1}^v \mathbb{E}\zeta_j$ and $\sum_{j=1}^v \mathbf{var}(\zeta_j)$

In this subsection, we discuss estimation of $\sum_{j=1}^v \mathbb{E}\zeta_j$ and $\sum_{j=1}^v \operatorname{var}(\zeta_j)$, the mean and variance parameters of the log normal distribution whose cumulative distribution function represents the spectral distribution of

$$\Theta^0 = \Theta_1^0 \otimes \Theta_2^0 \otimes \cdots \otimes \Theta_v^0 = \begin{bmatrix} 1 & \rho_1^0 \\ \rho_1^0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho_2^0 \\ \rho_2^0 & 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 & \rho_v^0 \\ \rho_v^0 & 1 \end{bmatrix}.$$

Note that

$$\sum_{j=1}^v \mathbb{E}\zeta_j = \sum_{j=1}^v \left[\frac{1}{2} \log(1 + \rho_j^0) + \frac{1}{2} \log(1 - \rho_j^0) \right] = \sum_{j=1}^v \frac{1}{2} \log(1 - [\rho_j^0]^2) = \theta_1^0,$$

where the first equality is because the eigenvalues of Θ_j^0 are $1 + \rho_j^0$ and $1 - \rho_j^0$ for $j = 1, \ldots, v$, and the last equality is due to the display above (4.3). Thus estimation of $\sum_{j=1}^{v_n} \mathbb{E}\zeta_j$ is trivial because θ_1^0 is just the first component of the $v + 1$ dimensional θ^0.

Now we consider $\sum_{j=1}^v \operatorname{var}(\zeta_j)$. Note that

$$\begin{aligned} &\sum_{j=1}^v \operatorname{var}(\zeta_j) = \sum_{j=1}^v \left[\mathbb{E}\zeta_j^2 - (\mathbb{E}\zeta_j)^2 \right] = \sum_{j=1}^v \frac{1}{4} \left[\log(1 + \rho_j^0) - \log(1 - \rho_j^0) \right]^2 \\ &= \sum_{j=1}^v \left[\frac{1}{2} \log \left(\frac{1 + \rho_j^0}{1 - \rho_j^0} \right) \right]^2 = \sum_{j=1}^v (\theta_{j+1}^0)^2, \end{aligned}$$

where the last equality is due to the display above (4.3). Estimation of $\sum_{j=1}^v \operatorname{var}(\zeta_j)$ is also manageable since it is a quadratic function of θ^0. We propose to estimate these quantities by the plug-in principle using $\hat{\theta}_T$ or $\hat{\theta}_{QMLE}$.

5.3.2 Estimation of Extreme Logarithmic Eigenvalues

Let $\omega^*_{(1)}$ and $\omega^*_{(n)}$ denote the largest and smallest logarithmic eigenvalues of Θ^0, respectively. For simplicity, assume $\rho^0_j \geq 0$ for $j = 1, \ldots, v$ (otherwise we need to add absolute values). Then it is easy to calculate that

$$\begin{aligned}\omega^*_{(1)} &= \sum_{j=1}^{v} \log(1+\rho^0_j) = \sum_{j=1}^{v} \log\left(\frac{2e^{2\theta^0_{j+1}}}{e^{2\theta^0_{j+1}}+1}\right) = \sum_{j=1}^{v}\left[\log 2 + 2\theta^0_{j+1} - \log\left(e^{2\theta^0_{j+1}}+1\right)\right] \\ &=: \sum_{j=1}^{v} f_1(\theta^0_{j+1}),\end{aligned}$$

where the second equality is due to (4.3). Similarly, we can calculate

$$\begin{aligned}\omega^*_{(n)} &= \sum_{j=1}^{v} \log(1-\rho^0_j) = \sum_{j=1}^{v} \log\left(\frac{2}{e^{2\theta^0_{j+1}}+1}\right) = \sum_{j=1}^{v}\left[\log 2 - \log\left(e^{2\theta^0_{j+1}}+1\right)\right] \\ &=: \sum_{j=1}^{v} f_2(\theta^0_{j+1}).\end{aligned}$$

Thus we see that $\omega^*_{(1)}$ and $\omega^*_{(n)}$ are non-linear functions of θ^0. We propose to estimate these quantities by the plug-in principle using $\hat{\theta}_T$ or $\hat{\theta}_{QMLE}$.

Note that when $\rho^0_j > 0$ for $j = 1, \ldots, v$,

$$\omega^*_{(1)} = \sum_{j=1}^{v} \log(1+\rho^0_j) \geq Cv,$$

for some positive constant C; the right-hand side of the preceding inequality tends to infinity at a rate v. This corresponds to the case $\alpha = 0$ in Example 1.

5.3.3 Estimation of $\mathbf{var}(w^\intercal_{MV} x_t)$

Recall that under correct specification (i.e, $\Theta = \Theta^0$ or $\theta_0 = \theta^0$)

$$\mathrm{var}(w^\intercal_{MV} x_t) = \frac{1}{\iota_2^\intercal [\Theta^0_1]^{-1}\iota_2 \times \cdots \times \iota_2^\intercal [\Theta^0_v]^{-1}\iota_2}.$$

First note that for $j = 1, \ldots, v$,

$$[\Theta^0_j]^{-1} = \begin{bmatrix} 1 & -\rho^0_j \\ -\rho^0_j & 1 \end{bmatrix} \frac{1}{1-[\rho^0_j]^2}, \qquad \iota_2^\intercal [\Theta^0_j]^{-1}\iota_2 = \frac{2}{1+\rho^0_j}.$$

Hence

$$\log \mathrm{var}(w^\intercal_{MV} x_t) = \sum_{j=1}^{v} -\log\left(\frac{2}{1+\rho^0_j}\right) = \sum_{j=1}^{v} -\log\left(1+e^{-2\theta^0_{j+1}}\right) =: \sum_{j=1}^{v} f_3(\theta^0_{j+1}),$$

where the second equality is due to (4.3). Thus we see that $\log \mathrm{var}(w^\intercal_{MV} x_t)$ is a non-linear function of θ^0. We propose to estimate these quantities by the plug-in principle using $\hat{\theta}_T$ or $\hat{\theta}_{QMLE}$.

6 Asymptotic Properties

In this section, we first derive the asymptotic properties of the two estimators, the minimum distance estimator $\hat{\theta}_T$ and the one-step QMLE $\tilde{\theta}_T$ which we define in Section 6.2. We consider the case where $n, T \to \infty$ simultaneously. In some results we assume that the Gaussian likelihood is correctly specified both with respect of the distribution and the covariance structure. In this case we expect that $\hat{\theta}_{QMLE}$ converges in probability to θ^0, where θ^0 is defined in (4.2) or (5.5). If the likelihood is not correctly specified, $\hat{\theta}_{QMLE}$ will converge in probability to the parameter of a Kronecker product structure which has a density closest to the density of the data generating process in terms of Kullback-Leibler divergence. Because of the special choice of the Gaussian likelihood, this parameter can be shown to coincide with θ^0, the value defined in Section 3.1. However, in general the asymptotic variance of $\hat{\theta}_{QMLE}$ will then have a sandwich form (see for instance van der Vaart (2010) Example 13.7). Our first main result (Theorem 1) establishes the rate of convergence of $\hat{\theta}_T$ around the limiting value θ^0 in the general setting where neither Gaussianity nor the Kronecker product structure is true. In Theorem 2 we derive the feasible CLT for $\hat{\theta}_T$ in the same case. We then establish the properties of the approximate QMLE in the Gaussian case. Then we work out the asymptotic properties of the estimators of $\sum_{j=1}^{v} \mathbb{E}\zeta_j$ and $\sum_{j=1}^{v} \mathrm{var}(\zeta_j)$, the mean and variance parameters of the log normal distribution whose cumulative distribution function represents the spectral distribution of Θ^0. Next, we provide the asymptotic properties of the estimators of the extreme logarithmic eigenvalues $\omega^*_{(1)}$ and $\omega^*_{(n)}$ defined in Section 5.3.2. We also give the asymptotic properties of the estimator of $\log \mathrm{var}(w^{\intercal}_{MV} x_t)$, the logarithm of the variance of the minimum variance portfolio formed using x_t, whose variance (correlation) matrix has a Kronecker product structure. Last, we formulate an over-identification test to allow us to test whether a correlation matrix has a Kronecker product structure.

6.1 The MD Estimator

6.1.1 Rate of Convergence

The following proposition linearizes the matrix logarithm.

Proposition 3. *Suppose both $n \times n$ matrices $A + B$ and A are positive definite for all n with the minimum eigenvalues bounded away from zero by absolute constants. Suppose the maximum eigenvalue of A is bounded from above by an absolute constant. Further suppose*

$$\left\| [t(A - I) + I]^{-1} tB \right\|_{\ell_2} \leq C < 1 \tag{6.1}$$

for all $t \in [0, 1]$ and some constant C. Then

$$\log(A + B) - \log A = \int_0^1 [t(A - I) + I]^{-1} B [t(A - I) + I]^{-1} dt + O(\|B\|^2_{\ell_2} \vee \|B\|^3_{\ell_2}).$$

Proof. See Appendix A. □

The conditions of the preceding proposition implies that for every $t \in [0, 1]$, $t(A-I)+I$ is positive definite for all n with the minimum eigenvalue bounded away from zero by an absolute constant (Horn and Johnson (1985) p181). Proposition 3 has a flavour of Frechet derivative because $\int_0^1 [t(A-I)+I]^{-1} B [t(A-I)+I]^{-1} dt$ is the Frechet derivative of

matrix logarithm at A in the direction B (Higham (2008) p272); however, this proposition is slightly stronger in the sense of a sharper bound on the remainder.

Assumption 1.

(i) $\{x_t\}_{t=1}^T$ *are subgaussian random vectors. That is, for all* t*, for every* $a \in \mathbb{R}^n$*, and every* $\epsilon > 0$

$$\mathbb{P}(|a^\intercal x_t| \geq \epsilon) \leq Ke^{-C\epsilon^2},$$

for positive absolute constants K *and* C*.*

(ii) $\{x_t\}_{t=1}^T$ *are normally distributed.*

Assumption 1(i) is standard in high-dimensional theoretical work. In essence it assumes that a random vector has exponential tail probabilities, which allows us to invoke some concentration inequality such as the Bernstein's inequality in Appendix B. Concentration inequalities are useful when one wants that a whole collection of events (here indexed by n) holds simultaneously with large probability.

Note that Assumption 1(ii) implies Assumption 1(i). We would like to remark that Assumption 1(ii) is not needed for Theorem 1 or Theorem 2.

Assumption 2.

(i) $n, T \to \infty$ *simultaneously, and* $n/T \to 0$*.*

(ii) $n, T \to \infty$ *simultaneously, and*

$$\frac{n^2\kappa^2(W)}{T}\left(T^{2/\gamma}\log^2 n \vee n^2\kappa^2(W)\log^5 n^4\right) = o(1), \qquad \text{for some } \gamma > 2,$$

where $\kappa(W)$ *is the condition number of* W *for matrix inversion with respect to the spectral norm, i.e.,*

$$\kappa(W) := \|W^{-1}\|_{\ell_2}\|W\|_{\ell_2}.$$

Assumption 2(i) is for the derivation of the rate of convergence of the minimum distance estimator $\hat{\theta}_T$ (Theorem 1). Assumption 2(ii) is *sufficient* for the asymptotic normality of $\hat{\theta}_T$ (Theorem 2). If Assumption 1(i) holds, we can choose the γ in Assumption 2(ii) arbitrarily large, so Assumption 2(ii) is roughly equivalent to $n^4\kappa^4(W)\log^5 n^4/T = o(1)$. In unreported work carried out by the authors, if one assumes normality and takes $W = I_{n(n+1)/2}$ (i.e., $\kappa(I_{n(n+1)/2}) = 1$), Assumption 2(ii) can be relaxed to

$$\frac{n^2}{T}\left(T^{2/\gamma}\log^2 n \vee n\right) = o(1), \qquad \text{for some } \gamma > 2.$$

Assumption 3.

(i) Recall that $D := diag(\sigma_1^2, \ldots, \sigma_n^2)$*, where* $\sigma_i^2 := \mathbb{E}(x_{t,i} - \mu_i)^2$*. Suppose* $\min_{1\leq i\leq n}\sigma_i^2$ *is bounded away from zero by an absolute constant.*

(ii) Recall that $\Sigma := \mathbb{E}(x_t - \mu)(x_t - \mu)^\intercal$*. Suppose its maximum eigenvalue bounded from above by an absolute constant.*

(iii) Suppose that Σ *is positive definite for all* n *with its minimum eigenvalue bounded away from zero by an absolute constant.*

(iv) $\max_{1\leq i\leq n}\sigma_i^2$ *is bounded from above by an absolute constant.*

We assume that $\min_{1\leq i\leq n}\sigma_i^2$ is bounded away from zero by an absolute constant in Assumption 3(i) otherwise $D^{-1/2}$ is not defined in the limit $n\to\infty$. Assumption 3(ii) is fairly standard in the high-dimensional literature. The assumption of positive definiteness of the covariance matrix Σ in Assumption 3(iii) is also standard, and, together with Assumption 3(iv), ensures that the correlation matrix $\Theta := D^{-1/2}\Sigma D^{-1/2}$ is positive definite for all n with its minimum eigenvalue bounded away from zero by an absolute constant via Observation 7.1.6 in Horn and Johnson (1985) p399. Similarly, Assumptions 3(i)-(ii) ensure that Θ has maximum eigenvalue bounded away from above by an absolute constant. To summarise, Assumption 3 ensures that Θ is well behaved; in particular, $\log\Theta$ is properly defined.

The following proposition is a stepping stone for the main results of this paper.

Proposition 4. *Suppose Assumptions 1(i), 2(i), and 3 hold. We have:*

(i)

$$\|M_T-\Theta\|_{\ell_2}=O_p\left(\sqrt{\frac{n}{T}}\right).$$

(ii) The bound (6.1) is satisfied with probability approaching 1 for $A=\Theta$ *and* $B=M_T-\Theta$*. That is,*

$$\|[t(\Theta-I)+I]^{-1}t(M_T-\Theta)\|_{\ell_2}\leq C<1 \qquad \textit{with probability approaching 1,}$$

for some constant C*.*

Proof. See Appendix A. □

Assumption 4. *Suppose* $M_T := D^{-1/2}\tilde{\Sigma}D^{-1/2}$ *defined in (5.2) is positive definite for all* n *with its minimum eigenvalue bounded away from zero by an absolute constant with probability approaching 1 as* $n,T\to\infty$*.*

Assumption 4 is the sample-analogue assumption as compared to Assumptions 3(iii)-(iv). In essence it ensures that $\log M_T$ is properly defined. More primitive conditions in terms of D and $\tilde{\Sigma}$ could easily be formulated to replace Assumption 4. Assumption 4, together with Proposition 4(i) ensure that the maximum eigenvalue of M_T is bounded from above by an absolute constant with probability approaching 1.

The following theorem gives the rate of convergence of the minimum distance estimator $\hat{\theta}_T$.

Theorem 1. *Let Assumptions 1(i), 2(i), 3, and 4 be satisfied. Then*

$$\|\hat{\theta}_T-\theta^0\|_2=O_p\left(\sqrt{\frac{n\kappa(W)}{T}}\right),$$

where $\hat{\theta}_T$ *and* θ^0 *are defined in (5.4) and (5.5), respectively.*

Proof. See Appendix A. □

Note that θ^0 contains the unique parameters of the Kronecker product Θ^0 which we have shown is closest to the correlation matrix Θ in some sense. The dimension of θ^0 is $v+1 = O(\log n)$ while the dimension of unique parameters of Θ is $O(n^2)$. If no structure whatsoever is imposed on covariance matrix estimation, the rate of convergence for Euclidean norm would be $(n^2/T)^{1/2}$ (square root of summing up n^2 terms each of which has a rate $1/T$). We have some rate improvement in Theorem 1 as compared to this crude rate, provided $\kappa(W)$ is not too large.

However, given the dimension of θ^0, one would conjecture that the optimal rate of convergence should be $(\log n/T)^{1/2}$. There are, perhaps, two reasons for the rate difference. First, the matrix W might not be sparse; a non-sparse W destroys the sparsity of $E^\intercal$ under multiplication. Of course in the special case $W = I_{n(n+1)/2}$, W is sparse. Second, linearisation of the matrix logarithm introduces another non-sparse matrix, the Frechet derivative, sandwiched between the sparse matrix $E^\intercal D_n^+$ (suppose $W = I_{n(n+1)/2}$ for the moment) and the vector $\mathrm{vec}(M_T - \Theta)$. Again we are unable to utilise the sparse structure of $E^\intercal$ except for the information about the eigenvalues (Proposition 2(iii)). If one makes some assumption directly on the entries of the matrix logarithm as well as imposes $W = I_{n(n+1)/2}$, we conjecture that one would achieve a better rate.

6.1.2 The Asymptotic Normality

Let H and $\hat{H}_T$ denote the $n^2 \times n^2$ matrices

$$H := \int_0^1 [t(\Theta - I) + I]^{-1} \otimes [t(\Theta - I) + I]^{-1} dt, \tag{6.2}$$

$$\hat{H}_T := \int_0^1 [t(M_T - I) + I]^{-1} \otimes [t(M_T - I) + I]^{-1} dt,$$

respectively.[4]

Note that $x \mapsto (\lceil \frac{x}{n} \rceil, x - \lfloor \frac{x}{n} \rfloor n)$ is a bijection from $\{1, \ldots, n^2\}$ to $\{1, \ldots, n\} \times \{1, \ldots, n\}$. Define the $n^2 \times n^2$ matrix

$$V := \mathrm{var}(\sqrt{T}\mathrm{vec}(\tilde{\Sigma} - \Sigma)).$$

It is easy to show that its (x, y)th entry is

$$V_{x,y} \equiv V_{i,j,k,l} =$$
$$\mathbb{E}[(x_{t,i} - \mu_i)(x_{t,j} - \mu_j)(x_{t,k} - \mu_k)(x_{t,l} - \mu_l)] - \mathbb{E}[(x_{t,i} - \mu_i)(x_{t,j} - \mu_j)]\mathbb{E}[(x_{t,k} - \mu_k)(x_{t,l} - \mu_l)],$$

where $x, y \in \{1, \ldots, n^2\}$ and $i, j, k, l \in \{1, \ldots, n\}$. Define its sample analogue $\hat{V}_T$ whose (x, y)th entry is

$$\hat{V}_{T,x,y} \equiv \hat{V}_{T,i,j,k,l} := \frac{1}{T}\sum_{t=1}^{T}(x_{t,i} - \mu_i)(x_{t,j} - \mu_j)(x_{t,k} - \mu_k)(x_{t,l} - \mu_l)$$
$$- \left(\frac{1}{T}\sum_{t=1}^{T}(x_{t,i} - \mu_i)(x_{t,j} - \mu_j)\right)\left(\frac{1}{T}\sum_{t=1}^{T}(x_{t,k} - \mu_k)(x_{t,l} - \mu_l)\right).$$

[4] In principle, both matrices depend on n as well but we suppress this subscript throughout the paper.

Finally for any $c \in \mathbb{R}^{v+1}$ define the scalar

$$\begin{aligned} G &:= c^\intercal J c \\ &:= c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H (D^{-1/2} \otimes D^{-1/2}) V (D^{-1/2} \otimes D^{-1/2}) H D_n^{+\intercal} W E (E^\intercal W E)^{-1} c. \end{aligned} \tag{6.3}$$

We also define the estimate $\hat{G}_T$:

$$\begin{aligned} \hat{G}_T &:= c^\intercal \hat{J}_T c \\ &:= c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ \hat{H}_T (D^{-1/2} \otimes D^{-1/2}) \hat{V}_T (D^{-1/2} \otimes D^{-1/2}) \hat{H}_T D_n^{+\intercal} W E (E^\intercal W E)^{-1} c. \end{aligned}$$

Assumption 5. *V is positive definite for all n, with its minimum eigenvalue bounded away from zero by an absolute constant and maximum eigenvalue bounded from above by an absolute constant.*

We remark that Assumption 5 could be relaxed to the case where the minimum (maximum) eigenvalue of V is drifting towards zero (infinity) at certain rate. The proof for Theorem 2 remains unchanged, but this rate will need to be incorporated in Assumption 2(ii).

Example 3. *In the special case of normality, $V = 2D_n D_n^+ (\Sigma \otimes \Sigma)$ (Magnus and Neudecker (1986) Lemma 9). Then G could be simplified into*

$$\begin{aligned} G = & \\ & 2c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H (D^{-1/2} \otimes D^{-1/2}) D_n D_n^+ (\Sigma \otimes \Sigma) (D^{-1/2} \otimes D^{-1/2}) H D_n^{+\intercal} W E (E^\intercal W E)^{-1} c \\ & = 2c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H (D^{-1/2} \otimes D^{-1/2}) (\Sigma \otimes \Sigma) (D^{-1/2} \otimes D^{-1/2}) H D_n^{+\intercal} W E (E^\intercal W E)^{-1} c \\ & = 2c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H (D^{-1/2} \Sigma D^{-1/2} \otimes D^{-1/2} \Sigma D^{-1/2}) H D_n^{+\intercal} W E (E^\intercal W E)^{-1} c \\ & = 2c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H (\Theta \otimes \Theta) H D_n^{+\intercal} W E (E^\intercal W E)^{-1} c, \end{aligned}$$

where the second equality is true because, given the structure of H, via Lemma 11 of Magnus and Neudecker (1986), we have the following identity:

$$D_n^+ H (D^{-1/2} \otimes D^{-1/2}) = D_n^+ H (D^{-1/2} \otimes D^{-1/2}) D_n D_n^+.$$

Note that Assumption 5 is automatically satisfied under normality given Assumption 3 (ii)-(iii).

Theorem 2. *Let Assumptions 1(i), 2(ii), 3, 4 and 5 be satisfied. Then*

$$\frac{\sqrt{T} c^\intercal (\hat{\theta}_T - \theta^0)}{\sqrt{\hat{G}_T}} \xrightarrow{d} N(0,1),$$

for any $(v+1) \times 1$ non-zero vector c with $\|c\|_2 = 1$.

Proof. See Appendix A. □

If one chooses, infeasibly,

$$W = \left[D_n^+ H(D^{-1/2} \otimes D^{-1/2})V(D^{-1/2} \otimes D^{-1/2})HD_n^{+\intercal}\right]^{-1},$$

the scalar G reduces to

$$c^\intercal \left(E^\intercal \left[D_n^+ H(D^{-1/2} \otimes D^{-1/2})V(D^{-1/2} \otimes D^{-1/2})HD_n^{+\intercal}\right]^{-1} E\right)^{-1} c.$$

Under the further assumption of normality (i.e., $V = 2D_n D_n^+(\Sigma \otimes \Sigma)$), the preceding display further simplifies to

$$c^\intercal \left(\frac{1}{2}E^\intercal D_n^\intercal H^{-1}(\Theta^{-1} \otimes \Theta^{-1})H^{-1}D_n E\right)^{-1} c,$$

by Lemma 14 of Magnus and Neudecker (1986).

We also give the following corollary which allows us to test multiple hypotheses like $H_0 : A^\intercal \theta^0 = a$.

Corollary 1. *Let Assumptions 1(i), 2(ii), 3, 4 and 5 be satisfied. Given a full-column-rank $(v+1) \times k$ matrix A where k is finite with $\|A\|_{\ell_2} = O_p(\sqrt{n\kappa(W)})$, we have*

$$\sqrt{T}(A^\intercal \hat{J}_T A)^{-1/2} A^\intercal (\hat{\theta}_T - \theta^0) \xrightarrow{d} N\left(0, I_k\right).$$

Proof. See Appendix A. □

The condition $\|A\|_{\ell_2} = O_p(\sqrt{n\kappa(W)})$ is trivial because the dimension of A is only of order $O(\log n) \times O(1)$. Moreover we can always rescale A when carrying out hypothesis testing.

6.2 An Approximate QMLE

We first define the score function and Hessian function of (5.1), which we give in the theorem below, since it is a non-trivial calculation.

Theorem 3. *The score function of the Gaussian quasi-likelihood takes the following form*

$$\frac{\partial \ell_T(\theta)}{\partial \theta^\intercal} = \frac{T}{2}E^\intercal D_n^\intercal \int_0^1 e^{t\Omega} \otimes e^{(1-t)\Omega} dt \left[vec\left([\exp(\Omega)]^{-1} D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega)]^{-1} - \left[\exp(\Omega)\right]^{-1}\right)\right],$$

where $\tilde{\Sigma}$ is defined in (5.2). The Hessian matrix takes the following form

$$\begin{aligned}\mathcal{H}(\theta) = \frac{\partial^2 \ell_T(\theta)}{\partial\theta\partial\theta^\intercal} = &-\frac{T}{2}E^\intercal D_n^\intercal \Psi_1 \left([\exp\Omega]^{-1} D^{-1/2}\tilde{\Sigma}D^{-1/2} \otimes I_n + I_n \otimes [\exp\Omega]^{-1} D^{-1/2}\tilde{\Sigma}D^{-1/2} - I_{n^2}\right) \cdot \\ &\left([\exp\Omega]^{-1} \otimes [\exp\Omega]^{-1}\right)\Psi_1 D_n E \\ &+ \frac{T}{2}(\Psi_2^\intercal \otimes E^\intercal D_n^\intercal) \int_0^1 P\left(I_{n^2} \otimes vece^{(1-t)\Omega}\right) \int_0^1 e^{st\Omega} \otimes e^{(1-s)t\Omega} ds \cdot t dt D_n E \\ &+ \frac{T}{2}(\Psi_2^\intercal \otimes E^\intercal D_n^\intercal) \int_0^1 P\left(vece^{t\Omega} \otimes I_{n^2}\right) \int_0^1 e^{s(1-t)\Omega} \otimes e^{(1-s)(1-t)\Omega} ds \cdot (1-t) dt D_n E,\end{aligned}$$

where

$$\Psi_1 = \Psi_1(\theta) := \int_0^1 e^{t\Omega(\theta)} \otimes e^{(1-t)\Omega(\theta)} dt,$$

$$\Psi_2 = \Psi_2(\theta) := vec\left([\exp \Omega(\theta)]^{-1} D^{-1/2} \tilde{\Sigma} D^{-1/2} [\exp \Omega(\theta)]^{-1} - \left[\exp \Omega(\theta)\right]^{-1} \right),$$

$$P := I_n \otimes K_{n,n} \otimes I_n.$$

Proof. See Appendix A. □

Under the assumption that a Kronecker product structure is correctly specified for the correlation matrix Θ (i.e., $D^{-1/2}\mathbb{E}\left[\tilde{\Sigma}\right] D^{-1/2} = \Theta = \Theta^0$), we have $\mathbb{E}\Psi_2(\theta^0) = 0$, so the normalized expected Hessian matrix evaluated at θ^0 takes the following form

$$\begin{aligned}\Upsilon &:= \mathbb{E}\left[\mathcal{H}(\theta^0)/T\right] = -\frac{1}{2} E^\intercal D_n^\intercal \Psi_1(\theta^0) \left([\exp \Omega(\theta^0)]^{-1} \otimes [\exp \Omega(\theta^0)]^{-1} \right) \Psi_1(\theta^0) D_n E \\ &= -\frac{1}{2} E^\intercal D_n^\intercal \Psi_1(\theta^0) \left([\Theta^0]^{-1} \otimes [\Theta^0]^{-1} \right) \Psi_1(\theta^0) D_n E.\end{aligned}$$

Therefore, define:

$$\hat{\Upsilon}_T := -\frac{1}{2} E^\intercal D_n^\intercal \hat{\Psi}_{1,T} \left(M_T^{-1} \otimes M_T^{-1} \right) \hat{\Psi}_{1,T} D_n E,$$

where

$$\hat{\Psi}_{1,T} := \int_0^1 M_T^t \otimes M_T^{1-t} dt.$$

We then propose the following one-step estimator in the spirit of van der Vaart (1998) p72 or Newey and McFadden (1994) p2150:

$$\tilde{\theta}_T := \hat{\theta}_T - \hat{\Upsilon}_T^{-1} \frac{\partial \ell_T(\hat{\theta}_T)}{\partial \theta^\intercal} / T. \tag{6.4}$$

We show in Appendix A that $\hat{\Upsilon}_T$ is invertible with probability approaching 1. We did not use the vanilla one-step estimator because the Hessian matrix is rather complicated to analyse. We next provide the large sample theory for $\tilde{\theta}_T$.

Assumption 6. *For every positive constant M and uniformly in $b \in \mathbb{R}^{v+1}$ with $\|b\|_2 = 1$,*

$$\sup_{\theta^*: \|\theta^* - \theta^0\|_2 \leq M\sqrt{n\kappa(W)/T}} \left| \sqrt{T} b^\intercal \left[\frac{1}{T} \frac{\partial \ell_T(\theta^*)}{\partial \theta^\intercal} - \frac{1}{T} \frac{\partial \ell_T(\theta^0)}{\partial \theta^\intercal} - \Upsilon(\theta^* - \theta^0) \right] \right| = o_p(1).$$

Assumption 6 is one of the sufficient conditions needed for Theorem 4. This kind of assumption is standard in the asymptotics of one-step estimators (see (5.44) of van der Vaart (1998) p71, Bickel (1975)) or of M-estimation (see (C3) of He and Shao (2000)). Roughly speaking, Assumption 6 implies that $\frac{1}{T}\frac{\partial \ell_T}{\partial \theta^\intercal}$ is differentiable at θ^0, with derivative tending to Υ in probability, but this is not an assumption. The radius of the shrinking neighbourhood $\sqrt{n\kappa(W)/T}$ is determined by the rate of convergence of any preliminary estimator, say, $\hat{\theta}_T$ in our case. The uniform requirement of the shrinking neighbourhood could be relaxed using Le Cam's *discretization* trick (see van der Vaart (1998) p72). It is possible to relax the $o_p(1)$ on the right side of Assumption 6 to $o_p(n^{1/2})$ by going through the proof of Theorem 4.

Theorem 4. *Suppose that a Kronecker product structure is correctly specified for the correlation matrix Θ. Let Assumptions 1(ii), 2(ii), 3, 4, and 6 be satisfied. Then*

$$\frac{\sqrt{T}b^\intercal(\tilde{\theta}_T - \theta^0)}{\sqrt{b^\intercal(-\hat{\Upsilon}_T)^{-1}b}} \xrightarrow{d} N(0,1)$$

for any $(v+1)\times 1$ vector b with $\|b\|_2 = 1$.

Proof. See Appendix A. □

Note that if we replace normality (Assumption 1(ii)) with the subgaussian assumption (Assumption 1(i)) - that is Gaussian likelihood is not correctly specified - although the norm consistency of $\tilde{\theta}_T$ should still hold, the asymptotic variance in Theorem 4 needs to be changed to have a sandwich formula.

Theorem 4 says that $\sqrt{T}b^\intercal(\tilde{\theta}_T - \theta^0) \xrightarrow{d} N\left(0, b^\intercal\left(-\mathbb{E}[\mathcal{H}(\theta^0)/T]\right)^{-1}b\right)$. In the finite n case, this estimator achieves the parametric efficiency bound. This shows that our one-step estimator $\tilde{\theta}_T$ is efficient when D (the variances) is known. When D is unknown, one has to differentiate (5.1) with respect to both θ and the diagonal elements of D. The analysis becomes considerably more involved and we leave it for future work.

By recognising that

$$H^{-1} = \int_0^1 e^{t\log\Theta} \otimes e^{(1-t)\log\Theta} dt,$$

(see Proposition 14 in Appendix A), we see that under Gaussianity and correct specification of the Kronecker product, $\tilde{\theta}_T$ and the optimal MD estimator have the same asymptotic variance, i.e.,

$$(-\Upsilon)^{-1} = \left(\frac{1}{2}E^\intercal D_n^\intercal H^{-1}(\Theta^{-1}\otimes\Theta^{-1})H^{-1}D_nE\right)^{-1}.$$

Likewise we have the following corollary which allows us to test multiple hypotheses like $H_0 : A^\intercal\theta^0 = a$.

Corollary 2. *Suppose that a Kronecker product structure is correctly specified for the correlation matrix Θ. Let Assumptions 1(ii), 2(ii), 3, 4, and 6 be satisfied. Given a full-column-rank $(v+1)\times k$ matrix A where k is finite with $\|A\|_{\ell_2} = O_p(\sqrt{n\kappa(W)})$, we have*

$$\sqrt{T}\left(A^\intercal(-\hat{\Upsilon}_T)^{-1}A\right)^{-1/2}A^\intercal(\tilde{\theta}_T - \theta^0) \xrightarrow{d} N\left(0, I_k\right).$$

Proof. Essentially same as that of Corollary 1. □

The condition $\|A\|_{\ell_2} = O_p(\sqrt{n\kappa(W)})$ is trivial because the dimension of A is only of order $O(\log n)\times O(1)$. Moreover we can always rescale A when carrying out hypothesis testing.

6.3 Estimators of Non-linear Functions of θ^0

6.3.1 The Estimators of $\sum_{j=1}^{v}\mathbb{E}\zeta_j$ and $\sum_{j=1}^{v}\mathbf{var}(\zeta_j)$

We have shown in Section 5.3.1 that $\sum_{j=1}^{v}\mathbb{E}\zeta_j = \theta_1^0$ and $\sum_{j=1}^{v}\mathrm{var}(\zeta_j) = \sum_{j=1}^{v}[\theta_{j+1}^0]^2$. Thus we can use either the minimum distance estimator $\hat{\theta}_T$ or the one-step estimator $\tilde{\theta}_T$ to estimate these two quantities. We give a result using $\hat{\theta}_T$; the proof for the parallel result of using $\tilde{\theta}_T$ should be roughly the same.

Theorem 5. *Let Assumptions 1(i), 2(ii), 3, 4 and 5 be satisfied. Assume $\rho_j^0 \neq 0$ for some $j \in \{1,\ldots,v\}$. Then*

(i)

$$\sqrt{T}\left(\hat{\theta}_{T,1} - \sum_{j=1}^{v}\mathbb{E}\zeta_j\right) \xrightarrow{d} N\left(0, G(e_1)\right),$$

where $G(e_1)$ is the matrix G defined in (6.3) with c evaluated at e_1, i.e., the $(v+1)$-dimensional vector with the first component being 1 and the remaining components being 0.

(ii)

$$\frac{\sqrt{T}}{\left(\sum_{j=1}^{v}\hat{\theta}_{T,j+1}^2\right)^{1/2}}\left(\sum_{j=1}^{v}\hat{\theta}_{T,j+1}^2 - \sum_{j=1}^{v}var(\zeta_j)\right) \xrightarrow{d} 2N\left(0, G(c')\right),$$

where $G(c')$ is the matrix G defined in (6.3) with c evaluated at c':

$$c_1' = 0, \qquad c_{j+1}' = \frac{\theta_{j+1}^0}{\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}}, \quad j = 1,\ldots,v.$$

Proof. See Appendix A. □

The requirement that $\rho_j^0 \neq 0$ for some $j \in \{1,\ldots,v\}$ ensures that at least one $\theta_{j+1}^0 \neq 0$ so c' is properly defined.

6.3.2 The Estimators of Extreme Logarithmic Eigenvalues

We have shown in Section 5.3.2 that when assuming $\rho_j^0 \geq 0$ for $j = 1,\ldots,v$ for simplicity, we have

$$\omega_{(1)}^* = \sum_{j=1}^{v}\log(1+\rho_j^0) = \sum_{j=1}^{v}\left[\log 2 + 2\theta_{j+1}^0 - \log\left(e^{2\theta_{j+1}^0}+1\right)\right] =: \sum_{j=1}^{v} f_1(\theta_{j+1}^0),$$

$$\omega_{(n)}^* = \sum_{j=1}^{v}\log(1-\rho_j^0) = \sum_{j=1}^{v}\left[\log 2 - \log\left(e^{2\theta_{j+1}^0}+1\right)\right] =: \sum_{j=1}^{v} f_2(\theta_{j+1}^0).$$

Again we shall, for simplicity, use the minimum distance estimator $\hat{\theta}_T$ to derive the asymptotic properties of the estimators of $\omega_{(1)}^*$ and $\omega_{(n)}^*$; a similar result should exist for the one-step estimator $\tilde{\theta}_T$.

Theorem 6. *Let Assumptions 1(i), 2(ii), 3, 4 and 5 be satisfied.*

(i) Assume at least one ρ_j^0 is bounded away from 1 by an absolute constant. Then

$$\frac{\sqrt{T}}{\left(\sum_{j=1}^{v}[f_1'(\hat{\theta}_{T,j+1})]^2\right)^{1/2}}\left(\sum_{j=1}^{v} f_1(\hat{\theta}_{T,j+1}) - \omega_{(1)}^*\right) \xrightarrow{d} N\left(0, G(c^U)\right),$$

where $G(c^U)$ is the matrix G defined in (6.3) with c evaluated at c^U:

$$c_1^U = 0, \qquad c_{j+1}^U = \frac{f_1'(\theta_{j+1}^0)}{\left(\sum_{j=1}^{v}[f_1'(\theta_{j+1}^0)]^2\right)^{1/2}}, \quad j = 1, \ldots, v.$$

(ii) Then

$$\frac{\sqrt{T}}{\left(\sum_{j=1}^{v}[f_2'(\hat{\theta}_{T,j+1})]^2\right)^{1/2}}\left(\sum_{j=1}^{v} f_2(\hat{\theta}_{T,j+1}) - \omega_{(n)}^*\right) \xrightarrow{d} N\left(0, G(c^L)\right),$$

where $G(c^L)$ is the matrix G defined in (6.3) with c evaluated at c^L:

$$c_1^L = 0, \qquad c_{j+1}^L = \frac{f_2'(\theta_{j+1}^0)}{\left(\sum_{j=1}^{v}[f_2'(\theta_{j+1}^0)]^2\right)^{1/2}}, \quad j = 1, \ldots, v.$$

Proof. See Appendix A. □

The requirement that at least one ρ_j^0 is bounded away from 1 by an absolute constant in Theorem 6(i) ensures that at least one $f_1'(\theta_{j+1}^0) > 0$ so c^U is properly defined. We do not need a similar assumption in Theorem 6(ii) because the case in Theorem 6(ii) is reversed: We need at least one ρ_j^0 is bounded away from -1 by an absolute constant, which is a weaker assumption than $\rho_j^0 \geq 0$ for all j.

6.3.3 The Estimator of $\log \operatorname{var}(w_{MV}^{\intercal} x_t)$

We have shown in Section 5.3.3 that

$$\log \operatorname{var}(w_{MV}^{\intercal} x_t) = \sum_{j=1}^{v} -\log\left(1 + e^{-2\theta_{j+1}^0}\right) =: \sum_{j=1}^{v} f_3(\theta_{j+1}^0).$$

Again we shall, for simplicity, use the minimum distance estimator $\hat{\theta}_T$ to derive the asymptotic properties of the estimator of $\log \operatorname{var}(w_{MV}^{\intercal} x_t)$; a similar result should exist for the one-step estimator $\tilde{\theta}_T$.

Theorem 7. *Let Assumptions 1(i), 2(ii), 3, 4 and 5 be satisfied. Assume at least one ρ_j^0 is bounded away from 1 by an absolute constant. Then*

$$\frac{\sqrt{T}}{\left(\sum_{j=1}^{v}[f_3'(\hat{\theta}_{T,j+1})]^2\right)^{1/2}}\left(\sum_{j=1}^{v} f_3(\hat{\theta}_{T,j+1}) - \log \operatorname{var}(w_{MV}^{\intercal} x_t)\right) \xrightarrow{d} N\left(0, G(c^*)\right),$$

where $G(c^*)$ is the matrix G defined in (6.3) with c evaluated at c^*:

$$c_1^* = 0, \qquad c_{j+1}^* = \frac{f_3'(\theta_{j+1}^0)}{\left(\sum_{j=1}^{v}[f_3'(\theta_{j+1}^0)]^2\right)^{1/2}}, \quad j = 1, \ldots, v.$$

Proof. See Appendix A. □

The requirement that at least one ρ_j^0 is bounded away from 1 by an absolute constant ensures that at least one $f_3'(\theta_{j+1}^0) > 0$ so c^* is properly defined.

6.4 An Over-Identification Test

In this section, we give an over-identification test based on the MD objective function in (5.3). Suppose we want to test whether the correlation matrix Θ has the Kronecker product structure Θ^0 defined in (2.3). That is,

$$H_0 : \Theta = \Theta^0 \quad (i.e., \theta_0 = \theta^0), \qquad H_1 : \Theta \neq \Theta^0.$$

We first *fix* n (and hence v). Recall (5.3):

$$\begin{aligned}\hat{\theta}_T = \hat{\theta}_T(W) &:= \arg\min_{b \in \mathbb{R}^{v+1}} [\mathrm{vech}(\log M_T) - Eb]^\intercal W [\mathrm{vech}(\log M_T) - Eb] \\ &=: \arg\min_{b \in \mathbb{R}^{v+1}} g_T(b)^\intercal W g_T(b).\end{aligned}$$

Theorem 8. *Fix n (and hence v). Let Assumptions 1(i), 3, 4 and 5 be satisfied. Thus, under H_0,*

$$T g_T(\hat{\theta}_T)^\intercal \hat{S}^{-1} g_T(\hat{\theta}_T) \xrightarrow{d} \chi^2_{n(n+1)/2-(v+1)}, \tag{6.5}$$

where

$$\hat{S} := D_n^+ \hat{H}_T (D^{-1/2} \otimes D^{-1/2}) \hat{V}_T (D^{-1/2} \otimes D^{-1/2}) \hat{H}_T D_n^{+\intercal}.$$

Proof. See Appendix A. □

From Theorem 8, we can easily get a result of *diagonal path* asymptotics, which is more general than *sequential* asymptotics but less general than *joint* asymptotics (see Phillips and Moon (1999)).

Corollary 3. *Let Assumptions 1(i), 3, 4 and 5 be satisfied. There exists a sequence $n_T \to \infty$ such that, under H_0,*

$$\frac{T g_{T,n_T}(\hat{\theta}_{T,n_T})^\intercal \hat{S}^{-1} g_{T,n_T}(\hat{\theta}_{T,n_T}) - \left[\frac{n_T(n_T+1)}{2} - (v_T+1)\right]}{\left[n_T(n_T+1) - 2(v_T+1)\right]^{1/2}} \xrightarrow{d} N(0,1),$$

as $T \to \infty$.

Proof. See Appendix A. □

7 Model Selection Issues

There are a number of model selection issues that arise in our context, and we briefly comment on them. In the absence of an explicit multiarray structure we may consider the choice of factorization in (2.1). Suppose that n has the unique prime factorization $n = p_1 p_2 \cdots p_v$ for some positive integer v and primes p_j for $j = 1, \ldots, v$. Then there are several different Kronecker product factorizations, which can be described by the dimensions of the square submatrices. The base model we have focussed on has dimensions:

$$p_1 \times p_1, \ldots, p_v \times p_v,$$

but there are many possible aggregations of this, for example

$$(p_1 + p_2) \times (p_1 + p_2), \ldots, (p_{v-1} + p_v) \times (p_{v-1} + p_v)$$

and so on. We may index the induced models by the dimensions $m_1, \ldots, m_v$ (where some could be zero dimensions), which are subject to the constraint that $\sum_{j=1}^{v} m_j = n$ and $m_j = \sum_{i=1}^{v} \pi_{ji} p_i$ with $\pi_{ji} \in \{0, 1\}$. Let the total number of free parameters be $q = \sum_{j=1}^{v} (m_j + 1) m_j / 2$ (minus identification restrictions). This includes the base model and the unrestricted $n \times n$ model as special cases. The Kronecker product structure is not invariant with respect to permutations of the series in the system, so we should also in principle consider all of the possible permutations of the series.[5]

We might choose between these models using some model choice criterion that penalizes the larger models. For example,

$$BIC = -2\ell_T(\hat{\theta}) + q \log T.$$

Typically, there are not so many subfactorizations to consider, so this is not computationally burdensome.

8 Simulation Study

We provide a small simulation study that evaluates the performance of the QMLE in two cases: when the Kronecker product structure is true for the covariance matrix; and when the Kronecker product structure is not present.

8.1 Kronecker Structure Is True

We simulate T random vectors x_t of dimension n according to

$$x_t = \Sigma^{1/2} z_t, \quad z_t \sim N(0, I_n)$$
$$\Sigma = \Sigma_1 \otimes \Sigma_2 \otimes \cdots \otimes \Sigma_v,$$

where $n = 2^v$ and $v \in \mathbb{N}$. The matrices Σ_j are 2×2. These matrices Σ_j are generated with unit variances and off-diagonal elements drawn from a uniform distribution on $(0, 1)$. This ensures positive definiteness of Σ. The sample size is set to $T = 300$. In the estimation procedure, the upper diagonal elements of $\Sigma_j, j \geq 2$, are set to 1 for identification. Altogether, there are $2v + 1$ parameters to estimate by maximum likelihood.

As in Ledoit and Wolf (2004), we use a *percentage relative improvement in average loss* (PRIAL) criterion, to measure the performance of the Kronecker estimator $\hat{\Sigma}$ with respect to the sample covariance estimator M_T. It is defined as

$$PRIAL1 = 1 - \frac{\mathbb{E}\|\hat{\Sigma} - \Sigma\|_F^2}{\mathbb{E}\|M_T - \Sigma\|_F^2}$$

[5]It is interesting to note that for particular functions of the covariance matrix, the ordering of the data does not matter. For example, the minimum variance portfolio (MVP) weights only depend on the covariance matrix through the row weights of its inverse, $\Sigma^{-1}\iota_n$, where ι_n is a vector of ones. If a Kronecker structure is imposed on Σ, then its inverse has the same structure. If the Kronecker factors are (2×2) and all variances are identical, then the row sums of Σ^{-1} are the same, leading to equal weights for the MVP: $w = (1/n)\iota_n$, and this is irrespective of the ordering of the data.

n	4	8	16	32	64	128	256
PRIAL1	0.33	0.69	0.86	0.94	0.98	0.99	0.99
PRIAL2	0.34	0.70	0.89	0.97	0.99	1.00	1.00
VR	0.997	0.991	0.975	0.944	0.889	0.768	0.386

Table 1: *PRIAL1 and PRIAL2 are the medians of the PRIAL1 and PRIAL2 criteria, respectively, for the Kronecker estimator with respect to the sample covariance estimator in the case of true Kronecker structure. VR is the median of the ratio of the variance of the MVP using the Kronecker estimator to that using the sample covariance estimator. The sample size is fixed at* $T = 300$.

where Σ is the true covariance matrix generated as above, $\hat{\Sigma}$ is Kronecker estimator estimated by quasi maximum likelihood, and M_T is the sample covariance matrix defined in (5.2). Often the estimator of the *precision matrix*, Σ^{-1}, is more important than that of Σ itself, so we also compute the PRIAL for the inverse covariance matrix, i.e.

$$PRIAL2 = 1 - \frac{\mathbb{E}\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_F^2}{\mathbb{E}\|M_T^{-1} - \Sigma^{-1}\|_F^2}.$$

Note that this requires invertibility of the sample covariance matrix $\tilde{\Sigma}$ and therefore can only be calculated for $n < T$.

Our final criterion is the minimum variance portfolio (MVP) constructed from an estimator of the covariance matrix, see Section 3.3. The first portfolio weights are constructed using the sample covariance matrix M_T and the second portfolio weights are constructed using the Kronecker factorized matrix $\hat{\Sigma}$. These two portfolios are then evaluated (by calculating the variance) using the out-of-sample returns generated using the same data generating mechanism. The ratio of the variance of the latter portfolio over that of the former (VR) is recorded. See Fan, Liao, and Shi (2015) for discussion of risk estimation for large dimensional portfolio choice problems.

We repeat the simulation 1000 times and obtain for each simulation PRIAL1, PRIAL2 and VR. Table 1 reports the median of the obtained PRIALs and VR for various dimensions. Clearly, as the dimension increases, the Kronecker estimator rapidly outperforms the sample covariance estimator. The relative performance of the precision matrix estimator (PRIAL2) is very similar. In terms of the ratio of MVP variances, the Kronecker estimator yields a 23.2 percent smaller variance for $n = 128$ and 61.4 percent for $n = 256$. The reduction becomes clear as n approaches T.

8.2 Kronecker Structure Is Not True

We now generate random vectors with covariance matrices that do not have a Kronecker structure. Similar to Ledoit and Wolf (2004), and without loss of generality, we generate diagonal covariance matrices with log-normally distributed diagonal elements. (Note that having a diagonal matrix does not necessary imply a Kronecker product structure.) The mean of the eigenvalues (i.e., the diagonal elements) is, without loss of generality, fixed at one, while their dispersion varies and is given by α^2. This dispersion can be viewed as measure for the distance from a Kronecker structure.

We report in Table 2 the results for $n/T \in \{0.5, 0.8\}$ and varying α^2. First, the relative performance of the Kronecker estimator of the precision matrix is better than that of the covariance matrix itself, comparing PRIAL2 with PRIAL1. Second, as n/T approaches 1

α^2	0.05	0.25	0.5	0.75	1	1.25	1.5	1.75	2
					$n/T = 0.5$				
PRIAL1	89.92	60.85	31.10	2.87	-23.70	-46.81	-65.29	-85.59	-106.30
PRIAL2	99.17	96.60	93.37	90.74	88.25	86.61	84.11	82.99	80.47
VR	73.33	77.90	84.44	89.49	93.83	97.14	100.78	102.85	108.71
					$n/T = 0.8$				
PRIAL1	92.78	74.29	54.05	36.36	20.31	4.59	-8.14	-15.60	-25.70
PRIAL2	99.97	99.89	99.78	99.71	99.61	99.56	99.49	99.45	99.38
VR	47.26	51.42	54.82	57.06	61.40	62.59	64.69	67.77	69.12

Table 2: *PRIAL1 and PRIAL2 are the medians of the PRIAL1 and PRIAL2 criteria (multiplied by 100), respectively, for the Kronecker estimator with respect to the sample covariance estimator in the case of true non-Kronecker structure. VR is the median of the ratio (multiplied by 100) of the variance of the MVP using the Kronecker estimator to that using the sample covariance estimator. α^2 is the dispersion of eigenvalues of the true covariance matrix.*

the PRIALs increase, while they decrease as the eigenvalue dispersion α^2 increases. This behaviour of the Kronecker estimator as a function of n/T and α^2 resembles that of the shrinkage estimator of Ledoit and Wolf (2004).

9 Application

We apply the model to a set of $n = 441$ daily stock returns x_t of the S&P 500 index, observed from January 3, 2005, to November 6, 2015. The number of trading days is $T = 2732$.

The Kronecker model is fitted to the correlation matrix $\Theta = D^{-1/2}\Sigma D^{-1/2}$, where D is the diagonal matrix containing the variances on the diagonal. The first model (M1) uses the factorization $2^9 = 512$ and assumes that

$$\Theta = \Theta_1 \otimes \Theta_2 \otimes \cdots \otimes \Theta_9,$$

where Θ_j are 2×2 correlation matrices. We add a vector of 71 independent pseudo variables $u_t \sim N(0, I_{71})$ such that $n + 71 = 2^9$, and then extract the upper left $(n \times n)$ block of Θ to obtain the correlation matrix of x_t.

The estimation is done in two steps: First, D is estimated using the sample variances, and then the correlation parameters are estimated by quasi maximum likelihood using the standardized returns, $\hat{D}^{-1/2}x_t$. Random permutations only lead to negligible improvements of the likelihood, so we keep the original order of the data. We experiment with more generous decompositions by looking at all factorizations of the numbers from 441 to 512, and selecting those yielding not more than 30 parameters. Table 3 gives a summary of these models including estimation results. The Schwarz information criterion favours the specification of model M6 with 27 parameters.

Next, we follow the approach of Fan et al. (2013) and estimate the model on windows of size m days that are shifted from the beginning to the end of the sample. After each estimation, the model is evaluated using the next 21 trading days (one month) out-of-sample. Then the estimation window of m days is shifted by one month, etc. After each estimation step, the estimated model yields an estimator of the covariance matrix that is used to construct minimum variance portfolio (MVP) weights. The same is done for

					Sample cov		SFM ($K=3$)		SFM ($K=4$)	
Model	p	decomp	logL/T	BIC/T	prop	impr	prop	impr	prop	impr
M1	9	$512 = 2^9$	-145.16	290.34	.89	27%	.25	-14%	.27	-15%
M2	16	$486 = 2 \times 3^5$	-141.85	283.74	.90	29%	.43	-4%	.42	-6 %
M3	17	$512 = 2^5 \times 4^2$	-140.91	281.87	.90	29%	.44	-2%	.41	-6%
M4	18	$480 = 2^5 \times 3 \times 5$	-139.63	279.31	.90	30%	.49	1%	.47	0%
M5	25	$512 = 4^4 \times 2$	-139.06	278.19	.91	30%	.53	5%	.53	4%
M6	27	$448 = 2^6 \times 7$	-134.27	268.61	.91	32%	.58	11%	.57	9%
M7	27	$450 = 2 \times 3^2 \times 5^2$	-137.33	274.73	.91	31%	.57	8%	.56	6%

Table 3: *Summary of Kronecker specifications of the correlation matrix. p is the number of parameters of the model, decomp is the factorization used for the full system including the additional pseudo variables, $logL/T$ the log-likelihood value, divided by the number of observations, and BIC/T is the value of the Schwarz information criterion, divided by the number of observations. Prop is the proportion of the time that the Kronecker MVP outperforms a competing model (sample covariance matrix, and a strict factor model (SFM) with $K = 3$ and $K = 4$ factors), and Impr is the percentage of average risk improvements.*

two competing devices: the sample covariance matrix and the strict factor model (SFM). For the SFM, the number of factors K is chosen as in Bai and Ng (2002), and equation (2.14) of Fan et al. (2013). The penalty functions IC1 and IC2 give optimal values K of 3 and 4, respectively, so we report results for both models. The last columns of Table 3 summarize the relative performance of the Kronecker model with respect to the sample covariance matrix and SFM.

All models outperform the sample covariance matrix, while only the more generous factorizations also outperform the SFM. Comparing the results with Table 6 of Fan et al. (2013) for similar data, it appears that the performance of the favored model M6 is quite close to their POET estimator. So our estimator may provide an alternative to high dimensional covariance modelling.

10 Conclusions

We have established the large sample properties of our estimation methods when the matrix dimensions increase. In particular, we obtained consistency and asymptotic normality. The method outperforms the sample covariance method theoretically, in a simulation study, and in an application to portfolio choice. It is possible to extend the framework in various directions to improve performance.

One extension concerns using the Kronecker factorization for general parameter matrices. For example, in the so called BEKK model for multivariate GARCH processes, the parameter matrices are of a Kronecker parameterization form $\mathcal{A} = A \otimes A$, where A is an $n \times n$ matrix, while $\mathcal{A}$ is an $n^2 \times n^2$ matrix that is a typical parameter of the dynamic process. In the case where n is composite one could consider further Kronecker factorizations that would allow one to treat very much larger systems. This approach has been considered in Hoff (2015) for vector autoregressions.

11 Appendix A

11.1 More Details about the Matrix E_*

Proposition 5. *If*

$$\Omega^0 = (\Omega_1^0 \otimes I_2 \otimes \cdots \otimes I_2) + (I_2 \otimes \Omega_2^0 \otimes \cdots \otimes I_2) + \cdots + (I_2 \otimes \cdots \otimes \Omega_v^0),$$

where Ω^0 *is* $n \times n \equiv 2^v \times 2^v$ *and* Ω_i^0 *is* 2×2 *for* $i = 1, \ldots, v$. *Then*

$$vech(\Omega^0) = \begin{bmatrix} E_1 & E_2 & \cdots & E_v \end{bmatrix} \begin{bmatrix} vech(\Omega_1^0) \\ vech(\Omega_2^0) \\ \vdots \\ vech(\Omega_v^0) \end{bmatrix},$$

where

$$E_i := D_n^+ (I_{2^i} \otimes K_{2^{v-i},2^i} \otimes I_{2^{v-i}}) \left(I_{2^{2i}} \otimes vecI_{2^{v-i}}\right) (I_{2^{i-1}} \otimes K_{2,2^{i-1}} \otimes I_2)(vecI_{2^{i-1}} \otimes I_4) D_2, \tag{11.1}$$

where D_n^+ *is the Moore-Penrose generalised inverse of* D_n, D_n *and* D_2 *are the* $n^2 \times n(n+1)/2$ *and* $2^2 \times 2(2+1)/2$ *duplication matrices, respectively, and* $K_{2^{v-i},2^i}$ *and* $K_{2,2^{i-1}}$ *are commutation matrices of various dimensions.*

Proof of Proposition 5. We first consider $\text{vec}(\Omega_1^0 \otimes I_2 \otimes \cdots \otimes I_2)$.

$$\begin{aligned}
&\text{vec}(\Omega_1^0 \otimes I_2 \otimes \cdots \otimes I_2) = \text{vec}(\Omega_1^0 \otimes I_{2^{v-1}}) = \left(I_2 \otimes K_{2^{v-1},2} \otimes I_{2^{v-1}}\right) \left(\text{vec}\Omega_1^0 \otimes \text{vec}I_{2^{v-1}}\right) \\
&= \left(I_2 \otimes K_{2^{v-1},2} \otimes I_{2^{v-1}}\right) \left(I_4 \text{vec}\Omega_1^0 \otimes \text{vec}I_{2^{v-1}} \cdot 1\right) \\
&= \left(I_2 \otimes K_{2^{v-1},2} \otimes I_{2^{v-1}}\right) \left(I_4 \otimes \text{vec}I_{2^{v-1}}\right) \text{vec}\Omega_1^0,
\end{aligned}$$

where the second equality is due to Magnus and Neudecker (2007) Theorem 3.10 p55. Thus,

$$\text{vech}(\Omega_1^0 \otimes I_2 \otimes \cdots \otimes I_2) = D_n^+ \left(I_2 \otimes K_{2^{v-1},2} \otimes I_{2^{v-1}}\right) \left(I_4 \otimes \text{vec}I_{2^{v-1}}\right) D_2 \text{vech}\Omega_1^0, \tag{11.2}$$

where D_n^+ is the Moore-Penrose inverse of D_n, i.e., $D_n^+ = (D_n^\intercal D_n)^{-1} D_n^\intercal$, and D_n and D_2 are the $n^2 \times n(n+1)/2$ and $2^2 \times 2(2+1)/2$ duplication matrices, respectively. We now consider $\text{vec}(I_2 \otimes \Omega_2^0 \otimes \cdots \otimes I_2)$.

$$\begin{aligned}
&\text{vec}(I_2 \otimes \Omega_2^0 \otimes \cdots \otimes I_2) = \text{vec}(I_2 \otimes \Omega_2^0 \otimes I_{2^{v-2}}) = (I_4 \otimes K_{2^{v-2},4} \otimes I_{2^{v-2}}) \left(\text{vec}(I_2 \otimes \Omega_2^0) \otimes \text{vec}I_{2^{v-2}}\right) \\
&= (I_4 \otimes K_{2^{v-2},4} \otimes I_{2^{v-2}}) \left(I_{2^4} \otimes \text{vec}I_{2^{v-2}}\right) \text{vec}(I_2 \otimes \Omega_2^0) \\
&= (I_4 \otimes K_{2^{v-2},4} \otimes I_{2^{v-2}}) \left(I_{2^4} \otimes \text{vec}I_{2^{v-2}}\right) (I_2 \otimes K_{2,2} \otimes I_2)(\text{vec}I_2 \otimes \text{vec}\Omega_2^0) \\
&= (I_4 \otimes K_{2^{v-2},4} \otimes I_{2^{v-2}}) \left(I_{2^4} \otimes \text{vec}I_{2^{v-2}}\right) (I_2 \otimes K_{2,2} \otimes I_2)(\text{vec}I_2 \otimes I_4)\text{vec}\Omega_2^0.
\end{aligned}$$

Thus

$$\begin{aligned}
&\text{vech}(I_2 \otimes \Omega_2^0 \otimes \cdots \otimes I_2) \\
&= D_n^+ (I_4 \otimes K_{2^{v-2},4} \otimes I_{2^{v-2}}) \left(I_{2^4} \otimes \text{vec}I_{2^{v-2}}\right) (I_2 \otimes K_{2,2} \otimes I_2)(\text{vec}I_2 \otimes I_4) D_2 \text{vech}\Omega_2^0.
\end{aligned} \tag{11.3}$$

Next we consider $\mathrm{vec}(I_2 \otimes I_2 \otimes \Omega_3^0 \otimes \cdots \otimes I_2)$.

$$\begin{aligned}
&\mathrm{vec}(I_2 \otimes I_2 \otimes \Omega_3^0 \otimes \cdots \otimes I_2) = \mathrm{vec}(I_4 \otimes \Omega_3^0 \otimes I_{2^{v-3}}) \\
&= (I_{2^3} \otimes K_{2^{v-3},2^3} \otimes I_{2^{v-3}}) \left(\mathrm{vec}(I_4 \otimes \Omega_3^0) \otimes \mathrm{vec} I_{2^{v-3}}\right) \\
&= (I_{2^3} \otimes K_{2^{v-3},2^3} \otimes I_{2^{v-3}}) \left(I_{2^6} \otimes \mathrm{vec} I_{2^{v-3}}\right) \mathrm{vec}(I_4 \otimes \Omega_3^0) \\
&= (I_{2^3} \otimes K_{2^{v-3},2^3} \otimes I_{2^{v-3}}) \left(I_{2^6} \otimes \mathrm{vec} I_{2^{v-3}}\right) (I_4 \otimes K_{2,4} \otimes I_2)(\mathrm{vec} I_4 \otimes \mathrm{vec}\Omega_3^0) \\
&= (I_{2^3} \otimes K_{2^{v-3},2^3} \otimes I_{2^{v-3}}) \left(I_{2^6} \otimes \mathrm{vec} I_{2^{v-3}}\right) (I_4 \otimes K_{2,4} \otimes I_2)(\mathrm{vec} I_4 \otimes I_4)\mathrm{vec}\Omega_3^0.
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathrm{vech}(I_2 \otimes I_2 \otimes \Omega_3^0 \otimes \cdots \otimes I_2) \\
&= D_n^+ (I_{2^3} \otimes K_{2^{v-3},2^3} \otimes I_{2^{v-3}}) \left(I_{2^6} \otimes \mathrm{vec} I_{2^{v-3}}\right) (I_4 \otimes K_{2,4} \otimes I_2)(\mathrm{vec} I_4 \otimes I_4) D_2 \mathrm{vech}\Omega_3^0.
\end{aligned} \tag{11.4}$$

By observing (11.2), (11.3) and (11.4), we deduce the following general formula: for $i = 1, 2, \ldots, v$

$$\begin{aligned}
&\mathrm{vech}(I_2 \otimes \cdots \otimes \Omega_i^0 \otimes \cdots \otimes I_2) \\
&= D_n^+ (I_{2^i} \otimes K_{2^{v-i},2^i} \otimes I_{2^{v-i}}) \left(I_{2^{2i}} \otimes \mathrm{vec} I_{2^{v-i}}\right) (I_{2^{i-1}} \otimes K_{2,2^{i-1}} \otimes I_2)(\mathrm{vec} I_{2^{i-1}} \otimes I_4) D_2 \mathrm{vech}\Omega_i^0 \\
&=: E_i \mathrm{vech}\Omega_i^0,
\end{aligned} \tag{11.5}$$

where E_i is a $n(n+1)/2 \times 3$ matrix. Using (11.5), we have

$$\begin{aligned}
\mathrm{vech}(\Omega^0) &= E_1 \mathrm{vech}(\Omega_1^0) + E_2 \mathrm{vech}(\Omega_2^0) + \cdots + E_v \mathrm{vech}(\Omega_v^0) \\
&= \begin{bmatrix} E_1 & E_2 & \cdots & E_v \end{bmatrix} \begin{bmatrix} \mathrm{vech}(\Omega_1^0) \\ \mathrm{vech}(\Omega_2^0) \\ \vdots \\ \mathrm{vech}(\Omega_v^0) \end{bmatrix}
\end{aligned}$$

□

Proof of Proposition 1. The $E_*^\intercal E_*$ can be written down using the analytical formula in (4.1). The R code for computing this is available upon request. The proofs of the claims (i) - (v) are similar to those in the observations made in Example 2. □

11.2 Proof of Proposition 3

Proof. Since both $A + B$ and A are positive definite for all n, with minimum eigenvalues real and bounded away from zero by absolute constants, by Theorem 9 in Appendix B, we have

$$\log(A+B) = \int_0^1 (A+B-I)[t(A+B-I)+I]^{-1} dt, \quad \log A = \int_0^1 (A-I)[t(A-I)+I]^{-1} dt.$$

Use (6.1) to invoke Proposition 15 in Appendix B to expand $[t(A-I)+I+tB]^{-1}$ to get

$$[t(A-I)+I+tB]^{-1} = [t(A-I)+I]^{-1} - [t(A-I)+I]^{-1} tB [t(A-I)+I]^{-1} + O(\|B\|_{\ell_2}^2)$$

and substitute into the expression of $\log(A+B)$

$$
\begin{aligned}
&\log(A+B) \\
&= \int_0^1 (A+B-I)\left\{[t(A-I)+I]^{-1} - [t(A-I)+I]^{-1}tB[t(A-I)+I]^{-1} + O(\|B\|_{\ell_2}^2)\right\} dt \\
&= \log A + \int_0^1 B[t(A-I)+I]^{-1}dt - \int_0^1 t(A+B-I)[t(A-I)+I]^{-1}B[t(A-I)+I]^{-1}dt \\
&\quad + (A+B-I)O(\|B\|_{\ell_2}^2) \\
&= \log A + \int_0^1 [t(A-I)+I]^{-1}B[t(A-I)+I]^{-1}dt - \int_0^1 tB[t(A-I)+I]^{-1}B[t(A-I)+I]^{-1}dt \\
&\quad + (A+B-I)O(\|B\|_{\ell_2}^2) \\
&= \log A + \int_0^1 [t(A-I)+I]^{-1}B[t(A-I)+I]^{-1}dt + O(\|B\|_{\ell_2}^2 \vee \|B\|_{\ell_2}^3),
\end{aligned}
$$

where the last equality follows from $\text{maxeval}(A) < C < \infty$ and $\text{mineval}[t(A-I)+I] > C' > 0$. □

11.3 Proof of Proposition 4

Denote $\hat{\mu} := \frac{1}{T}\sum_{t=1}^T x_t$.

Proposition 6. *Suppose Assumptions 1(i), 2(i), and 3(i) hold. We have*

(i)

$$\left\|\frac{1}{T}\sum_{t=1}^T x_t x_t^\intercal - \mathbb{E}x_t x_t^\intercal\right\|_{\ell_2} = O_p\left(\max\left(\frac{n}{T}, \sqrt{\frac{n}{T}}\right)\right) = O_p\left(\sqrt{\frac{n}{T}}\right).$$

(ii) $\|D^{-1}\|_{\ell_2} = O(1)$, $\|D^{-1/2}\|_{\ell_2} = O(1)$.

(iii)

$$\|2\mu\mu^\intercal - \hat{\mu}\mu^\intercal - \mu\hat{\mu}^\intercal\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right).$$

(iv)

$$\max_{1\le i\le n}|\mu_i| = O(1).$$

Proof. For part (i), invoke Lemma 2 in Appendix B with $\varepsilon = 1/4$:

$$
\begin{aligned}
\left\|\frac{1}{T}\sum_{t=1}^T x_t x_t^\intercal - \mathbb{E}x_t x_t^\intercal\right\|_{\ell_2} &\le 2\max_{a\in\mathcal{N}_{1/4}}\left|a^\intercal\left(\frac{1}{T}\sum_{t=1}^T x_t x_t^\intercal - \mathbb{E}x_t x_t^\intercal\right)a\right| \\
&=: 2\max_{a\in\mathcal{N}_{1/4}}\left|\frac{1}{T}\sum_{t=1}^T (z_{a,t}^2 - \mathbb{E}z_{a,t}^2)\right|,
\end{aligned}
$$

where $z_{a,t} := x_t^\intercal a$. By Assumption 1(i), $\{z_{a,t}\}_{t=1}^T$ are independent subgaussian random variables. For $\epsilon > 0$,

$$\mathbb{P}(|z_{a,t}^2| \ge \epsilon) = \mathbb{P}(|z_{a,t}| \ge \sqrt{\epsilon}) \le Ke^{-C\epsilon}.$$

We shall use Orlicz norms as defined in van der Vaart and Wellner (1996): Let ψ be a non-decreasing, convex function with $\psi(0) = 0$. Then, the Orlicz norm of a random variable X is given by

$$\|X\|_\psi = \inf\left\{C > 0 : \mathbb{E}\psi\left(|X|/C\right) \leq 1\right\},$$

where $\inf \emptyset = \infty$. We shall use Orlicz norms for $\psi(x) = \psi_p(x) = e^{x^p} - 1$ for $p = 1, 2$ in this paper. It follows from Lemma 2.2.1 in van der Vaart and Wellner (1996) that $\|z_{a,t}^2\|_{\psi_1} \leq (1 + K)/C$. Then

$$\|z_{a,t}^2 - \mathbb{E}z_{a,t}^2\|_{\psi_1} \leq \|z_{a,t}^2\|_{\psi_1} + \mathbb{E}\|z_{a,t}^2\|_{\psi_1} \leq \frac{2(1+K)}{C}.$$

Then, by the definition of the Orlicz norm, $\mathbb{E}\left[e^{C/(2+2K)|z_{a,t}^2 - \mathbb{E}z_{a,t}^2|}\right] \leq 2$. Use Fubini's theorem to expand out the exponential moment. It is easy to see that $z_{a,t}^2 - \mathbb{E}z_{a,t}^2$ satisfies the moment conditions of Bernstein's inequality in Appendix B with $A = \frac{2(1+K)}{C}$ and $\sigma_0^2 = \frac{8(1+K)^2}{C^2}$. Now invoke Bernstein's inequality for all $\epsilon > 0$

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}(z_{a,t}^2 - \mathbb{E}z_{a,t}^2)\right| \geq \sigma_0^2\left[A\epsilon + \sqrt{2\epsilon}\right]\right) \leq 2e^{-T\sigma_0^2\epsilon}.$$

Invoking Lemma 1 in Appendix B, we have $|\mathcal{N}_{1/4}| \leq 9^n$. Now we use the union bound:

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T} x_t x_t^\intercal - \mathbb{E}x_t x_t^\intercal\right\|_{\ell_2} \geq 2\sigma_0^2\left[A\epsilon + \sqrt{2\epsilon}\right]\right) \leq 2e^{n(\log 9 - \sigma_0^2\epsilon T/n)}.$$

Fix $\varepsilon > 0$. There exist $M_\varepsilon = M = \log 9 + 1$, T_ε, and $N_\varepsilon = -\log(\varepsilon/2)$. Setting $\epsilon = \frac{nM_\varepsilon}{T\sigma_0^2}$, the preceding inequality becomes, for all $n > N_\varepsilon$

$$\mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T} x_t x_t^\intercal - \mathbb{E}x_t x_t^\intercal\right\|_{\ell_2} \geq B_\varepsilon\frac{n}{T} + C_\varepsilon\sqrt{\frac{n}{T}}\right) \leq \varepsilon,$$

where $B_\varepsilon := 2AM_\varepsilon$ and $C_\varepsilon := \sigma_0\sqrt{8M_\varepsilon}$. Thus, for all $\varepsilon > 0$, there exist $D_\varepsilon := 2\max(B_\varepsilon, C_\varepsilon)$, T_ε and N_ε, such that for all $T > T_\varepsilon$ and all $n > N_\varepsilon$

$$\mathbb{P}\left(\frac{1}{\max\left(\frac{n}{T}, \sqrt{\frac{n}{T}}\right)}\left\|\frac{1}{T}\sum_{t=1}^{T} x_t x_t^\intercal - \mathbb{E}x_t x_t^\intercal\right\|_{\ell_2} \geq D_\varepsilon\right) \leq \varepsilon.$$

The result follows immediately from the definition of stochastic orders. Part (ii) follows trivially from Assumption 3(i). For part (iii), first recognise that $2\mu\mu^\intercal - \hat{\mu}\mu^\intercal - \mu\hat{\mu}^\intercal$ is symmetric. Invoking Lemma 2 in Appendix B for $\varepsilon = 1/4$, we have

$$\|2\mu\mu^\intercal - \hat{\mu}\mu^\intercal - \mu\hat{\mu}^\intercal\|_{\ell_2} \leq 2\max_{a\in\mathcal{N}_{1/4}} |a^\intercal\left(2\mu\mu^\intercal - \hat{\mu}\mu^\intercal - \mu\hat{\mu}^\intercal\right)a|.$$

It suffices to find a bound for the right hand side of the preceding inequality.

$$\begin{aligned}
&\max_{a\in\mathcal{N}_{1/4}} |a^\intercal\left(2\mu\mu^\intercal - \hat{\mu}\mu^\intercal - \mu\hat{\mu}^\intercal\right)a| = \max_{a\in\mathcal{N}_{1/4}} |a^\intercal\left((\mu - \hat{\mu})\mu^\intercal + \mu(\mu - \hat{\mu})^\intercal\right)a| \\
&\leq \max_{a\in\mathcal{N}_{1/4}} |a^\intercal\mu\left(\hat{\mu} - \mu\right)^\intercal a| + \max_{a\in\mathcal{N}_{1/4}} |a^\intercal\left(\hat{\mu} - \mu\right)\mu^\intercal a| \leq 2\max_{a\in\mathcal{N}_{1/4}} |a^\intercal(\hat{\mu} - \mu)| \max_{a\in\mathcal{N}_{1/4}} |\mu^\intercal a|
\end{aligned}$$

We bound $\max_{a\in\mathcal{N}_{1/4}}|(\hat{\mu}-\mu)^\intercal a|$ first.

$$(\hat{\mu}-\mu)^\intercal a = \frac{1}{T}\sum_{t=1}^{T}(x_t^\intercal a - \mathbb{E}x_t^\intercal a) =: \frac{1}{T}\sum_{t=1}^{T}(z_{a,t}-\mathbb{E}z_{a,t}).$$

By Assumption 1(i), $\{z_{a,t}\}_{t=1}^T$ are independent subgaussian random variables. For $\epsilon>0$, $\mathbb{P}(|z_{a,t}|\geq\epsilon)\leq Ke^{-C\epsilon^2}$. It follows from Lemma 2.2.1 in van der Vaart and Wellner (1996) that $\|z_{a,t}\|_{\psi_2}\leq(1+K)^{1/2}/C^{1/2}$. Then $\|z_{a,t}-\mathbb{E}z_{a,t}\|_{\psi_2}\leq\|z_{a,t}\|_{\psi_2}+\mathbb{E}\|z_{a,t}\|_{\psi_2}\leq\frac{2(1+K)^{1/2}}{C^{1/2}}$. Next, using the second last inequality in van der Vaart and Wellner (1996) p95, we have

$$\|z_{a,t}-\mathbb{E}z_{a,t}\|_{\psi_1}\leq\|z_{a,t}-\mathbb{E}z_{a,t}\|_{\psi_2}(\log 2)^{-1/2}\leq\frac{2(1+K)^{1/2}}{C^{1/2}}(\log 2)^{-1/2}=:\frac{1}{W}.$$

Then, by the definition of the Orlicz norm, $\mathbb{E}\left[e^{W|z_{a,t}-\mathbb{E}z_{a,t}|}\right]\leq 2$. Use Fubini's theorem to expand out the exponential moment. It is easy to see that $z_{a,t}-\mathbb{E}z_{a,t}$ satisfies the moment conditions of Bernstein's inequality in Appendix B with $A=\frac{1}{W}$ and $\sigma_0^2=\frac{2}{W^2}$. Now invoke Bernstein's inequality for all $\epsilon>0$

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}(z_{a,t}-\mathbb{E}z_{a,t})\right|\geq\sigma_0^2\left[A\epsilon+\sqrt{2\epsilon}\right]\right)\leq 2e^{-T\sigma_0^2\epsilon}.$$

Invoking Lemma 1 in Appendix B, we have $|\mathcal{N}_{1/4}|\leq 9^n$. Now we use the union bound:

$$\mathbb{P}\left(\max_{a\in\mathcal{N}_{1/4}}\left|\frac{1}{T}\sum_{t=1}^{T}(z_{a,t}-\mathbb{E}z_{a,t})\right|\geq 2\sigma_0^2\left[A\epsilon+\sqrt{2\epsilon}\right]\right)\leq 2e^{n(\log 9-\sigma_0^2\epsilon T/n)}.$$

Using the same argument as in part (i), we get

$$\max_{a\in\mathcal{N}_{1/4}}|(\hat{\mu}-\mu)^\intercal a| = O_p\left(\sqrt{\frac{n}{T}}\right). \tag{11.6}$$

Now $a^\intercal\mu=\mathbb{E}a^\intercal x_t=:\mathbb{E}y_{a,t}$. Again via Assumption 1(i) and Lemma 2.2.1 in van der Vaart and Wellner (1996), $\|y_{a,t}\|_{\psi_2}\leq C$. Hence

$$\begin{aligned}\max_{a\in\mathcal{N}_{1/4}}|\mathbb{E}y_{a,t}|&\leq\max_{a\in\mathcal{N}_{1/4}}\mathbb{E}|y_{a,t}|=\max_{a\in\mathcal{N}_{1/4}}\|y_{a,t}\|_{L_1}\leq\max_{a\in\mathcal{N}_{1/4}}\|y_{a,t}\|_{\psi_1}\leq\max_{a\in\mathcal{N}_{1/4}}\|y_{a,t}\|_{\psi_2}(\log 2)^{-1/2}\\&\leq C(\log 2)^{-1/2},\end{aligned}$$

where the second and third inequalities are from van der Vaart and Wellner (1996) p95. Thus we have

$$\max_{a\in\mathcal{N}_{1/4}}|a^\intercal\mu|=O(1).$$

The preceding display together with (11.6) deliver the result. For part (iv), via Assumption 1(i), we have $x_{t,i}$ to be subgaussian for all i:

$$\mathbb{P}(|x_{t,i}|\geq\epsilon)\leq Ke^{-C\epsilon^2},$$

for positive constants K and C. It follows from Lemma 2.2.1 in van der Vaart and Wellner (1996) that $\|x_{t,i}\|_{\psi_2}\leq(1+K)^{1/2}/C^{1/2}$. Now

$$\max_{1\leq i\leq n}|\mu_i|=\max_{1\leq i\leq n}|\mathbb{E}x_{t,i}|\leq\max_{1\leq i\leq n}\|x_{t,i}\|_{L_1}\leq\max_{1\leq i\leq n}\|x_{t,i}\|_{\psi_1}\leq\max_{1\leq i\leq n}\|x_{t,i}\|_{\psi_2}(\log 2)^{-1/2},$$

where the second and third inequalities follow from van der Vaart and Wellner (1996) p95. We have already shown that the ψ_2-Orlicz norms are uniformly bounded, so the result follows. □

Proof of Proposition 4. For part (i),

$$\begin{aligned}
&\|M_T - \Theta\|_{\ell_2} = \|D^{-1/2}\tilde{\Sigma}D^{-1/2} - D^{-1/2}\Sigma D^{-1/2}\|_{\ell_2} = \|D^{-1/2}(\tilde{\Sigma} - \Sigma)D^{-1/2}\|_{\ell_2} \\
&\leq \|D^{-1/2}\|_{\ell_2}^2\|\tilde{\Sigma} - \Sigma\|_{\ell_2} = O(1)\|\tilde{\Sigma} - \Sigma\|_{\ell_2} \\
&= O(1)\left\|\frac{1}{T}\sum_{t=1}^{T} x_t x_t^\intercal - \mathbb{E}x_t x_t^\intercal + 2\mu\mu^\intercal - \hat{\mu}\mu^\intercal - \mu\hat{\mu}^\intercal\right\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right), \qquad (11.7)
\end{aligned}$$

where the third and fifth equalities are due to Proposition 6. For part (ii),

$$\begin{aligned}
&\|[t(\Theta - I) + I]^{-1}t(M_T - \Theta)\|_{\ell_2} \leq t\|[t(\Theta - I) + I]^{-1}\|_{\ell_2}\|M_T - \Theta\|_{\ell_2} \\
&= \|[t(\Theta - I) + I]^{-1}\|_{\ell_2} O_p(\sqrt{n/T}) = O_p(\sqrt{n/T})/\text{mineval}(t(\Theta - I) + I) = o_p(1),
\end{aligned}$$

where the first equality is due to part (i), and the last equality is due to mineval$(t(\Theta - I) + I) > C > 0$ for some absolute constant C and Assumption 2(i). □

11.4 Proof of Theorem 1

Proof.

$$\begin{aligned}
&\|\hat{\theta}_T - \theta^0\|_2 = \|(E^\intercal W E)^{-1}E^\intercal W D_n^+ \text{vec}(\log M_T - \log\Theta)\|_2 \\
&\leq \|(E^\intercal W E)^{-1}E^\intercal W^{1/2}\|_{\ell_2}\|W^{1/2}\|_{\ell_2}\|D_n^+\|_{\ell_2}\|\text{vec}(\log M_T - \log\Theta)\|_2,
\end{aligned}$$

where $D_n^+ := (D_n^\intercal D_n)^{-1}D_n^\intercal$ and D_n is the duplication matrix. Since Proposition 4 holds under the assumptions of Theorem 1, together with Assumption 4 and Lemma 2.12 in van der Vaart (1998), we can invoke Proposition 3 stochastically with $A = \Theta$ and $B = M_T - \Theta$:

$$\log M_T - \log\Theta = \int_0^1 [t(\Theta - I) + I]^{-1}(M_T - \Theta)[t(\Theta - I) + I]^{-1}dt + O_p(\|M_T - \Theta\|_{\ell_2}^2). \quad (11.8)$$

(We can invoke Proposition 3 stochastically because the remainder of the log linearization is zero when the perturbation is zero. Moreover, we have $\|M_T - \Theta\|_{\ell_2} \xrightarrow{p} 0$ under Assumption 2(i).) Then

$$\begin{aligned}
&\|\text{vec}(\log M_T - \log\Theta)\|_2 \\
&\leq \left\|\int_0^1 [t(\Theta - I) + I]^{-1} \otimes [t(\Theta - I) + I]^{-1}dt\,\text{vec}(M_T - \Theta)\right\|_2 + \|\text{vec}O_p(\|M_T - \Theta\|_{\ell_2}^2)\|_2 \\
&\leq \left\|\int_0^1 [t(\Theta - I) + I]^{-1} \otimes [t(\Theta - I) + I]^{-1}dt\right\|_{\ell_2} \|M_T - \Theta\|_F + \|O_p(\|M_T - \Theta\|_{\ell_2}^2)\|_F \\
&\leq C\sqrt{n}\|M_T - \Theta\|_{\ell_2} + \sqrt{n}\|O_p(\|M_T - \Theta\|_{\ell_2}^2)\|_{\ell_2} \\
&\leq C\sqrt{n}\|M_T - \Theta\|_{\ell_2} + \sqrt{n}O_p(\|M_T - \Theta\|_{\ell_2}^2) = O_p(\sqrt{n^2/T}), \qquad (11.9)
\end{aligned}$$

where the third inequality is due to (11.12), and the last inequality is due to Proposition 4. Finally,

$$\begin{aligned}
&\|(E^\intercal WE)^{-1}E^\intercal W^{1/2}\|_{\ell_2} = \sqrt{\text{maxeval}\left(\left[(E^\intercal WE)^{-1}E^\intercal W^{1/2}\right]^\intercal (E^\intercal WE)^{-1}E^\intercal W^{1/2}\right)} \\
&= \sqrt{\text{maxeval}\left((E^\intercal WE)^{-1}E^\intercal W^{1/2}\left[(E^\intercal WE)^{-1}E^\intercal W^{1/2}\right]^\intercal\right)} \\
&= \sqrt{\text{maxeval}\left((E^\intercal WE)^{-1}E^\intercal W^{1/2}W^{1/2}E(E^\intercal WE)^{-1}\right)} \\
&= \sqrt{\text{maxeval}\left((E^\intercal WE)^{-1}\right)} = \sqrt{\frac{1}{\text{mineval}(E^\intercal WE)}} \leq \sqrt{\frac{1}{\text{mineval}(E^\intercal E)\text{mineval}(W)}} \\
&= \sqrt{2/n}\sqrt{\|W^{-1}\|_{\ell_2}},
\end{aligned}$$

where the second equality is due to the fact that for any matrix A, $AA^\intercal$ and $A^\intercal A$ have the same non-zero eigenvalues, the third equality is due to $(A^\intercal)^{-1} = (A^{-1})^\intercal$, and the last equality is due to Proposition 2. On the other hand, $D_n^\intercal D_n$ is a diagonal matrix with diagonal entries either 1 or 2, so

$$\|D_n^+\|_{\ell_2} = \|D_n^{+\intercal}\|_{\ell_2} = O(1), \qquad \|D_n\|_{\ell_2} = \|D_n^\intercal\|_{\ell_2} = O(1). \tag{11.10}$$

The result follows after assembling the rates. For future reference

$$\|(E^\intercal WE)^{-1}E^\intercal W\|_{\ell_2} = O(\sqrt{\kappa(W)/n}). \tag{11.11}$$

□

11.5 Proof of Theorem 2

Proposition 7. *Let Assumptions 1(i), 2(i), 3, and 4 be satisfied. Then we have*

$$\|H\|_{\ell_2} = O(1), \qquad \|\hat{H}_T\|_{\ell_2} = O_p(1), \qquad \|\hat{H}_T - H\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right). \tag{11.12}$$

Proof. The proofs for $\|H\|_{\ell_2} = O(1)$ and $\|\hat{H}_T\|_{\ell_2} = O_p(1)$ are exactly the same, so we only give the proof for the latter. Define $A_t := [t(M_T - I) + I]^{-1}$ and $B_t := [t(\Theta - I) + I]^{-1}$.

$$\begin{aligned}
\|\hat{H}_T\|_{\ell_2} &= \left\|\int_0^1 A_t \otimes A_t dt\right\|_{\ell_2} \leq \int_0^1 \left\|A_t \otimes A_t\right\|_{\ell_2} dt \leq \max_{t\in[0,1]}\left\|A_t \otimes A_t\right\|_{\ell_2} = \max_{t\in[0,1]}\|A_t\|_{\ell_2}^2 \\
&= \max_{t\in[0,1]}\{\text{maxeval}([t(M_T - I) + I]^{-1})\}^2 = \max_{t\in[0,1]}\left\{\frac{1}{\text{mineval}(t(M_T - I) + I)}\right\}^2 = O_p(1),
\end{aligned}$$

where the second equality is to Proposition 16 in Appendix B, and the last equality is due to Assumption 4. Now,

$$\begin{aligned}
&\|\hat{H}_T - H\|_{\ell_2} = \left\|\int_0^1 A_t \otimes A_t - B_t \otimes B_t dt\right\|_{\ell_2} \leq \int_0^1 \|A_t \otimes A_t - B_t \otimes B_t\|_{\ell_2} dt \\
&\leq \max_{t\in[0,1]}\|A_t \otimes A_t - B_t \otimes B_t\|_{\ell_2} = \max_{t\in[0,1]}\|A_t \otimes A_t - A_t \otimes B_t + A_t \otimes B_t - B_t \otimes B_t\|_{\ell_2} \\
&= \max_{t\in[0,1]}\left\|A_t \otimes (A_t - B_t) + (A_t - B_t) \otimes B_t\right\|_{\ell_2} \leq \max_{t\in[0,1]}\left(\left\|A_t \otimes (A_t - B_t)\right\|_{\ell_2} + \left\|(A_t - B_t) \otimes B_t\right\|_{\ell_2}\right) \\
&= \max_{t\in[0,1]}\left(\|A_t\|_{\ell_2}\|A_t - B_t\|_{\ell_2} + \|A_t - B_t\|_{\ell_2}\|B_t\|_{\ell_2}\right) = \max_{t\in[0,1]}\|A_t - B_t\|_{\ell_2}\left(\|A_t\|_{\ell_2} + \|B_t\|_{\ell_2}\right) \\
&= O_p(1)\max_{t\in[0,1]}\left\|[t(M_T - I) + I]^{-1} - [t(\Theta - I) + I]^{-1}\right\|_{\ell_2}
\end{aligned}$$

where the first inequality is due to Jensen's inequality, the third equality is due to special properties of Kronecker product, the fourth equality is due to Proposition 16 in Appendix B, and the last equality is because Assumption 4 and Assumption 3(iii)-(iv) implies

$$\|[t(M_T - I) + I]^{-1}\|_{\ell_2} = O_p(1) \qquad \|[t(\Theta - I) + I]^{-1}\|_{\ell_2} = O(1).$$

Now

$$\left\|[t(M_T - I) + I] - [t(\Theta - I) + I]\right\|_{\ell_2} = t\|M_T - \Theta\|_{\ell_2} = O_p(\sqrt{n/T}),$$

where the last equality is due to Proposition 4. The proposition then follows after invoking Lemma 3 in Appendix B. □

Proposition 8. *Let Assumptions 1(i), 2(i) be satisfied. Then*

$$\|\hat{V}_T - V\|_\infty = O_p\left(\sqrt{\frac{\log^5 n^4}{T}}\right).$$

Proof. Let $\dot{x}_{t,i}$ denote $x_{t,i} - \mu_i$, similarly for $\dot{x}_{t,j}, \dot{x}_{t,k}, \dot{x}_{t,l}$.

$$\|\hat{V}_T - V\|_\infty := \max_{1\le x,y\le n^2} |\hat{V}_{T,x,y} - V_{x,y}| = \max_{1\le i,j,k,l\le n} |\hat{V}_{T,i,j,k,l} - V_{i,j,k,l}| \le$$

$$\max_{1\le i,j,k,l\le n}\left|\frac{1}{T}\sum_{t=1}^{T} \dot{x}_{t,i}\dot{x}_{t,j}\dot{x}_{t,k}\dot{x}_{t,l} - \mathbb{E}[\dot{x}_{t,i}\dot{x}_{t,j}\dot{x}_{t,k}\dot{x}_{t,l}]\right| \tag{11.13}$$

$$+ \max_{1\le i,j,l,k\le n}\left|\left(\frac{1}{T}\sum_{t=1}^{T} \dot{x}_{t,i}\dot{x}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T} \dot{x}_{t,k}\dot{x}_{t,l}\right) - \mathbb{E}[\dot{x}_{t,i}\dot{x}_{t,j}]\mathbb{E}[\dot{x}_{t,k}\dot{x}_{t,l}]\right| \tag{11.14}$$

By Assumption 1(i), $x_{t,i}, x_{t,j}, x_{t,k}, x_{t,l}$ are subgaussian random variables. We now show that $\dot{x}_{t,i}, \dot{x}_{t,j}, \dot{x}_{t,k}, \dot{x}_{t,l}$ are also uniformly subgaussian. Without loss of generality consider $\dot{x}_{t,i}$.

$$\mathbb{P}\left(|\dot{x}_{t,i}| \ge \epsilon\right) = \mathbb{P}\left(|x_{t,i} - \mu_i| \ge \epsilon\right) \le \mathbb{P}\left(|x_{t,i}| \ge \epsilon - |\mu_i|\right) \le Ke^{-C(\epsilon - |\mu_i|)^2}$$
$$\le Ke^{-C\epsilon^2}e^{2C\epsilon|\mu_i|}e^{-C|\mu_i|^2} \le Ke^{-C\epsilon^2}e^{2C\epsilon|\mu_i|} \le Ke^{-C\epsilon^2}e^{C(\epsilon^2/2 + 2|\mu_i|^2)}$$
$$= Ke^{-\frac{1}{2}C\epsilon^2}e^{2C|\mu_i|^2} \le Ke^{-\frac{1}{2}C\epsilon^2}e^{2C(\max_{1\le i\le n}|\mu_i|)^2} = K'e^{-\frac{1}{2}C\epsilon^2},$$

where the fifth inequality is due to the decoupling inequality $2xy \le x^2/2 + 2y^2$, and the last equality is due to Proposition 6(iv). We consider (11.13) first. Invoke Proposition 17 in Appendix B:

$$\max_{1\le i,j,l,k\le n}\left|\frac{1}{T}\sum_{t=1}^{T} \dot{x}_{t,i}\dot{x}_{t,j}\dot{x}_{t,k}\dot{x}_{t,l} - \mathbb{E}\dot{x}_{t,i}\dot{x}_{t,j}\dot{x}_{t,k}\dot{x}_{t,l}\right| = O_p\left(\sqrt{\frac{\log^5 n^4}{T}}\right). \tag{11.15}$$

We now consider (11.14).

$$\max_{1\le i,j,l,k\le n}\left|\left(\frac{1}{T}\sum_{t=1}^{T} \dot{x}_{t,i}\dot{x}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T} \dot{x}_{t,k}\dot{x}_{t,l}\right) - \mathbb{E}[\dot{x}_{t,i}\dot{x}_{t,j}]\mathbb{E}[\dot{x}_{t,k}\dot{x}_{t,l}]\right|$$

$$\le \max_{1\le i,j,l,k\le n}\left|\left(\frac{1}{T}\sum_{t=1}^{T} \dot{x}_{t,i}\dot{x}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T} \dot{x}_{t,k}\dot{x}_{t,l} - \mathbb{E}[\dot{x}_{t,k}\dot{x}_{t,l}]\right)\right| \tag{11.16}$$

$$+ \max_{1\le i,j,l,k\le n}\left|\mathbb{E}[\dot{x}_{t,k}\dot{x}_{t,l}]\left(\frac{1}{T}\sum_{t=1}^{T} \dot{x}_{t,i}\dot{x}_{t,j} - \mathbb{E}[\dot{x}_{t,i}\dot{x}_{t,j}]\right)\right|. \tag{11.17}$$

Consider (11.16).

$$\begin{aligned}
&\max_{1\le i,j,l,k\le n}\left|\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,i}\dot{x}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,k}\dot{x}_{t,l}-\mathbb{E}\dot{x}_{t,k}\dot{x}_{t,l}\right)\right| \\
&\le \max_{1\le i,j\le n}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,i}\dot{x}_{t,j}-\mathbb{E}\dot{x}_{t,i}\dot{x}_{t,j}\right|+\left|\mathbb{E}\dot{x}_{t,i}\dot{x}_{t,j}\right|\right)\max_{1\le k,l\le n}\left|\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,k}\dot{x}_{t,l}-\mathbb{E}\dot{x}_{t,k}\dot{x}_{t,l}\right| \\
&= \left(O_p\left(\sqrt{\frac{\log^3 n^2}{T}}\right)+O(1)\right)O_p\left(\sqrt{\frac{\log^3 n^2}{T}}\right)=O_p\left(\sqrt{\frac{\log^3 n^2}{T}}\right) \qquad (11.18)
\end{aligned}$$

where the first equality is due to Proposition 17 in Appendix B and the last equality is due to Assumption 2(i). Now consider (11.17).

$$\begin{aligned}
&\max_{1\le i,j,k,l\le n}\left|\mathbb{E}[\dot{x}_{t,k}\dot{x}_{t,l}]\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,i}\dot{x}_{t,j}-\mathbb{E}[\dot{x}_{t,i}\dot{x}_{t,j}]\right)\right| \\
&\le \max_{1\le k,l\le n}\left|\mathbb{E}[\dot{x}_{t,k}\dot{x}_{t,l}]\right|\max_{1\le i,j\le n}\left|\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,i}\dot{x}_{t,j}-\mathbb{E}\dot{x}_{t,i}\dot{x}_{t,j}\right|=O_p\left(\sqrt{\frac{\log^3 n^2}{T}}\right) \qquad (11.19)
\end{aligned}$$

where the equality is due to Proposition 17 in Appendix B. The proposition follows after summing up the rates for (11.15), (11.18) and (11.19). □

Proof of Theorem 2.

$$\begin{aligned}
\frac{\sqrt{T}c^\intercal(\hat{\theta}_T-\theta^0)}{\sqrt{\hat{G}_T}} &= \frac{\sqrt{T}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\mathrm{vec}(\tilde{\Sigma}-\Sigma)}{\sqrt{\hat{G}_T}} \\
&\quad+\frac{\sqrt{T}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\mathrm{vec}O_p(\|M_T-\Theta\|_{\ell_2}^2)}{\sqrt{\hat{G}_T}} \\
&=: t_1+t_2.
\end{aligned}$$

Define

$$t_1' := \frac{\sqrt{T}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\mathrm{vec}(\tilde{\Sigma}-\Sigma)}{\sqrt{G}}.$$

To prove Theorem 2, it suffices to show $t_1' \xrightarrow{d} N(0,1)$, $t_1'-t_1=o_p(1)$, and $t_2=o_p(1)$.

11.5.1 $t_1' \xrightarrow{d} N(0,1)$

We now prove that t_1' is asymptotically distributed as a standard normal.

$$\begin{aligned}
t_1' &= \frac{\sqrt{T}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\mathrm{vec}\left(\frac{1}{T}\sum_{t=1}^{T}\left[(x_t-\mu)(x_t-\mu)^\intercal-\mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right)}{\sqrt{G}} \\
&= \sum_{t=1}^{T}\frac{T^{-1/2}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\mathrm{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal-\mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]}{\sqrt{G}} \\
&=: \sum_{t=1}^{T}U_{T,n,t}.
\end{aligned}$$

Trivially $\mathbb{E}[U_{T,n,t}] = 0$ and $\sum_{t=1}^{T} \mathbb{E}[U_{T,n,t}^2] = 1$. Then we just need to verify the following Lindeberg condition for a double indexed process (Phillips and Moon (1999) Theorem 2 p1070): for all $\varepsilon > 0$,

$$\lim_{n,T\to\infty} \sum_{t=1}^{T} \int_{\{|U_{T,n,t}|\geq\varepsilon\}} U_{T,n,t}^2 dP = 0.$$

For any $\gamma > 2$,

$$\int_{\{|U_{T,n,t}|\geq\varepsilon\}} U_{T,n,t}^2 dP = \int_{\{|U_{T,n,t}|\geq\varepsilon\}} U_{T,n,t}^2 |U_{T,n,t}|^{-\gamma} |U_{T,n,t}|^{\gamma} dP \leq \varepsilon^{2-\gamma} \int_{\{|U_{T,n,t}|\geq\varepsilon\}} |U_{T,n,t}|^{\gamma} dP$$
$$\leq \varepsilon^{2-\gamma} \mathbb{E}|U_{T,n,t}|^{\gamma}.$$

We first investigate at which rate the denominator $\sqrt{G}$ goes to zero:

$$G = c^{\intercal}(E^{\intercal}WE)^{-1}E^{\intercal}WD_n^{+}H(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})HD_n^{+\intercal}WE(E^{\intercal}WE)^{-1}c$$
$$\geq \text{mineval}(V)\text{mineval}(D^{-1}\otimes D^{-1})\text{mineval}(H^2)\text{mineval}(D_n^{+}D_n^{+\intercal})\text{mineval}(W)\text{mineval}((E^{\intercal}WE)^{-1})$$
$$= \frac{\text{mineval}(V)\text{mineval}^2(H)}{\text{maxeval}(D\otimes D)\text{maxeval}(D_n^{\intercal}D_n)\text{maxeval}(W^{-1})\text{maxeval}(E^{\intercal}WE)}$$
$$\geq \frac{\text{mineval}(V)\text{mineval}^2(H)}{\text{maxeval}(D\otimes D)\text{maxeval}(D_n^{\intercal}D_n)\text{maxeval}(W^{-1})\text{maxeval}(W)\text{maxeval}(E^{\intercal}E)}$$

where the first inequality is true by repeatedly invoking the Rayleigh-Ritz theorem. Since the minimum eigenvalue of H is bounded away from zero by an absolute constant by Assumption 3(i)-(ii), the maximum eigenvalue of D is bounded from above by an absolute constant (Assumption 3(iv)), and maxeval$[D_n^{\intercal}D_n]$ is bounded from above, we have

$$\frac{1}{\sqrt{G}} = O(\sqrt{n\kappa(W)}). \tag{11.20}$$

Then a sufficient condition for the Lindeberg condition is:

$$T^{1-\frac{\gamma}{2}}(n\kappa(W))^{\gamma/2}$$
$$\cdot\, \mathbb{E}\left|c^{\intercal}(E^{\intercal}WE)^{-1}E^{\intercal}WD_n^{+}H(D^{-1/2}\otimes D^{-1/2})\text{vec}\left[(x_t-\mu)(x_t-\mu)^{\intercal} - \mathbb{E}(x_t-\mu)(x_t-\mu)^{\intercal}\right]\right|^{\gamma}$$
$$= o(1), \tag{11.21}$$

for some $\gamma > 2$. We now verify (11.21).

$$\mathbb{E}\left|c^{\intercal}(E^{\intercal}WE)^{-1}E^{\intercal}WD_n^{+}H(D^{-1/2}\otimes D^{-1/2})\text{vec}\left[x_t-\mu)(x_t-\mu)^{\intercal} - \mathbb{E}(x_t-\mu)(x_t-\mu)^{\intercal}\right]\right|^{\gamma}$$
$$\leq \|c^{\intercal}(E^{\intercal}WE)^{-1}E^{\intercal}WD_n^{+}H(D^{-1/2}\otimes D^{-1/2})\|_2^{\gamma}\mathbb{E}\left\|\text{vec}\left[x_t-\mu)(x_t-\mu)^{\intercal} - \mathbb{E}(x_t-\mu)(x_t-\mu)^{\intercal}\right]\right\|_2^{\gamma}$$
$$= O((\kappa(W)/n)^{\gamma/2})\mathbb{E}\left\|x_t-\mu)(x_t-\mu)^{\intercal} - \mathbb{E}(x_t-\mu)(x_t-\mu)^{\intercal}\right\|_F^{\gamma}$$
$$\leq O((\kappa(W)/n)^{\gamma/2})\mathbb{E}\left|\|x_t-\mu)(x_t-\mu)^{\intercal}\|_F + \|\mathbb{E}(x_t-\mu)(x_t-\mu)^{\intercal}\|_F\right|^{\gamma}$$
$$\leq O((\kappa(W)/n)^{\gamma/2})2^{\gamma-1}\left(\mathbb{E}\|x_t-\mu)(x_t-\mu)^{\intercal}\|_F^{\gamma} + \mathbb{E}\|\mathbb{E}(x_t-\mu)(x_t-\mu)^{\intercal}\|_F^{\gamma}\right)$$
$$\leq O((\kappa(W)/n)^{\gamma/2})2^{\gamma}\mathbb{E}\|x_t-\mu)(x_t-\mu)^{\intercal}\|_F^{\gamma} \leq O((\kappa(W)/n)^{\gamma/2})2^{\gamma}\mathbb{E}\left(n\max_{1\leq i,j\leq n}\left|(x_t-\mu)_i(x_t-\mu)_j\right|\right)^{\gamma}$$
$$= O((\kappa(W)n)^{\gamma/2})\mathbb{E}\left(\max_{1\leq i,j\leq n}\left|(x_t-\mu)_i(x_t-\mu)_j\right|^{\gamma}\right) = O((\kappa(W)n)^{\gamma/2})\left\|\max_{1\leq i,j\leq n}\left|(x_t-\mu)_i(x_t-\mu)_j\right|\right\|_{L_\gamma}^{\gamma}$$

where the first equality is because of (11.11), (11.12), and Proposition 6(ii), the third inequality is due to Loeve's c_r inequality, the fourth inequality is due to Jensen's inequality, and the last equality is due to the definition of L_p norm. By Assumption 1(i), for any $i, j = 1, \ldots, n$,

$$\mathbb{P}(|x_{t,i}x_{t,j}| \geq \epsilon) \leq \mathbb{P}(|x_{t,i}| \geq \sqrt{\epsilon}) + \mathbb{P}(|x_{t,j}| \geq \sqrt{\epsilon}) \leq 2Ke^{-C\epsilon}.$$

It follows from Lemma 2.2.1 in van der Vaart and Wellner (1996) that $\|x_{t,i}x_{t,j}\|_{\psi_1} \leq (1+2K)/C$. Similarly we have $\mathbb{P}(|x_{t,i}| \geq \epsilon) \leq Ke^{-C\epsilon^2}$, so $\|x_{t,i}\|_{\psi_1} \leq \|x_{t,i}\|_{\psi_2}(\log 2)^{-1/2} \leq \left[\frac{1+K}{C}\right]^{1/2}(\log 2)^{-1/2}$. Recalling from Proposition 6(iv) that $\max_{1\leq i\leq n}|\mu_i| = O(1)$, we have

$$\|(x_t - \mu)_i(x_t - \mu)_j\|_{\psi_1} \leq \|x_{t,i}x_{t,j}\|_{\psi_1} + \mu_j\|x_{t,i}\|_{\psi_1} + \mu_i\|x_{t,j}\|_{\psi_1} + \mu_i\mu_j \leq C$$

for some constant C. Then invoke Lemma 2.2.2 in van der Vaart and Wellner (1996)

$$\left\|\max_{1\leq i,j\leq n}|(x_t - \mu)_i(x_t - \mu)_j|\right\|_{\psi_1} \lesssim \log(1+n^2)C = O(\log n).$$

Since $\|X\|_{L_r} \leq r!\|X\|_{\psi_1}$ for any random variable X (van der Vaart and Wellner (1996), p95), we have

$$\left\|\max_{1\leq i,j\leq n}|(x_t - \mu)_i(x_t - \mu)_j|\right\|_{L_\gamma}^{\gamma} \leq (\gamma!)^\gamma\left\|\max_{1\leq i,j\leq n}|(x_t - \mu)_i(x_t - \mu)_j|\right\|_{\psi_1}^{\gamma} = O(\log^\gamma n). \tag{11.22}$$

Summing up the rates, we have

$$\begin{aligned}
&T^{1-\frac{\gamma}{2}}(n\kappa(W))^{\gamma/2} \\
&\cdot \mathbb{E}\left|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\text{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal - \mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right|^\gamma \\
&= T^{1-\frac{\gamma}{2}}(\kappa(W)n)^\gamma O(\log^\gamma n) = O\left(\frac{\kappa(W)n\log n}{T^{\frac{1}{2}-\frac{1}{\gamma}}}\right)^\gamma = o(1)
\end{aligned}$$

by Assumption 2(ii). Thus, we have verified (11.21).

11.5.2 $t_1' - t_1 = o_p(1)$

We now show that $t_1' - t_1 = o_p(1)$. Since t_1' and t_1 have the same numerator, say denoted A, we have

$$\begin{aligned}
t_1' - t_1 = \frac{A}{\sqrt{G}} - \frac{A}{\sqrt{\hat{G}_T}} &= \frac{A}{\sqrt{G}}\left(\frac{\sqrt{n\kappa(W)\hat{G}_T} - \sqrt{n\kappa(W)G}}{\sqrt{n\kappa(W)\hat{G}_T}}\right) \\
&= \frac{A}{\sqrt{G}}\frac{1}{\sqrt{n\kappa(W)\hat{G}_T}}\left(\frac{n\kappa(W)\hat{G}_T - n\kappa(W)G}{\sqrt{n\kappa(W)\hat{G}_T} + \sqrt{n\kappa(W)G}}\right).
\end{aligned}$$

Since we have already shown in (11.20) that $n\kappa(W)G$ is bounded away from zero by an absolute constant and $A/\sqrt{G} = O_p(1)$, if in addition we show that $n\kappa(W)\hat{G}_T - n\kappa(W)G = o_p(1)$, then the right hand side of the preceding display is $o_p(1)$ by repeatedly invoking continuous mapping theorem. Now we show that $n\kappa(W)\hat{G}_T - n\kappa(W)G = o_p(1)$. Define

$$\tilde{G}_T := c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c.$$

By the triangular inequality: $|n\kappa(W)\hat{G}_T - n\kappa(W)G| \leq |n\kappa(W)\hat{G}_T - n\kappa(W)\tilde{G}_T| + |n\kappa(W)\tilde{G}_T - n\kappa(W)G|$. First, we prove $|n\kappa(W)\hat{G}_T - n\kappa(W)\tilde{G}_T| = o_p(1)$.

$$
\begin{aligned}
&n\kappa(W)|\hat{G}_T - \tilde{G}_T| \\
&= n\kappa(W)|c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T(D^{-1/2}\otimes D^{-1/2})\hat{V}_T(D^{-1/2}\otimes D^{-1/2})\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c \\
&\quad - c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c| \\
&= n\kappa(W) \\
&\quad \cdot |c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T(D^{-1/2}\otimes D^{-1/2})(\hat{V}_T - V)(D^{-1/2}\otimes D^{-1/2})\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c| \\
&\leq n\kappa(W)\|\hat{V}_T - V\|_\infty \|(D^{-1/2}\otimes D^{-1/2})\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_1^2 \\
&\leq n^3\kappa(W)\|\hat{V}_T - V\|_\infty \|(D^{-1/2}\otimes D^{-1/2})\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2^2 \\
&\leq n^3\kappa(W)\|\hat{V}_T - V\|_\infty \|(D^{-1/2}\otimes D^{-1/2})\|_{\ell_2}^2\|\hat{H}_T\|_{\ell_2}^2\|D_n^{+\intercal}\|_{\ell_2}^2\|WE(E^\intercal WE)^{-1}\|_{\ell_2}^2 \\
&= O_p(n^2\kappa^2(W))\|\hat{V}_T - V\|_\infty = O_p\left(\sqrt{\frac{n^4\kappa^4(W)\log^5 n^4}{T}}\right) = o_p(1),
\end{aligned}
$$

where $\|\cdot\|_\infty$ denotes the absolute elementwise maximum, the third equality is due to Assumptions 3(i), (11.12), (11.11), and (11.10, the second last equality is due to Proposition 8, and the last equality is due to Assumption 2(ii). We now prove $n\kappa(W)|\tilde{G}_T - G| = o_p(1)$.

$$
\begin{aligned}
&n\kappa(W)|\tilde{G}_T - G| \\
&= n\kappa(W)|c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c \\
&\quad - c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+ H(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})HD_n^{+\intercal}WE(E^\intercal WE)^{-1}c| \\
&\leq n\kappa(W)\big|\text{maxeval}\left[(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})\right]\big| \, \|(\hat{H}_T - H)D_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2^2 \\
&\quad + n\kappa(W)\|(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})HD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2 \\
&\quad \cdot \|(\hat{H}_T - H)D_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2 \qquad (11.23)
\end{aligned}
$$

where the inequality is due to Lemma 5 in Appendix B. We consider the first term of (11.23) first.

$$
\begin{aligned}
&n\kappa(W)\big|\text{maxeval}\left[(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})\right]\big| \, \|(\hat{H}_T - H)D_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2^2 \\
&= O(n\kappa(W))\|\hat{H}_T - H\|_{\ell_2}^2\|D_n^{+\intercal}\|_{\ell_2}^2\|WE(E^\intercal WE)^{-1}\|_{\ell_2}^2 \\
&= O_p(n\kappa^2(W)/T) = o_p(1),
\end{aligned}
$$

where the second last equality is due to (11.12), (11.10), and (11.11), and the last equality is due to Assumption 2(ii). We now consider the second term of (11.23).

$$
\begin{aligned}
&n\kappa(W)\|(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})HD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2 \\
&\quad \cdot \|(\hat{H}_T - H)D_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2 \\
&\leq O(n\kappa(W))\|H\|_{\ell_2}\|\hat{H}_T - H\|_{\ell_2}\|D_n^{+\intercal}\|_{\ell_2}^2\|WE(E^\intercal WE)^{-1}c\|_2^2 = O(\sqrt{n\kappa^4(W)/T}) = o_p(1),
\end{aligned}
$$

where the first equality is due to (11.12), (11.10), and (11.11), and the last equality is due to Assumption 2(ii). We have proved $|n\kappa(W)\tilde{G}_T - n\kappa(W)G| = o_p(1)$ and hence $|n\kappa(W)\hat{G}_T - n\kappa(W)G| = o_p(1)$.

11.5.3 $t_2 = o_p(1)$

Last, we prove that $t_2 = o_p(1)$. Write

$$t_2 = \frac{\sqrt{T}\sqrt{n\kappa(W)}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\text{vec}O_p(\|M_T - \Theta\|_{\ell_2}^2)}{\sqrt{n\kappa(W)\hat{G}_T}}.$$

Since the denominator of the preceding equation is bounded away from zero by an absolute constant with probability approaching one by (11.20) and that $|n\kappa(W)\hat{G}_T - n\kappa(W)G| = o_p(1)$, it suffices to show

$$\sqrt{T}\sqrt{n\kappa(W)}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\text{vec}O_p(\|M_T - \Theta\|_{\ell_2}^2) = o_p(1).$$

This is straightforward:

$$\begin{aligned}
&|\sqrt{Tn\kappa(W)}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\text{vec}O_p(\|M_T - \Theta\|_{\ell_2}^2)| \\
&\leq \sqrt{Tn\kappa(W)}\|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\|_2\|\text{vec}O_p(\|M_T - \Theta\|_{\ell_2}^2)\|_2 \\
&= O(\sqrt{T}\kappa(W))\|O_p(\|M_T - \Theta\|_{\ell_2}^2)\|_F = O(\sqrt{Tn}\kappa(W))\|O_p(\|M_T - \Theta\|_{\ell_2}^2)\|_{\ell_2} \\
&= O(\sqrt{Tn}\kappa(W))O_p(\|M_T - \Theta\|_{\ell_2}^2) = O_p\left(\frac{\kappa(W)\sqrt{Tn}n}{T}\right) = O_p\left(\sqrt{\frac{n^3\kappa^2(W)}{T}}\right) = o_p(1),
\end{aligned}$$

where the last equality is due to Assumption 2(ii). □

11.6 Proof of Corollary 1

Proof of Corollary 1. Theorem 2 and a result we proved before, namely,

$$|\hat{G}_T - G| = |c^\intercal \hat{J}_T c - c^\intercal Jc| = o_p\left(\frac{1}{n\kappa(W)}\right), \tag{11.24}$$

imply

$$\sqrt{T}c^\intercal(\hat{\theta}_T - \theta^0) \xrightarrow{d} N(0, c^\intercal Jc). \tag{11.25}$$

Consider an arbitrary, non-zero vector $b \in \mathbb{R}^k$. Then

$$\left\|\frac{Ab}{\|Ab\|_2}\right\|_2 = 1,$$

so we can invoke (11.25) with $c = Ab/\|Ab\|_2$:

$$\sqrt{T}\frac{1}{\|Ab\|_2}b^\intercal A^\intercal(\hat{\theta}_T - \theta^0) \xrightarrow{d} N\left(0, \frac{b^\intercal A^\intercal}{\|Ab\|_2}J\frac{Ab}{\|Ab\|_2}\right),$$

which is equivalent to

$$\sqrt{T}b^\intercal A^\intercal(\hat{\theta}_T - \theta^0) \xrightarrow{d} N\left(0, b^\intercal A^\intercal JAb\right).$$

Since $b \in \mathbb{R}^k$ is non-zero and arbitrary, via the Cramer-Wold device, we have

$$\sqrt{T}A^\intercal(\hat{\theta}_T - \theta^0) \xrightarrow{d} N\left(0, A^\intercal JA\right).$$

Since we have shown in the paragraph above (11.20) that J is positive definite and A has full-column rank, $A^\intercal JA$ is positive definite and its negative square root exists. Hence,

$$\sqrt{T}(A^\intercal JA)^{-1/2}A^\intercal(\hat{\theta}_T - \theta^0) \xrightarrow{d} N\left(0, I_k\right).$$

Next from (11.24),

$$\left|b^\intercal Bb\right| := \left|b^\intercal A^\intercal \hat{J}_T Ab - b^\intercal A^\intercal JAb\right| = o_p\left(\frac{1}{n\kappa(W)}\right)\|Ab\|_2^2 \leq o_p\left(\frac{1}{n\kappa(W)}\right)\|A\|_{\ell_2}^2\|b\|_2^2.$$

By choosing $b = e_j$ where e_j is a vector in $\mathbb{R}^k$ with jth component being 1 and the rest of components being 0, we have for $j = 1, \ldots, k$

$$\left|B_{jj}\right| \leq o_p\left(\frac{1}{n\kappa(W)}\right)\|A\|_{\ell_2}^2 = o_p(1),$$

where the equality is due to $\|A\|_{\ell_2} = O_p(\sqrt{n\kappa(W)})$. By choosing $b = e_{ij}$, where e_{ij} is a vector in $\mathbb{R}^k$ with ith and jth components being $1/\sqrt{2}$ and the rest of components being 0, we have

$$\left|B_{ii}/2 + B_{jj}/2 + B_{ij}\right| \leq o_p\left(\frac{1}{n\kappa(W)}\right)\|A\|_{\ell_2}^2 = o_p(1).$$

Then

$$|B_{ij}| \leq |B_{ij} + B_{ii}/2 + B_{jj}/2| + |-(B_{ii}/2 + B_{jj}/2)| = o_p(1).$$

Thus we proved

$$B = A^\intercal \hat{J}_T A - A^\intercal JA = o_p(1),$$

because the dimension of the matrix B, k, is finite. By Slutsky's lemma

$$\sqrt{T}(A^\intercal \hat{J}_T A)^{-1/2}A^\intercal(\hat{\theta}_T - \theta^0) \xrightarrow{d} N\left(0, I_k\right).$$

□

11.7 Proof of Theorem 3

Proof of Theorem 3. At each step, we take the symmetry of $\Omega(\theta)$ into account.

$$
\begin{aligned}
&d\ell_T(\theta) \\
&= -\frac{T}{2} d \log \left| D^{1/2} \exp(\Omega) D^{1/2} \right| - \frac{T}{2} d \mathrm{tr} \left(\frac{1}{T} \sum_{t=1}^{T} (x_t - \mu)^\intercal D^{-1/2} [\exp(\Omega)]^{-1} D^{-1/2} (x_t - \mu) \right) \\
&= -\frac{T}{2} d \log \left| D^{1/2} \exp(\Omega) D^{1/2} \right| - \frac{T}{2} d \mathrm{tr} \left(D^{-1/2} \tilde{\Sigma} D^{-1/2} [\exp(\Omega)]^{-1} \right) \\
&= -\frac{T}{2} \mathrm{tr} \left(\left[D^{1/2} \exp(\Omega) D^{1/2} \right]^{-1} D^{1/2} d \exp(\Omega) D^{1/2} \right) - \frac{T}{2} d \mathrm{tr} \left(D^{-1/2} \tilde{\Sigma} D^{-1/2} [\exp(\Omega)]^{-1} \right) \\
&= -\frac{T}{2} \mathrm{tr} \left(\left[\exp(\Omega) \right]^{-1} d \exp(\Omega) \right) - \frac{T}{2} \mathrm{tr} \left(D^{-1/2} \tilde{\Sigma} D^{-1/2} d [\exp(\Omega)]^{-1} \right) \\
&= -\frac{T}{2} \mathrm{tr} \left(\left[\exp(\Omega) \right]^{-1} d \exp(\Omega) \right) + \frac{T}{2} \mathrm{tr} \left(D^{-1/2} \tilde{\Sigma} D^{-1/2} [\exp(\Omega)]^{-1} d \exp(\Omega) [\exp(\Omega)]^{-1} \right) \\
&= \frac{T}{2} \mathrm{tr} \left(\left\{ [\exp(\Omega)]^{-1} D^{-1/2} \tilde{\Sigma} D^{-1/2} [\exp(\Omega)]^{-1} - \left[\exp(\Omega) \right]^{-1} \right\} d \exp(\Omega) \right) \\
&= \frac{T}{2} \left[\mathrm{vec} \left(\left\{ [\exp(\Omega)]^{-1} D^{-1/2} \tilde{\Sigma} D^{-1/2} [\exp(\Omega)]^{-1} - \left[\exp(\Omega) \right]^{-1} \right\}^\intercal \right) \right]^\intercal \mathrm{vec} d \exp(\Omega) \\
&= \frac{T}{2} \left[\mathrm{vec} \left([\exp(\Omega)]^{-1} D^{-1/2} \tilde{\Sigma} D^{-1/2} [\exp(\Omega)]^{-1} - \left[\exp(\Omega) \right]^{-1} \right) \right]^\intercal \mathrm{vec} d \exp(\Omega),
\end{aligned}
$$

where in the second equality we used the definition of $\tilde{\Sigma}$ (5.2), the third equality is due to that $d \log |X| = \mathrm{tr}(X^{-1} dX)$, the fifth equality is due to that $dX^{-1} = -X^{-1}(dX)X^{-1}$, the seventh equality is due to that $\mathrm{tr}(AB) = (\mathrm{vec}[A^\intercal])^\intercal \mathrm{vec} B$, and the eighth equality is due to that matrix function preserves symmetry and we can interchange inverse and transpose operators. The following Frechet derivative of matrix exponential can be found in Higham (2008) p238:

$$
d \exp(\Omega) = \int_0^1 e^{(1-t)\Omega} (d\Omega) e^{t\Omega} dt.
$$

Therefore,

$$
\begin{aligned}
\mathrm{vec} d \exp(\Omega) &= \int_0^1 e^{t\Omega} \otimes e^{(1-t)\Omega} dt \mathrm{vec}(d\Omega) = \int_0^1 e^{t\Omega} \otimes e^{(1-t)\Omega} dt D_n \mathrm{vech}(d\Omega) \\
&= \int_0^1 e^{t\Omega} \otimes e^{(1-t)\Omega} dt D_n E d\theta,
\end{aligned}
$$

where the last equality is due to (4.2). Hence,

$$
\begin{aligned}
&d\ell_T(\theta) \\
&= \frac{T}{2} \left[\mathrm{vec} \left([\exp(\Omega)]^{-1} D^{-1/2} \tilde{\Sigma} D^{-1/2} [\exp(\Omega)]^{-1} - \left[\exp(\Omega) \right]^{-1} \right) \right]^\intercal \int_0^1 e^{t\Omega} \otimes e^{(1-t)\Omega} dt D_n E d\theta
\end{aligned}
$$

and

$$
\begin{aligned}
y &:= \frac{\partial \ell_T(\theta)}{\partial \theta^\intercal} \\
&= \frac{T}{2} E^\intercal D_n^\intercal \int_0^1 e^{t\Omega} \otimes e^{(1-t)\Omega} dt \left[\mathrm{vec}\left([\exp(\Omega)]^{-1} D^{-1/2} \tilde{\Sigma} D^{-1/2} [\exp(\Omega)]^{-1} - \left[\exp(\Omega)\right]^{-1} \right) \right] \\
&=: \frac{T}{2} E^\intercal D_n^\intercal \Psi_1 \Psi_2.
\end{aligned}
$$

Now we derive the Hessian matrix.

$$
dy = \frac{T}{2} E^\intercal D_n^\intercal (d\Psi_1)\Psi_2 + \frac{T}{2} E^\intercal D_n^\intercal \Psi_1 d\Psi_2 = \frac{T}{2} (\Psi_2^\intercal \otimes E^\intercal D_n^\intercal) \mathrm{vec} d\Psi_1 + \frac{T}{2} E^\intercal D_n^\intercal \Psi_1 d\Psi_2. \tag{11.26}
$$

Consider $d\Psi_1$ first.

$$
\begin{aligned}
d\Psi_1 &= d \int_0^1 e^{t\Omega} \otimes e^{(1-t)\Omega} dt = \int_0^1 de^{t\Omega} \otimes e^{(1-t)\Omega} dt + \int_0^1 e^{t\Omega} \otimes de^{(1-t)\Omega} dt \\
&=: \int_0^1 A \otimes e^{(1-t)\Omega} dt + \int_0^1 e^{t\Omega} \otimes B dt,
\end{aligned}
$$

where

$$
A := \int_0^1 e^{(1-s)t\Omega} d(t\Omega) e^{st\Omega} ds, \quad B := \int_0^1 e^{(1-s)(1-t)\Omega} d((1-t)\Omega) e^{s(1-t)\Omega} ds.
$$

Therefore,

$$
\begin{aligned}
\mathrm{vec} d\Psi_1 &= \int_0^1 \mathrm{vec}\left(A \otimes e^{(1-t)\Omega} \right) dt + \int_0^1 \mathrm{vec}\left(e^{t\Omega} \otimes B \right) dt \\
&= \int_0^1 P\left(\mathrm{vec} A \otimes \mathrm{vec} e^{(1-t)\Omega} \right) dt + \int_0^1 P\left(\mathrm{vec} e^{t\Omega} \otimes \mathrm{vec} B \right) dt \\
&= \int_0^1 P\left(I_{n^2} \otimes \mathrm{vec} e^{(1-t)\Omega} \right) \mathrm{vec} A dt + \int_0^1 P\left(\mathrm{vec} e^{t\Omega} \otimes I_{n^2} \right) \mathrm{vec} B dt \\
&= \int_0^1 P\left(I_{n^2} \otimes \mathrm{vec} e^{(1-t)\Omega} \right) \int_0^1 e^{st\Omega} \otimes e^{(1-s)t\Omega} ds \cdot \mathrm{vec} d(t\Omega) dt \\
&\quad + \int_0^1 P\left(\mathrm{vec} e^{t\Omega} \otimes I_{n^2} \right) \int_0^1 e^{s(1-t)\Omega} \otimes e^{(1-s)(1-t)\Omega} ds \cdot \mathrm{vec} d((1-t)\Omega) dt \\
&= \int_0^1 P\left(I_{n^2} \otimes \mathrm{vec} e^{(1-t)\Omega} \right) \int_0^1 e^{st\Omega} \otimes e^{(1-s)t\Omega} ds \cdot t dt D_n E d\theta \\
&\quad + \int_0^1 P\left(\mathrm{vec} e^{t\Omega} \otimes I_{n^2} \right) \int_0^1 e^{s(1-t)\Omega} \otimes e^{(1-s)(1-t)\Omega} ds \cdot (1-t) dt D_n E d\theta
\end{aligned} \tag{11.27}
$$

where $P := I_n \otimes K_{n,n} \otimes I_n$, the second equality is due to Lemma 6 in Appendix B. We now consider $d\Psi_2$.

$$
\begin{aligned}
d\Psi_2 &= d\text{vec}\left([\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega)]^{-1} - \left[\exp(\Omega)\right]^{-1}\right) \\
&= \text{vec}\left(d[\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega)]^{-1}\right) \\
&\quad + \left([\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}d[\exp(\Omega)]^{-1}\right) - \text{vec}\left(d\left[\exp(\Omega)\right]^{-1}\right) \\
&= \text{vec}\left(-[\exp(\Omega)]^{-1}d\exp(\Omega)[\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega)]^{-1}\right) \\
&\quad + \text{vec}\left(-[\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega)]^{-1}d\exp(\Omega)[\exp(\Omega)]^{-1}\right) \\
&\quad + \text{vec}\left([\exp(\Omega)]^{-1}d\exp(\Omega)[\exp(\Omega)]^{-1}\right) \\
&= \left([\exp(\Omega)]^{-1} \otimes [\exp(\Omega)]^{-1}\right)\text{vec}d\exp(\Omega) \\
&\quad - \left([\exp(\Omega)]^{-1} \otimes [\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega)]^{-1}\right)\text{vec}d\exp(\Omega) \\
&\quad - \left([\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega)]^{-1} \otimes [\exp(\Omega)]^{-1}\right)\text{vec}d\exp(\Omega) \qquad (11.28)
\end{aligned}
$$

Substituting (11.27) and (11.28) into (11.26) yields the result:

$$
\begin{aligned}
&\frac{\partial^2 \ell_T(\theta)}{\partial\theta\partial\theta^\intercal} = \\
&-\frac{T}{2}E^\intercal D_n^\intercal \Psi_1 \left([\exp\Omega]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2} \otimes I_n + I_n \otimes [\exp\Omega]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2} - I_{n^2}\right) \cdot \\
&\quad \left([\exp\Omega]^{-1} \otimes [\exp\Omega]^{-1}\right)\Psi_1 D_n E \\
&\quad + \frac{T}{2}(\Psi_2^\intercal \otimes E^\intercal D_n^\intercal)\int_0^1 P\left(I_{n^2} \otimes \text{vec}e^{(1-t)\Omega}\right)\int_0^1 e^{st\Omega} \otimes e^{(1-s)t\Omega}ds \cdot tdtD_nE \\
&\quad + \frac{T}{2}(\Psi_2^\intercal \otimes E^\intercal D_n^\intercal)\int_0^1 P\left(\text{vec}e^{t\Omega} \otimes I_{n^2}\right)\int_0^1 e^{s(1-t)\Omega} \otimes e^{(1-s)(1-t)\Omega}ds \cdot (1-t)dtD_nE.
\end{aligned}
$$

□

11.8 Proof of Theorem 4

Under Assumptions 3 - 4 and Proposition 4(i), $\Theta^{-1} \otimes \Theta^{-1}$ and $M_T^{-1} \otimes M_T^{-1}$ are positive definite for all n with minimum eigenvalues bounded away from zero by absolute constants and maximum eigenvalues bounded from above by absolute constants (with probability approaching 1 for $M_T^{-1} \otimes M_T^{-1}$) , so their unique positive definite square roots $\Theta^{-1/2} \otimes \Theta^{-1/2}$ and $M_T^{-1/2} \otimes M_T^{-1/2}$ exist, whose minimum eigenvalues also bounded away from zero by absolute constants and maximum eigenvalues bounded from above by absolute constants. Define

$$
\mathcal{X} := (\Theta^{-1/2} \otimes \Theta^{-1/2})\Psi_1(\theta^0)D_nE, \qquad \hat{\mathcal{X}}_T := (M_T^{-1/2} \otimes M_T^{-1/2})\hat{\Psi}_{1,T}D_nE.
$$

Therefore

$$
\Upsilon = -\frac{1}{2}\mathcal{X}^\intercal\mathcal{X}, \qquad \hat{\Upsilon}_T = -\frac{1}{2}\hat{\mathcal{X}}_T^\intercal\hat{\mathcal{X}}_T.
$$

Proposition 9. *Suppose Assumptions 1(i), 2(i), 3 and 4 hold. Then $\Psi_1 = \Psi_1(\theta^0)$ is positive definite for all n with its minimum eigenvalue bounded away from zero by an absolute constant and maximum eigenvalue bounded from above by an absolute constant. $\hat{\Psi}_{1,T}$ is positive definite for all n with its minimum eigenvalue bounded away from zero by an absolute constant and maximum eigenvalue bounded from above by an absolute constant with probability approaching 1.*

Proof. Since the proofs for the sample analogue and population are exactly the same, we only give a proof for the sample analogue. The idea is to re-express $\hat{\Psi}_{1,T}$ into the diagonalised form, as in Linton and McCrorie (1995):

$$\hat{\Psi}_{1,T} = \int_0^1 e^{t \log M_T} \otimes e^{(1-t) \log M_T} dt = \int_0^1 \left(\sum_{k=0}^{\infty} \frac{1}{k!} t^k \log^k M_T \right) \otimes \left(\sum_{l=0}^{\infty} \frac{1}{l!} (1-t)^l \log^l M_T \right) dt$$

$$= \int_0^1 \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \frac{t^k (1-t)^l}{k!l!} (\log^k M_T \otimes \log^l M_T) dt = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} (\log^k M_T \otimes \log^l M_T) \frac{1}{k!l!} \int_0^1 t^k (1-t)^l dt$$

$$= \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \frac{1}{(k+l+1)!} (\log^k M_T \otimes \log^l M_T) = \sum_{n=0}^{\infty} \frac{1}{(n+1)!} \sum_{l=0}^{n} \log^{n-l} M_T \otimes \log^l M_T,$$

where the fourth equality is true because the infinite series is absolutely convergent (infinite radius of convergence) so we can interchange $\sum$ and $\int$, the fifth equality is due to Lemma 7 in Appendix B. Suppose that M_T has eigenvalues $\lambda_1, \ldots, \lambda_n$, then $\log M_T$ has eigenvalues $\log \lambda_1, \ldots, \log \lambda_n$ (Higham (2008) p10). Suppose that $\log M_T = Q^\intercal \Xi Q$ (orthogonal diagonalization). Then

$$\log^{n-l} M_T \otimes \log^l M_T = (Q^\intercal \Xi^{n-l} Q) \otimes (Q^\intercal \Xi^l Q) = (Q^\intercal \otimes Q^\intercal)(\Xi^{n-l} \otimes \Xi^l)(Q \otimes Q).$$

The matrix $\sum_{l=0}^{n} \Xi^{n-l} \otimes \Xi^l$ is a $n^2 \times n^2$ diagonal matrix with the $[(i-1)n+j]$th entry equal to

$$\begin{cases} \sum_{l=0}^{n} (\log \lambda_i)^{n-l} (\log \lambda_j)^l = \frac{(\log \lambda_i)^{n+1} - (\log \lambda_j)^{n+1}}{\log \lambda_i - \log \lambda_j} & \text{if } i \neq j, \lambda_i \neq \lambda_j \\ (n+1)(\log \lambda_i)^n & \text{if } i \neq j, \lambda_i = \lambda_j \\ (n+1)(\log \lambda_i)^n & \text{if } i = j \end{cases}$$

for $i, j = 1, \ldots, n$. Therefore $\hat{\Psi}_{1,T} = (Q^\intercal \otimes Q^\intercal)[\sum_{n=0}^{\infty} \frac{1}{(n+1)!} \sum_{l=0}^{n} (\Xi^{n-l} \otimes \Xi^l)](Q \otimes Q)$ whose $[(i-1)n+j]$th eigenvalue equal to

$$\begin{cases} \frac{\exp(\log \lambda_i) - \exp(\log \lambda_j)}{\log \lambda_i - \log \lambda_j} = \frac{\lambda_i - \lambda_j}{\log \lambda_i - \log \lambda_j} & \text{if } i \neq j, \lambda_i \neq \lambda_j \\ \exp \log \lambda_i = \lambda_i & \text{if } i \neq j, \lambda_i = \lambda_j \\ \exp \log \lambda_i = \lambda_i & \text{if } i = j \end{cases}$$

The proposition then follows from the assumptions of the proposition. □

Proposition 10. *For any $(v+1) \times 1$ non-zero vector b, with $\|b\|_2 = 1$,*

$$\|b^\intercal (\mathcal{X}^\intercal \mathcal{X})^{-1} \mathcal{X}^\intercal\|_2 = O\left(\frac{1}{\sqrt{n}}\right).$$

Proof. Note that

$$\|b^\intercal(\mathcal{X}^\intercal\mathcal{X})^{-1}\mathcal{X}^\intercal\|_2^2 = b^\intercal(\mathcal{X}^\intercal\mathcal{X})^{-1}b \leq \text{maxeval}(\mathcal{X}^\intercal\mathcal{X})^{-1} = \frac{1}{\text{mineval}(\mathcal{X}^\intercal\mathcal{X})}.$$

Note that for any $(v+1)\times 1$ a with $\|a\| = 1$

$$\begin{aligned} &a^\intercal\mathcal{X}^\intercal\mathcal{X}a = a^\intercal E^\intercal D_n^\intercal \Psi_1 \left(\Theta^{-1}\otimes\Theta^{-1}\right)\Psi_1 D_n E a \\ &\geq \text{mineval}(\Theta^{-1}\otimes\Theta^{-1})\text{mineval}(\Psi_1^2)\text{mineval}(D_n^\intercal D_n)\text{mineval}(E^\intercal E) \geq Cn, \end{aligned}$$

for some positive constant C. □

Proposition 11. *Let Assumptions 1(i), 2(i), 3 and 4 be satisfied. Then*

(i)

$$\|M_T^{-1}\otimes M_T^{-1} - \Theta^{-1}\otimes\Theta^{-1}\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right).$$

(ii)

$$\|M_T^{-1/2}\otimes M_T^{-1/2} - \Theta^{-1/2}\otimes\Theta^{-1/2}\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right).$$

Proof. For (i)

$$\begin{aligned} &\|M_T^{-1}\otimes M_T^{-1} - \Theta^{-1}\otimes\Theta^{-1}\|_{\ell_2} \\ &= \|M_T^{-1}\otimes M_T^{-1} - M_T^{-1}\otimes\Theta^{-1} + M_T^{-1}\otimes\Theta^{-1} - \Theta^{-1}\otimes\Theta^{-1}\|_{\ell_2} \\ &= \|M_T^{-1}\otimes(M_T^{-1}-\Theta^{-1}) + (M_T^{-1}-\Theta^{-1})\otimes\Theta^{-1}\|_{\ell_2} \\ &\leq \|M_T^{-1}\|_{\ell_2}\|M_T^{-1}-\Theta^{-1}\|_{\ell_2} + \|M_T^{-1}-\Theta^{-1}\|_{\ell_2}\|\Theta^{-1}\|_{\ell_2} \\ &= (\|M_T^{-1}\|_{\ell_2} + \|\Theta^{-1}\|_{\ell_2})\|M_T^{-1}-\Theta^{-1}\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right) \end{aligned}$$

where the inequality is due to Proposition 16 in Appendix B, and the last equality is due to Lemma 3 in Appendix B given Proposition 4(i) and Assumption 2(i). For part (ii), invoke Lemma 4 in Appendix B:

$$\begin{aligned} &\|M_T^{-1/2}\otimes M_T^{-1/2} - \Theta^{-1/2}\otimes\Theta^{-1/2}\|_{\ell_2} \leq \\ &\frac{\|M_T^{-1}\otimes M_T^{-1} - \Theta^{-1}\otimes\Theta^{-1}\|_{\ell_2}}{\text{mineval}(M_T^{-1/2}\otimes M_T^{-1/2}) + \text{mineval}(\Theta^{-1/2}\otimes\Theta^{-1/2})} = O_p\left(\sqrt{\frac{n}{T}}\right). \end{aligned}$$

□

Proposition 12. *Let Assumptions 1(i), 2(i), 3 and 4 be satisfied. Then*

$$\|\hat{\Psi}_{1,T} - \Psi_1\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right).$$

Proof.

$$\begin{aligned}
\|\hat{\Psi}_{1,T} - \Psi_1\|_{\ell_2} &= \left\| \int_0^1 (M_T^t \otimes M_T^{1-t} - \Theta^t \otimes \Theta^{1-t}) dt \right\|_{\ell_2} \\
&\leq \int_0^1 \|M_T^t \otimes M_T^{1-t} - \Theta^t \otimes \Theta^{1-t}\|_{\ell_2} dt \\
&\leq \int_0^1 \left\|M_T^t \otimes (M_T^{1-t} - \Theta^{1-t})\right\|_{\ell_2} dt + \int_0^1 \left\|(M_T^t - \Theta^t) \otimes \Theta^{1-t}\right\|_{\ell_2} dt \\
&\leq \max_{t\in[0,1]} \left\|M_T^t\right\|_{\ell_2} \left\|M_T^{1-t} - \Theta^{1-t}\right\|_{\ell_2} + \max_{t\in[0,1]} \left\|M_T^t - \Theta^t\right\| \left\|\Theta^{1-t}\right\|_{\ell_2} \\
&= \max_{t\in[0,1]} \left(\|M_T^{1-t}\|_{\ell_2} + \|\Theta^{1-t}\|_{\ell_2}\right) \|M_T^t - \Theta^t\|_{\ell_2}.
\end{aligned}$$

The lemma follows trivially for the boundary cases $t = 0$ and $t = 1$, so we only need to consider the case $t \in (0, 1)$. We first show that for any $t \in (0, 1)$, $\|M_T^{1-t}\|_{\ell_2}$ and $\|\Theta^{1-t}\|_{\ell_2}$ are $O_p(1)$. This is obvious: diagonalize Θ, apply the function $f(x) = x^{1-t}$, and take the spectral norm. The lemma would then follow if we show that $\max_{t\in(0,1)} \|M_T^t - \Theta^t\|_{\ell_2} = O_p(\sqrt{n/T})$.

$$\begin{aligned}
\|M_T^t - \Theta^t\|_{\ell_2} &= \left\|e^{t \log M_T} - e^{t \log \Theta}\right\| \\
&\leq \|t(\log M_T - \log \Theta)\|_{\ell_2} \exp[t\| \log M_T - \log \Theta\|_{\ell_2}] \exp[t\| \log \Theta\|_{\ell_2}] \\
&= \|t(\log M_T - \log \Theta)\|_{\ell_2} \exp[t\| \log M_T - \log \Theta\|_{\ell_2}] O(1),
\end{aligned}$$

where the first inequality is due to Theorem 11 in Appendix B, and the second equality is due to the fact that all the eigenvalues of Θ are bounded away from zero and infinity by absolute constants. Now use (11.8):

$$\begin{aligned}
&\| \log M_T - \log \Theta\|_{\ell_2} \leq \max_{t\in[0,1]} \|[t(\Theta - I) + I]^{-1}\|_{\ell_2}^2 \|M_T - \Theta\|_{\ell_2} + O_p(\|M_T - \Theta\|_{\ell_2}^2) \\
&= O_p(\|M_T - \Theta\|_{\ell_2}) + O_p(\|M_T - \Theta\|_{\ell_2}^2) = O_p\left(\sqrt{\frac{n}{T}}\right)
\end{aligned}$$

where the first inequality is due to the triangular inequality and the submultiplicative property of matrix norm, the first equality is due to the minimum eigenvalue of $t\Theta+(1-t)I$ is bounded away from zero by an absolute constant for any $t \in (0, 1)$, and the last equality is due to Proposition 4(i). The result follows after recognising $\exp(o_p(1)) = O_p(1)$. □

Proposition 13. *Let Assumptions 1(i), 2(i), 3 and 4 be satisfied. Then*

(i)

$$\|\hat{\mathcal{X}}_T\|_{\ell_2} = \|\hat{\mathcal{X}}_T^\intercal\|_{\ell_2} = O_p(\sqrt{n}), \quad \|\mathcal{X}\|_{\ell_2} = \|\mathcal{X}^\intercal\|_{\ell_2} = O(\sqrt{n}).$$

(ii)

$$\|\hat{\mathcal{X}}_T - \mathcal{X}\|_{\ell_2} = O_p\left(\sqrt{\frac{n^2}{T}}\right).$$

(iii)

$$\left\|\frac{\hat{\Upsilon}_T}{n} - \frac{\Upsilon}{n}\right\|_{\ell_2} = \left\|\frac{\hat{\mathcal{X}}_T^\intercal \hat{\mathcal{X}}_T}{2n} - \frac{\mathcal{X}^\intercal \mathcal{X}}{2n}\right\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right).$$

(iv)

$$\left\|n\hat{\Upsilon}_T^{-1} - n\Upsilon^{-1}\right\|_{\ell_2} = \left\|2n(\hat{\mathcal{X}}_T^\intercal\hat{\mathcal{X}}_T)^{-1} - 2n(\mathcal{X}^\intercal\mathcal{X})^{-1}\right\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right).$$

Proof. For part (i), it suffices to give a proof for $\|\hat{\mathcal{X}}_T\|_{\ell_2}$.

$$\begin{aligned}\|\hat{\mathcal{X}}_T\|_{\ell_2} &= \|(M_T^{-1/2}\otimes M_T^{-1/2})\hat{\Psi}_{1,T}D_nE\|_{\ell_2} \le \|M_T^{-1/2}\otimes M_T^{-1/2}\|_{\ell_2}\|\hat{\Psi}_{1,T}\|_{\ell_2}\|D_n\|_{\ell_2}\|E\|_{\ell_2}\\ &= O_p(\sqrt{n}),\end{aligned}$$

where the last equality is due to Propositions 2(iii) and 9 and (11.10). Now

$$\begin{aligned}\|\hat{\mathcal{X}}_T - \mathcal{X}\|_{\ell_2} &= \|(M_T^{-1/2}\otimes M_T^{-1/2})\hat{\Psi}_{1,T}D_nE - (\Theta^{-1/2}\otimes\Theta^{-1/2})\Psi_1D_nE\|_{\ell_2}\\ &\le \|(M_T^{-1/2}\otimes M_T^{-1/2} - \Theta^{-1/2}\otimes\Theta^{-1/2})\hat{\Psi}_{1,T}D_nE\|_{\ell_2}\\ &\quad + \|(\Theta^{-1/2}\otimes\Theta^{-1/2})(\hat{\Psi}_{1,T}-\Psi_1)D_nE\|_{\ell_2}\\ &\le \|(M_T^{-1/2}\otimes M_T^{-1/2} - \Theta^{-1/2}\otimes\Theta^{-1/2})\|_{\ell_2}\|\hat{\Psi}_{1,T}\|_{\ell_2}\|D_n\|_{\ell_2}\|E\|_{\ell_2}\\ &\quad + \|\Theta^{-1/2}\otimes\Theta^{-1/2}\|_{\ell_2}\|\hat{\Psi}_{1,T}-\Psi_1\|_{\ell_2}\|D_n\|_{\ell_2}\|E\|_{\ell_2}.\end{aligned}$$

The proposition result (ii) follows after invoking Propositions 11 and 12. For part (iii),

$$\|\hat{\mathcal{X}}_T^\intercal\hat{\mathcal{X}}_T - \mathcal{X}^\intercal\mathcal{X}\|_{\ell_2} = \|\hat{\mathcal{X}}_T^\intercal\hat{\mathcal{X}}_T - \hat{\mathcal{X}}_T^\intercal\mathcal{X} + \hat{\mathcal{X}}_T^\intercal\mathcal{X} - \mathcal{X}^\intercal\mathcal{X}\|_{\ell_2} \le \|\hat{\mathcal{X}}_T^\intercal(\hat{\mathcal{X}}_T-\mathcal{X})\|_{\ell_2} + \|(\hat{\mathcal{X}}_T-\mathcal{X})^\intercal\mathcal{X}\|_{\ell_2}.$$

Therefore part (iii) follows from parts (i) and (ii). Part (iv) follows from result (iii) via Lemma 3 in Appendix B and the fact that $\|2n(\mathcal{X}^\intercal\mathcal{X})^{-1}\|_{\ell_2} = O(1)$. □

Proof of Theorem 4. We first show that $\hat{\Upsilon}_T$ is invertible with probability approaching 1, so that our estimator $\tilde{\theta}_T := \hat{\theta}_T + (-\hat{\Upsilon}_T)^{-1}\frac{\partial\ell_T(\hat{\theta}_T)}{\partial\theta^\intercal}/T$ is well defined. It suffices to show that $-\hat{\Upsilon}_T = \frac{1}{2}E^\intercal D_n^\intercal\hat{\Psi}_{1,T}\left(M_T^{-1}\otimes M_T^{-1}\right)\hat{\Psi}_{1,T}D_nE$ has minimum eigenvalue bounded away from zero by an absolute constant with probability approaching one. For any $(v+1)\times 1$ vector a with $\|a\|_2 = 1$,

$$\begin{aligned}&a^\intercal E^\intercal D_n^\intercal\hat{\Psi}_{1,T}\left(M_T^{-1}\otimes M_T^{-1}\right)\hat{\Psi}_{1,T}D_nEa/2\\ &\ge \text{mineval}(M_T^{-1}\otimes M_T^{-1})\text{mineval}(\hat{\Psi}_{1,T}^2)\text{mineval}(D_n^\intercal D_n)\text{mineval}(E^\intercal E)/2 \ge Cn,\end{aligned}$$

for some absolute constant C with probability approaching one. Hence $-\hat{\Upsilon}_T$ has minimum eigenvalue bounded away from zero by an absolute constant with probability approaching one. Also as a by-product

$$\|(-\hat{\Upsilon}_T)^{-1}\|_{\ell_2} = \frac{1}{\text{mineval}(-\hat{\Upsilon}_T)} = O_p(n^{-1}). \tag{11.29}$$

From the definition of $\tilde{\theta}_T$, for any $b\in\mathbb{R}^{v+1}$ with $\|b\|_2 = 1$ we can write

$$\begin{aligned}\sqrt{T}b^\intercal(-\hat{\Upsilon}_T)(\tilde{\theta}_T-\theta^0) &= \sqrt{T}b^\intercal(-\hat{\Upsilon}_T)(\hat{\theta}_T-\theta^0) + \sqrt{T}b^\intercal\frac{1}{T}\frac{\partial\ell_T(\hat{\theta}_T)}{\partial\theta^\intercal}\\ &= \sqrt{T}b^\intercal(-\hat{\Upsilon}_T)(\hat{\theta}_T-\theta^0) + \sqrt{T}b^\intercal\frac{1}{T}\frac{\partial\ell_T(\theta^0)}{\partial\theta^\intercal} + \sqrt{T}b^\intercal\Upsilon(\hat{\theta}_T-\theta^0) + o_p(1)\\ &= \sqrt{T}b^\intercal(\Upsilon-\hat{\Upsilon}_T)(\hat{\theta}_T-\theta^0) + b^\intercal\sqrt{T}\frac{1}{T}\frac{\partial\ell_T(\theta^0)}{\partial\theta^\intercal} + o_p(1)\end{aligned}$$

where the second equality is due to Assumption 6 and the fact that $\hat{\theta}_T$ is $\sqrt{T/(n\kappa(W))}$-consistent. Defining $a^\intercal = b^\intercal(-\hat{\Upsilon}_T)$, we write

$$\sqrt{T}\frac{a^\intercal}{\|a\|_2}(\tilde{\theta}_T-\theta^0) = \sqrt{T}\frac{a^\intercal}{\|a\|_2}(-\hat{\Upsilon}_T)^{-1}(\Upsilon-\hat{\Upsilon}_T)(\hat{\theta}_T-\theta^0)+\frac{a^\intercal}{\|a\|_2}(-\hat{\Upsilon}_T)^{-1}\sqrt{T}\frac{1}{T}\frac{\partial \ell_T(\theta^0)}{\partial \theta^\intercal}+\frac{o_p(1)}{\|a\|_2}.$$

By recognising that $\|a^\intercal\|_2 = \|b^\intercal(-\hat{\Upsilon}_T)\|_2 \geq \text{mineval}(-\hat{\Upsilon}_T)$, we have

$$\frac{1}{\|a\|_2} = O_p(n^{-1}).$$

Thus without loss of generality, we have

$$\sqrt{T}b^\intercal(\tilde{\theta}_T - \theta^0) = \sqrt{T}b^\intercal(-\hat{\Upsilon}_T)^{-1}(\Upsilon - \hat{\Upsilon}_T)(\hat{\theta}_T - \theta^0) + b^\intercal(-\hat{\Upsilon}_T)^{-1}\sqrt{T}\frac{1}{T}\frac{\partial \ell_T(\theta^0)}{\partial \theta^\intercal} + o_p(n^{-1}).$$

We now show that the first term on the right side is $o_p(n^{-1/2})$. This is straightforward

$$\begin{aligned}&\sqrt{T}|b^\intercal(-\hat{\Upsilon}_T)^{-1}(\Upsilon - \hat{\Upsilon}_T)(\hat{\theta}_T - \theta^0)| \leq \sqrt{T}\|b\|_2\|(-\hat{\Upsilon}_T)^{-1}\|_{\ell_2}\|\Upsilon - \hat{\Upsilon}_T\|_{\ell_2}\|\hat{\theta}_T - \theta^0\|_2 \\ &= \sqrt{T}O_p(n^{-1})nO_p(\sqrt{n/T})O_p(\sqrt{n\kappa(W)/T}) = O_p(\sqrt{n^3\kappa(W)/T}n^{-1/2}) = o_p(n^{-1/2}),\end{aligned}$$

where the first equality is due to (11.29), Proposition 13 (iii) and Theorem 1, and the last equation is due to Assumption 2(ii). Thus

$$\sqrt{T}b^\intercal(\tilde{\theta}_T - \theta^0) = -b^\intercal\hat{\Upsilon}_T^{-1}\sqrt{T}\frac{1}{T}\frac{\partial \ell_T(\theta^0)}{\partial \theta^\intercal} + o_p(n^{-1/2}),$$

whence, if we divide by $\sqrt{b^\intercal(-\hat{\Upsilon}_T)^{-1}b}$, we have

$$\begin{aligned}\frac{\sqrt{T}b^\intercal(\tilde{\theta}_T - \theta^0)}{\sqrt{b^\intercal(-\hat{\Upsilon}_T)^{-1}b}} &= \frac{-b^\intercal\hat{\Upsilon}_T^{-1}\sqrt{T}\frac{\partial \ell_T(\theta^0)}{\partial \theta^\intercal}/T}{\sqrt{b^\intercal(-\hat{\Upsilon}_T)^{-1}b}} + \frac{o_p(n^{-1/2})}{\sqrt{b^\intercal(-\hat{\Upsilon}_T)^{-1}b}} \\ &=: t_{2,1} + t_{2,2}.\end{aligned}$$

Define

$$t'_{2,1} := \frac{-b^\intercal\Upsilon^{-1}\sqrt{T}\frac{\partial \ell_T(\theta^0)}{\partial \theta^\intercal}/T}{\sqrt{b^\intercal(-\Upsilon)^{-1}b}}.$$

To prove Theorem 4, it suffices to show $t'_{2,1} \xrightarrow{d} N(0,1)$, $t'_{2,1} - t_{2,1} = o_p(1)$, and $t_{2,2} = o_p(1)$.

11.8.1 $t'_{2,1} \xrightarrow{d} N(0,1)$

We now prove that $t'_{2,1}$ is asymptotically distributed as a standard normal.

$$\begin{aligned}&t'_{2,1} = \frac{-b^\intercal\Upsilon^{-1}\sqrt{T}\frac{\partial \ell_T(\theta^0)}{\partial \theta^\intercal}/T}{\sqrt{b^\intercal(-\Upsilon)^{-1}b}} = \\ &\sum_{t=1}^{T}\frac{b^\intercal(\mathcal{X}^\intercal\mathcal{X})^{-1}\mathcal{X}^\intercal(\Theta^{-1/2}\otimes\Theta^{-1/2})(D^{-1/2}\otimes D^{-1/2})T^{-1/2}\text{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal - \mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]}{\sqrt{b^\intercal(-\Upsilon)^{-1}b}} \\ &=: \sum_{t=1}^{T}U_{T,n,t}.\end{aligned}$$

The proof is very similar to that of $t_1' \xrightarrow{d} N(0,1)$ in Section 11.5.1. It is not difficult to show $\mathbb{E}[U_{T,n,t}] = 0$ and $\sum_{t=1}^{T} \mathbb{E}[U_{T,n,t}^2] = 1$. Then we just need to verify the following Lindeberg condition for a double indexed process: for all $\varepsilon > 0$,

$$\lim_{n,T\to\infty} \sum_{t=1}^{T} \int_{\{|U_{T,n,t}|\geq\varepsilon\}} U_{T,n,t}^2 dP = 0.$$

For any $\gamma > 2$,

$$\int_{\{|U_{T,n,t}|\geq\varepsilon\}} U_{T,n,t}^2 dP = \int_{\{|U_{T,n,t}|\geq\varepsilon\}} U_{T,n,t}^2 |U_{T,n,t}|^{-\gamma} |U_{T,n,t}|^{\gamma} dP \leq \varepsilon^{2-\gamma} \int_{\{|U_{T,n,t}|\geq\varepsilon\}} |U_{T,n,t}|^{\gamma} dP$$
$$\leq \varepsilon^{2-\gamma} \mathbb{E}|U_{T,n,t}|^{\gamma},$$

We first investigate that at what rate the denominator $\sqrt{b^\intercal(-\Upsilon)^{-1}b}$ goes to zero.

$$\begin{aligned} b^\intercal(-\Upsilon)^{-1}b &= 2b^\intercal \left(E^\intercal D_n^\intercal \Psi_1 \left(\Theta^{-1} \otimes \Theta^{-1}\right) \Psi_1 D_n E\right)^{-1} b \\ &\geq 2\text{mineval}\left(\left(E^\intercal D_n^\intercal \Psi_1 \left(\Theta^{-1} \otimes \Theta^{-1}\right) \Psi_1 D_n E\right)^{-1}\right) \\ &= \frac{2}{\text{maxeval}\left(E^\intercal D_n^\intercal \Psi_1 \left(\Theta^{-1} \otimes \Theta^{-1}\right) \Psi_1 D_n E\right)}. \end{aligned}$$

For an arbitrary $(v+1) \times 1$ vector a with $\|a\|_2 = 1$, we have

$$\begin{aligned} &a^\intercal E^\intercal D_n^\intercal \Psi_1 \left(\Theta^{-1} \otimes \Theta^{-1}\right) \Psi_1 D_n E a \\ &\leq \text{maxeval}(\Theta^{-1} \otimes \Theta^{-1})\text{maxeval}(\Psi_1^2)\text{maxeval}(D_n^\intercal D_n)\text{maxeval}(E^\intercal E) \leq Cn, \end{aligned}$$

for some constant C. Thus we have

$$\frac{1}{\sqrt{b^\intercal(-\Upsilon)^{-1}b}} = O(\sqrt{n}). \tag{11.30}$$

Then a sufficient condition for the Lindeberg condition is:

$T^{1-\frac{\gamma}{2}} n^{\gamma/2}$.

$$\begin{aligned} &\mathbb{E}\left|b^\intercal(\mathcal{X}^\intercal\mathcal{X})^{-1}\mathcal{X}^\intercal(\Theta^{-1/2} \otimes \Theta^{-1/2})(D^{-1/2} \otimes D^{-1/2})\text{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal - \mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right|^{\gamma} \\ &= o(1), \end{aligned} \tag{11.31}$$

for some $\gamma > 2$. We now verify (11.31). We shall be concise as the proof is very similar to that in Section 11.5.1.

$$\begin{aligned} &\mathbb{E}\left|b^\intercal(\mathcal{X}^\intercal\mathcal{X})^{-1}\mathcal{X}^\intercal(\Theta^{-1/2} \otimes \Theta^{-1/2})(D^{-1/2} \otimes D^{-1/2})\text{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal - \mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right|^{\gamma} \\ &\lesssim \|b^\intercal(\mathcal{X}^\intercal\mathcal{X})^{-1}\mathcal{X}^\intercal\|_2^{\gamma}\mathbb{E}\|(x_t-\mu)(x_t-\mu)^\intercal\|_F^{\gamma} = O\left(n^{-\gamma/2}\right) n^{\gamma} \left\| \max_{1\leq i,j\leq n} \left|(x_t-\mu)_i(x_t-\mu)_j\right| \right\|_{L_\gamma}^{\gamma} \\ &= O\left(n^{-\gamma/2}\right) n^{\gamma} O(\log^{\gamma} n), \end{aligned}$$

where the second last equality is due to Proposition 10 and the last equality is due to (11.22). Summing up the rates, we have

$$T^{1-\frac{\gamma}{2}} n^{\gamma/2} O\left(n^{-\gamma/2}\right) n^{\gamma} O(\log^{\gamma} n) = o\left(\frac{n \log n}{T^{\frac{1}{2}-\frac{1}{\gamma}}}\right)^{\gamma} = o(1),$$

by Assumption 2(ii). Thus, we have verified (11.31).

11.8.2 $t'_{2,1} - t_{2,1} = o_p(1)$

Let A and $\hat{A}$ denote the numerators of $t'_{2,1}$ and $t_{2,1}$, respectively. Let $\sqrt{G}$ and $\sqrt{\hat{G}}$ denote the denominators of $t'_{2,1}$ and $t_{2,1}$, respectively. Write

$$\begin{aligned}
t'_{2,1} - t_{2,1} &= \frac{\sqrt{n}A}{\sqrt{nG}} - \frac{\sqrt{n}\hat{A}}{\sqrt{nG}} + \frac{\sqrt{n}\hat{A}}{\sqrt{nG}} - \frac{\sqrt{n}\hat{A}}{\sqrt{n\hat{G}}} \\
&= \frac{1}{\sqrt{nG}}(\sqrt{n}A - \sqrt{n}\hat{A}) + \sqrt{n}\hat{A}\left(\frac{1}{\sqrt{nG}} - \frac{1}{\sqrt{n\hat{G}}}\right) \\
&= \frac{1}{\sqrt{nG}}(\sqrt{n}A - \sqrt{n}\hat{A}) + \sqrt{n}\hat{A}\frac{1}{\sqrt{nG}\sqrt{n\hat{G}}}\frac{n\hat{G} - nG}{\sqrt{n\hat{G}} + \sqrt{nG}}.
\end{aligned}$$

Note that we have shown in (11.30) that $\sqrt{nG}$ is uniformly (in n) bounded away from zero, that is, $1/\sqrt{nG} = O(1)$. Also we have shown that $t'_{2,1} = A/\sqrt{G} = O_p(1)$. Hence

$$\sqrt{n}A = \sqrt{n}O_p(\sqrt{G}) = \sqrt{n}O_p\left(\frac{1}{\sqrt{n}}\right) = O_p(1),$$

where the second last equality is due to Proposition 10. Then to show that $t'_{2,1} - t_{2,1} = o_p(1)$, it suffices to show

$$\sqrt{n}A - \sqrt{n}\hat{A} = o_p(1) \tag{11.32}$$

$$n\hat{G} - nG = o_p(1). \tag{11.33}$$

11.8.3 Proof of (11.32)

We now show that

$$\left|b^\intercal \hat{\Upsilon}_T^{-1}\sqrt{Tn}\frac{\partial \ell_T(\theta^0)}{\partial \theta^\intercal}/T - b^\intercal \Upsilon^{-1}\sqrt{Tn}\frac{\partial \ell_T(\theta^0)}{\partial \theta^\intercal}/T\right| = o_p(1).$$

This is straightforward.

$$\begin{aligned}
&\left|b^\intercal \hat{\Upsilon}_T^{-1}\sqrt{Tn}\frac{\partial \ell_T(\theta^0)}{\partial \theta^\intercal}/T - b^\intercal \Upsilon^{-1}\sqrt{Tn}\frac{\partial \ell_T(\theta^0)}{\partial \theta^\intercal}/T\right| \\
&= \left|b^\intercal(\hat{\Upsilon}_T^{-1} - \Upsilon^{-1})\frac{\sqrt{Tn}}{2}\mathcal{X}^\intercal(\Theta^{-1/2}\otimes\Theta^{-1/2})(D^{-1/2}\otimes D^{-1/2})\mathrm{vec}(\tilde{\Sigma} - \Sigma)\right| \\
&\leq O(\sqrt{Tn})\|\hat{\Upsilon}_T^{-1} - \Upsilon^{-1}\|_{\ell_2}\|\mathcal{X}^\intercal\|_{\ell_2}\sqrt{n}\|\tilde{\Sigma} - \Sigma\|_{\ell_2} = O_p\left(\sqrt{\frac{n^3}{T}}\right) = o_p(1),
\end{aligned}$$

where the second equality is due to Proposition 13(iv) and (11.7), and the last equality is due to Assumption 2(ii).

11.8.4 Proof of (11.33)

We now show that

$$n\left|b^\intercal(-\hat{\Upsilon}_T)^{-1}b - b^\intercal(-\Upsilon)^{-1}b\right| = o_p(1).$$

This is also straight-forward.

$$n\left|b^\intercal(-\hat{\Upsilon}_T)^{-1}b - b^\intercal(-\Upsilon)^{-1}b\right| = n\left|b^\intercal(\hat{\Upsilon}_T^{-1} - \Upsilon^{-1})b\right| \leq n\|\hat{\Upsilon}_T^{-1} - \Upsilon^{-1}\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right) = o_p(1),$$

where the second equality is due to Proposition 13(iv) and the last equality is due to Assumption 2(i).

11.8.5 $t_{2,2} = o_p(1)$

We now prove $t_{2,2} = o_p(1)$. It suffices to prove

$$\frac{1}{\sqrt{b^\intercal(-\hat{\Upsilon}_T)^{-1}b}} = O_p(n^{1/2}).$$

This follows from (11.30) and (11.33). □

11.9 Proof of Proposition 14

Proposition 14. *For any positive definite matrix* Θ*,*

$$\left(\int_0^1 [t(\Theta - I) + I]^{-1} \otimes [t(\Theta - I) + I]^{-1} dt\right)^{-1} = \int_0^1 e^{t\log\Theta} \otimes e^{(1-t)\log\Theta} dt.$$

Proof. (11.9) and (11.10) of Higham (2008) p272 give, respectively, that

$$\mathrm{vec}E = \int_0^1 e^{t\log\Theta} \otimes e^{(1-t)\log\Theta} dt \mathrm{vec}L(\Theta, E),$$

$$\mathrm{vec}L(\Theta, E) = \int_0^1 [t(\Theta - I) + I]^{-1} \otimes [t(\Theta - I) + I]^{-1} dt \mathrm{vec}E.$$

Substitute the preceding equation into the second last

$$\mathrm{vec}E = \int_0^1 e^{t\log\Theta} \otimes e^{(1-t)\log\Theta} dt \int_0^1 [t(\Theta - I) + I]^{-1} \otimes [t(\Theta - I) + I]^{-1} dt \mathrm{vec}E.$$

Since E is arbitrary, the result follows. □

11.10 Proof of Theorem 5

Proof. Theorem 2 and a result we proved before, namely,

$$|\hat{G}_T - G| = |c^\intercal \hat{J}_T c - c^\intercal J c| = o_p\left(\frac{1}{n\kappa(W)}\right),$$

imply

$$\sqrt{T}c^\intercal(\hat{\theta}_T - \theta^0) \xrightarrow{d} N(0, G).$$

Part (i) then follows trivially from Theorem 2. We now prove part (ii). Write

$$\frac{\sqrt{T}}{\left(\sum_{j=1}^{v}\hat{\theta}_{T,j+1}^2\right)^{1/2}}\left(\sum_{j=1}^{v}\hat{\theta}_{T,j+1}^2-\sum_{j=1}^{v}\mathrm{var}(\zeta_j)\right)$$
$$=\frac{\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}}{\left(\sum_{j=1}^{v}\hat{\theta}_{T,j+1}^2\right)^{1/2}}\frac{\sqrt{T}}{\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}}\left(\sum_{j=1}^{v}\hat{\theta}_{T,j+1}^2-\sum_{j=1}^{v}\mathrm{var}(\zeta_j)\right)=:\frac{\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}}{\left(\sum_{j=1}^{v}\hat{\theta}_{T,j+1}^2\right)^{1/2}}A.$$

Note that

$$\sum_{j=1}^{v}\hat{\theta}_{T,j+1}^2-\sum_{j=1}^{v}[\theta_{j+1}^0]^2=2\sum_{j=1}^{v}\theta_{j+1}^0(\hat{\theta}_{T,j+1}-\theta_{j+1}^0)+\sum_{j=1}^{v}(\hat{\theta}_{T,j+1}-\theta_{j+1}^0)^2.$$

Hence

$$A=2\sqrt{T}\sum_{j=1}^{v}\frac{\theta_{j+1}^0}{\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}}(\hat{\theta}_{T,j+1}-\theta_{j+1}^0)+\frac{\sqrt{T}}{\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}}\sum_{j=1}^{v}(\hat{\theta}_{T,j+1}-\theta_{j+1}^0)^2$$
$$=:A_1+A_2.$$

It is easy to see that A_1 converges weakly by Theorem 2:

$$A_1=2\sqrt{T}c'^{\intercal}(\hat{\theta}_T-\theta^0)\xrightarrow{d}2N(0,G(c')).$$

Next

$$A_2=\frac{\sqrt{T}}{\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}}\sum_{j=1}^{v}(\hat{\theta}_{T,j+1}-\theta_{j+1}^0)^2\le\frac{\sqrt{T}}{\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}}\|\hat{\theta}_T-\theta^0\|_2^2$$
$$=\frac{1}{\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}}O_p\left(\sqrt{\frac{n^2\kappa^2(W)}{T}}\right)=O_p\left(\sqrt{\frac{n^2\kappa^2(W)}{T}}\right)=o_p(1),$$

where the second last equality is true because $\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}$ is bounded away from zero by an absolute constant as long as $\rho_j^0\neq 0$ for some $j=1,\ldots,v$. Last, since $\|\hat{\theta}_{T,-1}-\theta_{-1}^0\|_2\le\|\hat{\theta}_T-\theta^0\|_2\xrightarrow{p}0$,

$$\sum_{j=1}^{v}\hat{\theta}_{T,j+1}^2=\|\hat{\theta}_{T,-1}\|_2^2\xrightarrow{p}\|\theta_{-1}^0\|_2^2=\sum_{j=1}^{v}[\theta_{j+1}^0]^2,$$

where θ_{-1}^0 denotes θ^0 excluding its first component, similarly for $\hat{\theta}_{T,-1}$. Then

$$\frac{\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}}{\left(\sum_{j=1}^{v}\hat{\theta}_{T,j+1}^2\right)^{1/2}}\xrightarrow{p}\frac{\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}}{\left(\sum_{j=1}^{v}[\theta_{j+1}^0]^2\right)^{1/2}}=1,$$

given that $\sum_{j=1}^{v}[\theta_{j+1}^0]^2\neq 0$ via Slutsky's lemma. The result then follows after invoking Slutsky's lemma again. □

11.11 Proof of Theorems 6 and 7

The proofs for Theorems 6(i)-(ii) and 7 are exactly the same. The following proof holds for any f and $c \in \mathbb{R}^{v+1}$ satisfying

(i) $f''(\cdot)$ has a bounded range,

(ii) $f'(\theta^0_{j+1})$ is bounded away from zero for some $j = 1, \ldots, v$,

(iii)
$$c_1 = 0, \qquad c_{j+1} = \frac{f'(\theta^0_{j+1})}{\left(\sum_{j=1}^v [f'(\theta^0_{j+1})]^2\right)^{1/2}}, \quad j = 1, \ldots, v.$$

Proof. Then by Taylor's theorem

$$f(\hat{\theta}_{T,j+1}) - f(\theta^0_{j+1}) = f'(\theta^0_{j+1})(\hat{\theta}_{T,j+1} - \theta^0_{j+1}) + \frac{f''(\dot{\theta}_{j+1})}{2}(\hat{\theta}_{T,j+1} - \theta^0_{j+1})^2,$$

where $\dot{\theta}_{j+1}$ is a point interior to the interval joining $\hat{\theta}_{T,j+1}$ and θ^0_{j+1}. Then

$$\begin{aligned}
&\frac{\sqrt{T}}{\left(\sum_{j=1}^v [f'(\theta^0_{j+1})]^2\right)^{1/2}} \left(\sum_{j=1}^v f(\hat{\theta}_{T,j+1}) - \sum_{j=1}^v f(\theta^0_{j+1})\right) \\
&= \sqrt{T} \sum_{j=1}^v \frac{f'(\theta^0_{j+1})}{\left(\sum_{j=1}^v [f'(\theta^0_{j+1})]^2\right)^{1/2}} (\hat{\theta}_{T,j+1} - \theta^0_{j+1}) + \frac{\sqrt{T}}{\left(\sum_{j=1}^v [f'(\theta^0_{j+1})]^2\right)^{1/2}} \sum_{j=1}^v \frac{f''(\dot{\theta}_{j+1})}{2} (\hat{\theta}_{T,j+1} - \theta^0_{j+1})^2 \\
&=: B_1 + B_2.
\end{aligned}$$

We consider B_1 first. It is easy to see that B_1 converges weakly by Theorem 2:

$$B_1 = \sqrt{T} c^\intercal (\hat{\theta}_T - \theta^0) \xrightarrow{d} N(0, G(c)).$$

Next

$$\begin{aligned}
B_2 &= \frac{\sqrt{T}}{\left(\sum_{j=1}^v [f'(\theta^0_{j+1})]^2\right)^{1/2}} \sum_{j=1}^v \frac{f''(\dot{\theta}_{j+1})}{2} (\hat{\theta}_{T,j+1} - \theta^0_{j+1})^2 \leq \frac{\frac{1}{2} \max_{1 \leq j \leq v} f''(\dot{\theta}_{j+1}) \sqrt{T}}{\left(\sum_{j=1}^v [f'(\theta^0_{j+1})]^2\right)^{1/2}} \sum_{j=1}^v (\hat{\theta}_{T,j+1} - \theta^0_{j+1})^2 \\
&\leq \frac{\frac{1}{2} \max_{1 \leq j \leq v} f''(\dot{\theta}_{j+1})}{\left(\sum_{j=1}^v [f'(\theta^0_{j+1})]^2\right)^{1/2}} \sqrt{T} \|\hat{\theta}_{T,j+1} - \theta^0_{j+1}\|_2^2 \lesssim \frac{1}{\left(\sum_{j=1}^v [f'(\theta^0_{j+1})]^2\right)^{1/2}} \sqrt{T} \|\hat{\theta}_{T,j+1} - \theta^0_{j+1}\|_2^2 \\
&\lesssim \sqrt{T} \|\hat{\theta}_{T,j+1} - \theta^0_{j+1}\|_2^2 = O_p\left(\sqrt{\frac{n^2 \kappa^2(W)}{T}}\right) = o_p(1),
\end{aligned}$$

where the first "$\lesssim$" is because $f''(\dot{\theta}_{j+1})$ is bounded for all j, and the second "$\lesssim$" is because $f'(\theta^0_{j+1})$ is bounded away from zero for some $j = 1, \ldots, v$. Last, since $\|\hat{\theta}_T - \theta^0\|_2 \xrightarrow{p} 0$,

$$\frac{\left(\sum_{j=1}^v [f'(\theta^0_{j+1})]^2\right)^{1/2}}{\left(\sum_{j=1}^v [f'(\hat{\theta}_{T,j+1})]^2\right)^{1/2}} \xrightarrow{p} 1,$$

given that $\sum_{j=1}^v [f'(\theta^0_{j+1})]^2 \neq 0$ via the continuous mapping theorem and Slutsky's lemma. The result then follows after invoking Slutsky's lemma again. □

Obviously c^U, c^L and c^* satisfy the condition (iii) of our generic proof. We are only left to verify that f_1, f_2 and f_3 satisfy the conditions (i)-(ii) of our generic proof.

(a) It is easy to calculate that for $j = 1, \ldots, v$

$$f_1'(\theta_{j+1}^0) = \frac{2}{1 + e^{2\theta_{j+1}^0}}, \qquad f_1''(\theta_{j+1}^0) = -\frac{4e^{2\theta_{j+1}^0}}{(1 + e^{2\theta_{j+1}^0})^2}.$$

Then we see that $f_1''(\cdot)$ has a bounded range $[-1, 0]$ and $f_1'(\theta_{j+1}^0)$ is bounded away from zero if $\rho_j^0 = (e^{2\theta_{j+1}^0} - 1)/(e^{2\theta_{j+1}^0} + 1)$ is bounded away from 1 for some j.

(b) It is easy to calculate that for $j = 1, \ldots, v$

$$f_2'(\theta_{j+1}^0) = -\frac{2e^{2\theta_{j+1}^0}}{1 + e^{2\theta_{j+1}^0}}, \qquad f_2''(\theta_{j+1}^0) = \frac{4e^{4\theta_{j+1}^0}}{(1 + e^{2\theta_{j+1}^0})^2}.$$

Then we see that $f_2''(\cdot)$ has a bounded range $[0, 4]$ and $f_2'(\theta_{j+1}^0)$ is bounded away from zero if $\rho_j^0 = (e^{2\theta_{j+1}^0} - 1)/(e^{2\theta_{j+1}^0} + 1)$ is bounded away from -1 for some j.

(c) It is easy to calculate that for $j = 1, \ldots, v$

$$f_3'(\theta_{j+1}^0) = \frac{2e^{-2\theta_{j+1}^0}}{1 + e^{-2\theta_{j+1}^0}}, \qquad f_3''(\theta_{j+1}^0) = \frac{-4e^{-2\theta_{j+1}^0}}{(1 + e^{-2\theta_{j+1}^0})^2}.$$

Then we see that $f_3''(\cdot)$ has a bounded range $[-1, 0]$ and $f_3'(\theta_{j+1}^0)$ is bounded away from zero if $\rho_j^0 = (e^{2\theta_{j+1}^0} - 1)/(e^{2\theta_{j+1}^0} + 1)$ is bounded away from 1 for some j.

11.12 Proof of Theorem 8

Proof. Note that under H_0,

$$\begin{aligned}\sqrt{T} g_T(\theta^0) &= \sqrt{T}[\text{vech}(\log M_T) - E\theta^0] = \sqrt{T}[\text{vech}(\log M_T) - \text{vech}(\log \Theta^0)] \\ &= \sqrt{T}[\text{vech}(\log M_T) - \text{vech}(\log \Theta)] = \sqrt{T} D_n^+ \text{vec}(\log M_T - \log \Theta),\end{aligned}$$

where the third equality is true under H_0. Thus we can adopt the same method as in Theorem 2 to establish the asymptotic distribution of $\sqrt{T} g_T(\theta^0)$. In fact, it will be much simpler here because we fixed n. We should have

$$\sqrt{T} g_T(\theta^0) \xrightarrow{d} N(0, S), \qquad S := D_n^+ H (D^{-1/2} \otimes D^{-1/2}) V (D^{-1/2} \otimes D^{-1/2}) H D_n^{+\intercal}, \quad (11.34)$$

where S is positive definite given Assumptions 3 and 5. The close-form solution for $\hat{\theta}_T = \hat{\theta}_T(W)$ has been given in (5.4), but this is not important. We only need that $\hat{\theta}_T$ sets the first derivation of the objective function to zero:

$$E^\intercal W g_T(\hat{\theta}_T) = 0. \quad (11.35)$$

Notice that

$$g_T(\hat{\theta}_T) - g_T(\theta^0) = -E(\hat{\theta}_T - \theta^0). \quad (11.36)$$

Pre-multiply (11.36) by $\frac{\partial g_T(\hat{\theta}_T)}{\partial \theta^\intercal} W = -E^\intercal W$ to give

$$-E^\intercal W[g_T(\hat{\theta}_T) - g_T(\theta^0)] = E^\intercal W E(\hat{\theta}_T - \theta^0),$$

whence we obtain

$$\hat{\theta}_T - \theta^0 = -(E^\intercal W E)^{-1} E^\intercal W [g_T(\hat{\theta}_T) - g_T(\theta^0)]. \tag{11.37}$$

Substitute (11.37) into (11.36)

$$\begin{aligned}\sqrt{T} g_T(\hat{\theta}_T) &= \left[I_{n(n+1)/2} - E(E^\intercal W E)^{-1} E^\intercal W\right] \sqrt{T} g_T(\theta^0) + E(E^\intercal W E)^{-1} \sqrt{T} E^\intercal W g_T(\hat{\theta}_T) \\ &= \left[I_{n(n+1)/2} - E(E^\intercal W E)^{-1} E^\intercal W\right] \sqrt{T} g_T(\theta^0),\end{aligned}$$

where the second equality is due to (11.35). Using (11.34), we have

$$\begin{aligned}&\sqrt{T} g_T(\hat{\theta}_T) \xrightarrow{d} \\ &N\left(0, \left[I_{n(n+1)/2} - E(E^\intercal W E)^{-1} E^\intercal W\right] S \left[I_{n(n+1)/2} - E(E^\intercal W E)^{-1} E^\intercal W\right]^\intercal\right).\end{aligned}$$

Now choosing $W = S^{-1}$, we can simplify the asymptotic covariance matrix in the preceding display to

$$S^{1/2} \left(I_{n(n+1)/2} - S^{-1/2} E (E^\intercal S^{-1} E)^{-1} E^\intercal S^{-1/2}\right) S^{1/2}.$$

Thus

$$\sqrt{T} \hat{S}^{-1/2} g_T(\hat{\theta}_T) \xrightarrow{d} N\left(0, I_{n(n+1)/2} - S^{-1/2} E (E^\intercal S^{-1} E)^{-1} E^\intercal S^{-1/2}\right),$$

because $\hat{S}$ is a consistent estimate of S given Propositions 7 and 8, which hold under the assumptions of this theorem. The asymptotic covariance matrix in the preceding display is idempotent and has rank $n(n+1)/2 - (v+1)$. Thus, under H_0,

$$T g_T(\hat{\theta}_T)^\intercal \hat{S}^{-1} g_T(\hat{\theta}_T) \xrightarrow{d} \chi^2_{n(n+1)/2-(v+1)}.$$

□

11.13 Proof of Corollary 3

Proof. From (6.5) and the Slutsky lemma, we have for every fixed n (and hence v)

$$\frac{T g_T(\hat{\theta}_T)^\intercal \hat{S}^{-1} g_T(\hat{\theta}_T) - \left[\frac{n(n+1)}{2} - (v+1)\right]}{\left[n(n+1) - 2(v+1)\right]^{1/2}} \xrightarrow{d} \frac{\chi^2_{n(n+1)/2-(v+1)} - \left[\frac{n(n+1)}{2} - (v+1)\right]}{\left[n(n+1) - 2(v+1)\right]^{1/2}},$$

as $T \to \infty$. Then invoke Lemma 8 in Appendix B

$$\frac{\chi^2_{n(n+1)/2-(v+1)} - \left[\frac{n(n+1)}{2} - (v+1)\right]}{\left[n(n+1) - 2(v+1)\right]^{1/2}} \xrightarrow{d} N(0,1),$$

as $n \to \infty$ under H_0. Next invoke Lemma 9 in Appendix B, there exists a sequence $n_T \to \infty$ such that

$$\frac{T g_{T,n_T}(\hat{\theta}_{T,n_T})^\intercal \hat{S}^{-1} g_{T,n_T}(\hat{\theta}_{T,n_T}) - \left[\frac{n_T(n_T+1)}{2} - (v_T+1)\right]}{\left[n_T(n_T+1) - 2(v_T+1)\right]^{1/2}} \xrightarrow{d} N(0,1), \qquad \text{under } H_0$$

as $T \to \infty$. □

12 Appendix B

12.1 The MD Estimator

Proposition 15. *Let A, B be $n \times n$ complex matrices. Suppose that A is positive definite for all n and its minimum eigenvalue is uniformly bounded away from zero by an absolute constant. Assume $\|A^{-1}B\|_{\ell_2} \leq C < 1$ for some constant C. Then $A+B$ is invertible for every n and*

$$(A+B)^{-1} = A^{-1} - A^{-1}BA^{-1} + O(\|B\|_{\ell_2}^2).$$

Proof. We write $A + B = A[I - (-A^{-1}B)]$. Since $\| - A^{-1}B\|_{\ell_2} \leq C < 1$, $I - (-A^{-1}B)$ and hence $A + B$ are invertible (Horn and Johnson (1985) p301). We then can expand

$$(A+B)^{-1} = \sum_{k=0}^{\infty}(-A^{-1}B)^k A^{-1} = A^{-1} - A^{-1}BA^{-1} + \sum_{k=2}^{\infty}(-A^{-1}B)^k A^{-1}.$$

Then

$$\left\|\sum_{k=2}^{\infty}(-A^{-1}B)^k A^{-1}\right\|_{\ell_2} \leq \left\|\sum_{k=2}^{\infty}(-A^{-1}B)^k\right\|_{\ell_2} \|A^{-1}\|_{\ell_2} \leq \sum_{k=2}^{\infty}\left\|(-A^{-1}B)^k\right\|_{\ell_2} \|A^{-1}\|_{\ell_2}$$
$$\leq \sum_{k=2}^{\infty}\left\|-A^{-1}B\right\|_{\ell_2}^k \|A^{-1}\|_{\ell_2} = \frac{\left\|A^{-1}B\right\|_{\ell_2}^2 \|A^{-1}\|_{\ell_2}}{1 - \|A^{-1}B\|_{\ell_2}} \leq \frac{\|A^{-1}\|_{\ell_2}^3 \|B\|_{\ell_2}^2}{1-C},$$

where the first and third inequalities are due to the submultiplicative property of a matrix norm, the second inequality is due to the triangular inequality. Since A is positive definite with the minimum eigenvalue bounded away from zero by an absolute constant, $\|A^{-1}\|_{\ell_2} = \text{maxeval}(A^{-1}) = 1/\text{mineval}(A) < D < \infty$ for some absolute constant D. Hence the result follows. □

Theorem 9 (Higham (2008) p269; Dieci, Morini, and Papini (1996))**.** *For $A \in \mathbb{C}^{n\times n}$ with no eigenvalues lying on the closed negative real axis $(-\infty, 0]$,*

$$\log A = \int_0^1 (A-I)[t(A-I)+I]^{-1}dt.$$

Definition 1 (Nets and covering numbers)**.** *Let (T, d) be a metric space and fix $\varepsilon > 0$.*

(i) A subset $\mathcal{N}_\varepsilon$ of T is called an ε-net of T if every point $x \in T$ satisfies $d(x, y) \leq \varepsilon$ for some $y \in \mathcal{N}_\varepsilon$.

(ii) The minimal cardinality of an ε-net of T is denote $\mathcal{N}(\varepsilon, d)$ and is called the covering number of T (at scale ε). Equivalently, $\mathcal{N}(\varepsilon, d)$ is the minimal number of balls of radius ε and with centers in T needed to cover T.

Lemma 1. *The unit Euclidean sphere* $\{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ *equipped with the Euclidean metric* d *satisfies for every* $\varepsilon > 0$ *that*

$$\mathcal{N}(\varepsilon, d) \leq \left(1 + \frac{2}{\varepsilon}\right)^n.$$

Proof. See Vershynin (2011) p8. □

Recall that for a symmetric $n \times n$ matrix A, its ℓ_2 spectral norm can be written as: $\|A\|_{\ell_2} = \max_{\|x\|_2=1} |x^\intercal A x|$.

Lemma 2. *Let* A *be a symmetric* $n \times n$ *matrix, and let* $\mathcal{N}_\varepsilon$ *be an* ε*-net of the unit sphere* $\{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ *for some* $\varepsilon \in [0, 1)$*. Then*

$$\|A\|_{\ell_2} \leq \frac{1}{1 - 2\varepsilon} \max_{x \in \mathcal{N}_\varepsilon} |x^\intercal A x|.$$

Proof. See Vershynin (2011) p8. □

Theorem 10 (Bernstein's inequality)**.** *We let* $Z_1, \ldots, Z_T$ *be independent random variables, satisfying for positive constants* A *and* σ_0^2

$$\mathbb{E}Z_t = 0 \quad \forall t, \quad \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}|Z_t|^m \leq \frac{m!}{2} A^{m-2}\sigma_0^2, \quad m = 2, 3, \ldots.$$

Let $\epsilon > 0$ *be arbitrary. Then*

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Z_t\right| \geq \sigma_0^2\left[A\epsilon + \sqrt{2\epsilon}\right]\right) \leq 2e^{-T\sigma_0^2\epsilon}.$$

Proof. Slightly adapted from Bühlmann and van de Geer (2011) p487. □

Lemma 3. *Let* $\hat{\Omega}_n$ *and* Ω_n *be invertible (both possibly stochastic) square matrices whose dimensions could be growing. Let* T *be the sample size. For any matrix norm, suppose that* $\|\Omega_n^{-1}\| = O_p(1)$ *and* $\|\hat{\Omega}_n - \Omega_n\| = O_p(a_{n,T})$ *for some sequence* $a_{n,T}$ *with* $a_{n,T} \to 0$ *as* $n \to \infty$*,* $T \to \infty$ *simultaneously (joint asymptotics). Then* $\|\hat{\Omega}_n^{-1} - \Omega_n^{-1}\| = O_p(a_{n,T})$.

Proof. The original proof could be found in Saikkonen and Lutkepohl (1996) Lemma A.2.

$$\|\hat{\Omega}_n^{-1} - \Omega_n^{-1}\| \leq \|\hat{\Omega}_n^{-1}\|\|\Omega_n - \hat{\Omega}_n\|\|\Omega_n^{-1}\| \leq \left(\|\Omega_n^{-1}\| + \|\hat{\Omega}_n^{-1} - \Omega_n^{-1}\|\right)\|\Omega_n - \hat{\Omega}_n\|\|\Omega_n^{-1}\|.$$

Let $v_{n,T}$, $z_{n,T}$ and $x_{n,T}$ denote $\|\Omega_n^{-1}\|$, $\|\hat{\Omega}_n^{-1} - \Omega_n^{-1}\|$ and $\|\Omega_n - \hat{\Omega}_n\|$, respectively. From the preceding equation, we have

$$w_{n,T} := \frac{z_{n,T}}{(v_{n,T} + z_{n,T})v_{n,T}} \leq x_{n,T} = O_p(a_{n,T}) = o_p(1).$$

We now solve for $z_{n,T}$:

$$z_{n,T} = \frac{v_{n,T}^2 w_{n,T}}{1 - v_{n,T}w_{n,T}} = O_p(a_{n,T}).$$

□

Lemma 4. *Let A, B be $n \times n$ positive semidefinite matrices and not both singular. Then*

$$\|A - B\|_{\ell_2} \leq \frac{\|A^2 - B^2\|_{\ell_2}}{mineval(A) + mineval(B)}.$$

Proof. See Horn and Johnson (1985) p410. □

Proposition 16. *Consider real matrices A $(m \times n)$ and B $(p \times q)$. Then*

$$\|A \otimes B\|_{\ell_2} = \|A\|_{\ell_2}\|B\|_{\ell_2}.$$

Proof.

$$\begin{aligned}\|A \otimes B\|_{\ell_2} &= \sqrt{\text{maxeval}[(A \otimes B)^\intercal (A \otimes B)]} = \sqrt{\text{maxeval}[(A^\intercal \otimes B^\intercal)(A \otimes B)]} \\ &= \sqrt{\text{maxeval}[A^\intercal A \otimes B^\intercal B]} = \sqrt{\text{maxeval}[A^\intercal A]\text{maxeval}[B^\intercal B]} = \|A\|_{\ell_2}\|B\|_{\ell_2},\end{aligned}$$

where the fourth equality is due to the fact that both $A^\intercal A$ and $B^\intercal B$ are positive semidefinite. □

Lemma 5. *Let A be a $p \times p$ symmetric matrix and $\hat{v}, v \in \mathbb{R}^p$. Then*

$$|\hat{v}^\intercal A \hat{v} - v^\intercal A v| \leq |maxeval(A)|\|\hat{v} - v\|_2^2 + 2(\|Av\|_2\|\hat{v} - v\|_2).$$

Proof. See Lemma 3.1 in the supplementary material of van de Geer, Buhlmann, Ritov, and Dezeure (2014). □

Proposition 17. *Suppose we have subgaussian random variables $Z_{l,t,j}$ for $l = 1, \ldots, L$ ($L \geq 2$ fixed), $t = 1, \ldots, T$ and $j = 1, \ldots, p$. Z_{l_1,t_1,j_1} and Z_{l_2,t_2,j_2} are independent as long as $t_1 \neq t_2$ regardless of the values of other subscripts. Then,*

$$\max_{1 \leq j \leq p} \max_{1 \leq t \leq T} \mathbb{E}\left|\prod_{l=1}^{L} Z_{l,t,j}\right| \leq A = O(1),$$

for some positive constant A and

$$\max_{1 \leq j \leq p}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\prod_{l=1}^{L} Z_{l,t,j} - \mathbb{E}\left[\prod_{l=1}^{L} Z_{l,t,j}\right]\right)\right| = O_p\left(\sqrt{\frac{(\log p)^{L+1}}{T}}\right).$$

Proof. See Proposition 3 of Kock and Tang (2016). □

12.2 The QMLE

Lemma 6. *Let A and B be $m \times n$ and $p \times q$ matrices, respectively. There exists a unique permutation matrix $P := I_n \otimes K_{q,m} \otimes I_p$, where $K_{q,m}$ is the commutation matrix, such that*

$$vec(A \otimes B) = P(vecA \otimes vecB).$$

Proof. Magnus and Neudecker (2007) Theorem 3.10 p55. □

Lemma 7. *For $m, n \geq 0$, we have*

$$\int_0^1 (1-s)^n s^m ds = \frac{m!n!}{(m+n+1)!}.$$

Theorem 11. *For arbitrary $n \times n$ complex matrices A and E, and for any matrix norm $\|\cdot\|$,*

$$\|e^{A+E} - e^A\| \leq \|E\| \exp(\|E\|) \exp(\|A\|).$$

Proof. See Horn and Johnson (1991) p430. □

12.3 The Over-Identification Test

Lemma 8 (van der Vaart (1998) p27)**.**

$$\frac{\chi_k^2 - k}{\sqrt{2k}} \xrightarrow{d} N(0,1),$$

as $k \to \infty$.

Lemma 9 (van der Vaart (2010) p41)**.** *For $T, n \in \mathbb{N}$ let $X_{T,n}$ be random vectors such that $X_{T,n} \rightsquigarrow X_n$ as $T \to \infty$ for every fixed n such that $X_n \rightsquigarrow X$ as $n \rightsquigarrow \infty$. Then there exists a sequence $n_T \to \infty$ such that $X_{T,n_T} \rightsquigarrow X$ as $T \to \infty$.*

References

Akdemir, D. and A. K. Gupta (2011): "Array Variate Random Variables with Multiway Kronecker Delta Covariance Matrix Structure," *Journal of Algebraic Statistics*, 2, 98–113.

Allen, G. I. (2012): "Regularized Tensor Factorizations and Higher-Order Principal Components Analysis," *arXiv preprint arXiv:1202.2476v1.*

Amemiya, T. (1983): "Partially Generalized Least Squares and Two-Stage Least Squares Estimators," *Journal of Econometrics*, 23, 275–283.

Anderson, T. W. (1984): *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons.

Bai, J. (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica*, 77, 1229–1279.

Bai, J. and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

Bickel, P. J. (1975): "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association*, 70, 428–434.

Bickel, P. J. and E. Levina (2008): "Covariance Regularization by Thresholding," *The Annals of Statistics*, 36, 2577–2604.

Billingsley, P. (1995): *Probability and Measure*, Wiley.

Browne, M. W., R. C. MacCallum, C.-T. Kim, B. L. Andersen, and R. Glaser (2002): "When Fit Indices and Residuals are Incompatible," *Psychological Methods*, 7, 403–421.

Browne, M. W. and A. Shapiro (1991): "Invariance of Covariance Structures under Groups of Transformations," *Metrika*, 38, 345–355.

Bühlmann, P. and S. van de Geer (2011): *Statistics for High-Dimensional Data*, Springer.

Cai, T. T., Z. Ren, and H. H. Zhou (2014): "Estimating Structured High-Dimensional Covariance and Precision Matrices: Optimal Rates and Adaptive Estimation," *Working Paper*.

Cai, T. T., C.-H. Zhang, and H. H. Zhou (2010): "Optimal Rates of Convergence for Covariance Matrix Estimation," *The Annals of Statistics*, 38, 2118–2144.

Cai, T. T. and H. H. Zhou (2012): "Optimal Rates of Convergence for Sparse Covariance Matrix Estimation," *The Annals of Statistics*, 40, 2389–2420.

Campbell, D. T. and E. J. O'Connell (1967): "Method Factors in Multitrait-Multimethod Matrices: Multiplicative Rather than Additive?" *Multivariate Behavioral Research*, 2, 409–426.

Chan, L., J. Karceski, and J. Lakonishok (1999): "On Portfolio Optimization: Forecasting Covariances and Choosing the Risk Model," *Review of Financial Studies*, 12, 937–974.

Cohen, J. E., K. Usevich, and P. Comon (2016): "A Tour of Constrained Tensor Canonical Polyadic Decomposition," *Research Report*.

Constantinou, P., P. Kokoszka, and M. Reimherr (2015): "Testing Separability of Space-Time Functional Processes," *arXiv preprint arXiv: 1509.07017v1*.

Cudeck, R. (1988): "Multiplicative Models and MTMM Matrices," *Journal of Educational Statistics*, 13, 131–147.

Dieci, L., B. Morini, and A. Papini (1996): "Computational Techniques for Real Logarithms of Matrices," *SIAM Journal on Matrix Analysis and Applications*, 17, 570–593.

Dobra, A. (2014): "Graphical Modeling of Spatial Health Data," *arXiv preprint arXiv: 1411.6512v1*.

Fan, J., Y. Fan, and J. Lv (2008): "High Dimensional Covariance Matrix Estimation Using a Factor Model," *Journal of Econometrics*, 147, 186–197.

Fan, J., Y. Liao, and M. Mincheva (2013): "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 603–680.

Fan, J., Y. Liao, and X. Shi (2015): "Risks of Large Portfolios," *Journal of Econometrics*, 186, 367–387.

Fosdick, B. K. and P. D. Hoff (2014): "Separable Factor Analysis with Applications to Mortality Data," *The Annals of Applied Statistics*, 8, 120–147.

Gerard, D. and P. Hoff (2015): "Equivariant Minimax Dominators of the MLE in the Array Normal Model," *Journal of Multivariate Analysis*, 137, 42–49.

Gibrat, R. (1931): *Les Inégalités économiques*, Paris, France.

He, X. and Q. Shao (2000): "On Parameters of Increasing Dimensions," *Journal of Multivariate Analysis*, 73, 120–135.

Higham, N. J. (2008): *Functions of Matrices: Theory and Computation*, SIAM.

Hoff, P. D. (2011): "Separable Covariance Arrays via the Tucker Product, with Applications to Multivariate Relational Data," *Bayesian Analysis*, 6, 179–196.

——— (2015): "Multilinear Tensor Regression for Longitudinal Relational Data," *The Annals of Applied Statistics*, 9, 1169–1193.

——— (2016): "Equivariant and Scale-Free Tucker Decomposition Models," *Bayesian Analysis*, 11, 627–648.

Horn, R. and C. Johnson (1985): *Matrix Analysis*, Cambridge University Press.

Horn, R. A. and C. R. Johnson (1991): *Topics in Matrix Analysis*, Cambridge University Press.

Kock, A. B. and H. Tang (2016): "Uniform Inference in High-Dimensional Dynamic Panel Data Models," *arXiv preprint arXiv:1501.00478*.

Krijnen, W. P. (2004): "Convergence in Mean Square of Factor Predictors," *British Journal of Mathematical and Statistical Psychology*, 57, 311–326.

Ledoit, O. and M. Wolf (2003): "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection," *Journal of Empirical Finance*, 10, 603–621.

——— (2004): "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices," *Journal of Multivariate Analysis*, 88, 365–411.

Leiva, R. and A. Roy (2014): "Classification of Higher-Order Data with Separable Covariance and Structured Multiplicative or Additive Mean Models," *Communications in Statistics - Theory and Methods*, 43, 989–1012.

Leng, C. and C. Y. Tang (2012): "Sparse Matrix Graphical Models," *Journal of the American Statistical Association*, 107, 1187–1200.

Li, L. and X. Zhang (2016): "Parsimonious Tensor Response Regression," *Journal of the American Statistical Association*.

Linton, O. and J. R. McCrorie (1995): "Differentiation of an Exponential Matrix Function," *Econometric Theory*, 11, 1182–1185.

Magnus, J. R. and H. Neudecker (1986): "Symmetry, 0-1 Matrices and Jacobians a Review," *Econometric Theory*, 157–190.

——— (2007): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley and Sons Ltd.

Manceura, A. M. and P. Dutilleul (2013): "Maximum Likelihood Estimation for the Tensor Normal Distribution: Algorithm, Minimum Sample Size, and Empirical Bias and Dispersion," *Journal of Computational and Applied Mathematics*, 239, 37–49.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, IV.

Ning, Y. and H. Liu (2013): "High-Dimensional Semiparametric Bigraphical Models," *Biometrika*, 100, 655–670.

Ohlson, M., M. R. Ahmada, and D. von Rosen (2013): "The Multilinear Normal Distribution: Introduction and Some Basic Properties," *Journal of Multivariate Analysis*, 113, 37–47.

Onatski, A. (2009): "Testing Hypotheses about the Number of Factors in Large Factor Models," *Econometrica*, 77, 1447–1479.

Phillips, P. C. and H. R. Moon (1999): "Linear Regression Limit Theory for Nonstationary Panel Data," *Econometrica*, 67, 1057–1111.

Pitsianis, N. P. (1997): "The Kronecker Product in Approximation and Fast Transform Generation," *PhD Thesis*.

Rao, P. S. (1997): *Variance and Covariance Components: Mixed Models, Methodologies and Applications*, Chapman Hall.

Saikkonen, P. and H. Lutkepohl (1996): "Infinite-order Cointegrated Vector Autoregressive Processes," *Econometric Theory*, 12, 814–844.

Shephard, N. (1996): "Statistical Aspects of ARCH and Stochastic Volatility," *Time Series Models in Econometrics, Finance and Other Fields*, London: Chapman & Hall, 1–67.

Singull, M., M. R. Ahmad, and D. von Rosen (2012): "More on the Kronecker Structured Covariance Matrix," *Communications in Statistics-Theory and Methods*, 41, 2512–2523.

Swain, A. J. (1975): "Analysis of Parametric Structures for Variance Matrices," *PhD Thesis, University of Adelaide*.

van de Geer, S., P. Buhlmann, Y. Ritov, and R. Dezeure (2014): "On Asymptotically Optimal Confidence Regions and Tests for High-dimensional Models," *The Annals of Statistics*, 42, 1166–1202.

van der Vaart, A. (1998): *Asymptotic Statistics*, Cambridge University Press.

——— (2010): *Time Series*.

van der Vaart, A. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

van Loan, C. F. (2000): "The Ubiquitous Kronecker Product," *Journal of Computational and Applied Mathematics*, 123, 85–100.

Verhees, J. and T. J. Wansbeek (1990): "A Multimode Direct Product Model for Covariance Structure Analysis," *British Journal of Mathematical and Statistical Psychology*, 43, 231–240.

Vershynin, R. (2011): "Introduction to the non-asymptotic analysis of random matrices," *Chapter 5 of: Compressed Sensing, Theory and Applications*.

Volfovsky, A. and P. D. Hoff (2014): "Hierarchical Array Priors for ANOVA Decompositions," *The Annals of Applied Statistics*, 8, 19–41.

——— (2015): "Testing for Nodal Dependence in Relational Data Matrices," *Journal of the American Statistical Association*, 110, 1037–1046.

Wang, W. and J. Fan (2016): "Asymptotics of Empirical Eigen-Structure for High Dimensional Spiked Covariance," *The Annals of Statistics*, To appear.

Yao, J., S. Zheng, and Z. Bai (2015): *Large Sample Covariance Matrices and High-Dimensional Data Analysis*, Cambridge University Press.

Yin, J. and H. Li (2012): "Model Selection and Estimation in the Matrix Normal Graphical Model," *Journal of Multivariate Analysis*, 107, 119–140.