cemmap

centre for microdata methods and practice

Estimation of a Multiplicative Covariance Structure

Christian M. Hafner
Oliver Linton
Haihan Tang

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP23/16

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

Estimation of a Multiplicative Covariance Structure*

Christian M. Hafner†
Université catholiquede Louvain

Oliver Linton‡
University of Cambridge

Haihan Tang§
University of Cambridge

May 16, 2016

Abstract

We consider a Kronecker product structure for large covariance matrices, which has the feature that the number of free parameters increases logarithmically with the dimensions of the matrix. We propose an estimation method of the free parameters based on the log linear property of this structure, and also a Quasi-Likelihood method. We establish the rate of convergence of the estimated parameters when the size of the matrix diverges. We also establish a CLT for our method. We apply the method to portfolio choice for S&P500 daily returns and compare with sample covariance based methods and with the recent Fan et al. (2013) method.

Some key words: Correlation Matrix; Kronecker Product; MTMM; Portfolio Choice
AMS 2000 subject classification: 62F12

*We would like to thank 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, for example. In psychology there is a long history of modelling unobserved traits through factor models that imply specific structure 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, see also Wansbeek and Meijer (2000) for an extensive discussion of models and applications in social sciences. More recently, theoretical and empirical work has considered the case where the covariance matrix is large, see for example Ledoit and Wolf (2003), Bickel and Levina (2008), Onatski (2009), and Fan et al. (2013). The general approach is to impose some sparsity on the model or to use a shrinkage method that achieves effectively the same dimensionality reduction.

We consider a parametric model for the covariance matrix or the 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. We reconsider this model in the setting where the matrix dimension n is large, i.e., increases with the sample size T. In this setting, the model effectively imposes sparseness on the covariance matrix, since the number of free parameters in the covariance matrix grows logarithmically with dimensions. We propose a closed-form minimum distance estimator of the parameters of this model as well as an approximate MLE. We establish the rate of convergence and asymptotic normality of the estimated parameters when n and T diverge. In the following two sections we discuss this model and its motivation.

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 ℓ_2 (Euclidean) norm. For any real matrix A, let $\text{maxeval}(A)$ and $\text{mineval}(A)$ denote its maximum and minimum eigenvalues, respectively. Let $\|A\|_F := [\text{tr}(A^\intercal A)]^{1/2} \equiv [\text{tr}(AA^\intercal)]^{1/2} \equiv \|\text{vec}A\|_2$ and $\|A\|_{\ell_2} := \max_{\|x\|_2=1} \|Ax\|_2 \equiv \sqrt{\text{maxeval}(A^\intercal A)}$ denote the Frobenius norm and spectral norm of A, respectively.

Let A be a $m \times n$ matrix. $\text{vec}A$ 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 $\text{vec}A$ to $\text{vec}(A^\intercal)$, i.e., $\text{vec}(A^\intercal) = K_{m,n}\text{vec}(A)$. 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 $\text{vec}A$, this defines a new $n(n+1)/2 \times 1$ vector, denoted $\text{vech}A$. They are related by the full-column-rank, $n^2 \times n(n+1)/2$ *duplication matrix* D_n: $\text{vec}A = D_n\text{vech}A$. Conversely, $\text{vech}A = D_n^+\text{vec}A$, 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 $a_T \leq Cb_T$ for some positive real number C for all $T \geq 1$.

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 On the Covariance Matrix

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

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

where the covariance matrix 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 Σ_j is positive semidefinite, then so is Σ^*. The total number of free parameters in Σ^* is $\sum_{j=1}^{v} r_j$ where $r_j := n_j(n_j+1)/2$, 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. We will discuss this further below.

This Kronecker type of 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, 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 $u_{i,t}$, while $\gamma = (\gamma_1, \ldots, \gamma_n)^\intercal$ 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

$$\mathrm{var}(u) = \mathbb{E}[uu^\intercal] = \Gamma \otimes \Phi,$$

where $\Gamma = \mathbb{E}\gamma\gamma^\intercal$ is $n \times n$ and $\Phi = \mathbb{E}ff^\intercal$ is $T \times T$.[2] We can think of our more general structure (2.1) arising from a multi-index setting with v multiplicative factors. The interpretation is that there are v different indexes that define an observation and

$$u_{i_1, i_2, \ldots, i_v} = \varepsilon_{i_1} \varepsilon_{i_2} \cdots \varepsilon_{i_v},$$

[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 $u_t \sim N(0, \tau I_k)$, 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

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

where B_j are 2×2 positive definite matrices.

[2]In the so called BEKK model for multivariate GARCH processes, the authors consider a similar Kronecker parameterization of the 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.

where the errors $\varepsilon_{i_1}, \ldots, \varepsilon_{i_v}$ are mutually independent. The 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 multimethod (MTMM) context in psychometrics (see e.g. Campbell and O'Connell (1967) and Cudeck (1988)). This structure has been considered before in Swain (1975) and Verhees and Wansbeek (1990), where they emphasize the case where v is small and where the structure is known and correct. Our thinking is more along the lines that we allow v to be large, and use Σ^* in (2.1) as an approximation device to Σ. In some sense as we shall see the Kronecker product structure corresponds to a kind of additive structure on the log of the covariance matrix, and so from a mathematical point of view it has some advantages.

There are two issues with the model (2.1). First, there is also 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 upper diagonal element to be 1. Second, if the matrices Σ_js 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 begs the question of how one chooses the correct permutation, and we discuss this briefly below.

2.3 On the Correlation Matrix

In this paper, we will mainly approximate the *correlation matrix*, instead of the covariance matrix, with a Kronecker product structure. 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 $\Sigma := \mathbb{E}(x_t - \mu)(x_t - \mu)^\intercal$. Define $D := \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, where $\sigma_i^2 := \mathbb{E}(x_{t,i} - \mu_i)^2$. That is, D is a diagonal matrix with the ith diagonal entry being Σ_{ii}. Define

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

such that $\mathbb{E}y_t = 0$ and $\text{var}[y_t] = D^{-1/2}\Sigma D^{-1/2} =: \Theta$. Note that Θ is the correlation matrix; that is, it has all its diagonal entries to be 1. This is the matrix which we will estimate using our Kronecker product method.

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

$$\Theta^0 = \Theta_1^0 \otimes \Theta_2^0 \otimes \cdots \otimes \Theta_v^0 = \begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 1 & \rho_v \\ \rho_v & 1 \end{bmatrix} \tag{2.2}$$

which minimizes $\|\log\Theta - \log\Theta^*\|_F$ among $\log\Theta^*$. Define

$$\begin{aligned}
\Omega^0 &:= \log\Theta^0 \\
&= (\log\Theta_1^0 \otimes I_2 \otimes \cdots \otimes I_2) + (I_2 \otimes \log\Theta_2^0 \otimes \cdots \otimes I_2) + \cdots + (I_2 \otimes \cdots \otimes \log\Theta_v^0), \\
&=: (\Omega_1 \otimes I_2 \otimes \cdots \otimes I_2) + (I_2 \otimes \Omega_2 \otimes \cdots \otimes I_2) + \cdots + (I_2 \otimes \cdots \otimes \Omega_v), \qquad (2.3)
\end{aligned}$$

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

$$\begin{aligned}\Omega_1 = \log\Theta_1^0 &= \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} \log(1+\rho_1) & 0 \\ 0 & \log(1-\rho_1) \end{pmatrix}\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2}\log(1-\rho_1^2) & \frac{1}{2}\log\left(\frac{1+\rho_1}{1-\rho_1}\right) \\ \frac{1}{2}\log\left(\frac{1+\rho_1}{1-\rho_1}\right) & \frac{1}{2}\log(1-\rho_1^2) \end{pmatrix} =: \begin{pmatrix} a_1 & b_1 \\ b_1 & a_1 \end{pmatrix}, \end{aligned} \tag{2.4}$$

whence we see that ρ_1 generates two distinct entries - one negative and one positive - in Ω_1. We also see that Ω_1 is not only symmetric about the diagonal, but also symmetric about the cross-diagonal (from the upper right to the lower left). In this paper, for simplicity we will not utilize the information about the signs of the entries of Ω_1; we merely use the estimates of entries of Ω_1 to recover ρ_1 (in some over-identified sense). The same reasoning applies to $\Omega_2, \ldots, \Omega_v$. Therefore we obtain an estimate $\hat{\Theta}^0$, which we will use to approximate Θ. 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 4.

3 Some Motivating Properties of the Model

In this section we give three motivational reasons for considering the Kronecker product model. First, we show that for any given covariance matrix (or 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 arbitrary large covariance matrix well. 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 to ensure that Σ^* is positive definite. For an identification issue we also impose the first diagonal of Σ_1 is 1. Since Σ_1 and Σ_2 are symmetric, we can orthogonally diagonalize them:

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

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 ensure 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 function. Now

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

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 $\log\Sigma^*$.

Let $\mathcal{M}_n$ denote the set of all $n \times n$ real symmetric matrices. Define the inner product $\langle A, B\rangle = \mathrm{tr}(A^\intercal B) = \mathrm{tr}(AB)$, inducing the Frobenius norm $\|\cdot\|_F$. $\mathcal{M}_n$ with this inner product can be identified by $\mathbb{R}^{n^2}$ with Euclidean inner product. Since $\mathbb{R}^{n^2}$ with Euclidean inner product is a Hilbert space (for finite n), so is $\mathcal{M}_n$.

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

$$\Sigma_{1,1} \otimes \Sigma_{2,1} + \Sigma_{1,2} \otimes \Sigma_{2,2} = \Sigma_{1,3} \otimes \Sigma_{2,3},$$

where $\Sigma_{1,j}$ are $n_1 \times n_1$ and $\Sigma_{2,j}$ are $n_2 \times n_2$ for $j = 1, 2$. Second, $\mathcal{C}_n$ is a positive cone, hence not necessarily a subspace. Third, 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

$$\begin{aligned}
&(\log\Sigma_{1,1} \otimes I_{n_2} + I_{n_1} \otimes \log\Sigma_{2,1}) + (\log\Sigma_{1,2} \otimes I_{n_2} + I_{n_1} \otimes \log\Sigma_{2,2}) \\
&= \big(\log\Sigma_{1,1} + \log\Sigma_{1,2}\big) \otimes I_{n_2} + I_{n_1} \otimes \big(\log\Sigma_{2,1} + \log\Sigma_{2,2}\big) \in \mathcal{D}_n.
\end{aligned}$$

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

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

Note also that since $\log\Sigma^{-1} = -\log\Sigma$, 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 form. 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.

3.2 Large n Approximation Properties

We next consider what happens to the eigenstructure of large covariance matrices. In general, a covariance matrix can have a wide variety of eigenstructures. Suppose we have a sequence of covariance matrix Σ_n with $\lambda_{n,i}$ denoting its ith largest eigenvalue. Let $\omega_{n,i} := \log \lambda_{n,i}$ and suppose that there exists a bounded continuous decreasing function $\omega(\cdot)$ defined on $[0,1]$ such that $\omega(i/n) = \omega_{n,i}$ for $i = 1,\ldots,n$. We may further suppose without loss of generality that $\omega(0) = 1$ and $\omega(1) = 0$, but otherwise ω can be anything. In some sense this class of matrices Σ_n is very large with potentially a very rich eigenstructure. On the other hand, 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 $n = 2^{v_n}$. We consider the 2×2 matrices $\{\Sigma_j^n : j = 1,2,\ldots,v_n\}$. Let $\overline{\omega}_j^n$ and $\underline{\omega}_j^n$ denote the logarithms of the larger and smaller eigenvalues of Σ_j^n, respectively. The logarithms of the eigenvalues of the Kronecker product matrix

$$\Sigma_n^* = \Sigma_1^n \otimes \cdots \otimes \Sigma_{v_n}^n$$

are of the form $\sum_{j=1}^{v_n} l_j$, where $l_j \in \{\overline{\omega}_j^n, \underline{\omega}_j^n\}$ for $j = 1,\ldots,v_n$. That is, the logarithms of the eigenvalues of Σ_n^* of the form

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

for some $I \subset \{1,2,\ldots,v_n\}$. In fact, we can think of l_j as a binary random variable that takes the two values with equal probability. Therefore, we may expect that

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

as $n \to \infty$. This says that the spectral distribution of Σ_n^* can be represented by the cumulative distribution function of the log normal distribution whose mean parameter is $\sum_{j=1}^{v_n}\mathbb{E}l_j$ and variance parameter $\sum_{j=1}^{v_n}\mathrm{var}(l_j)$, provided these two quantities stabilize. For example, suppose that $\overline{\omega}_j^n + \underline{\omega}_j^n = 0$, then $\mathbb{E}l_j = 0$, and provided

$$\sum_{j=1}^{v_n}\mathrm{var}(l_j) = \frac{1}{2}\sum_{j=1}^{v_n}\left[\left(\overline{\omega}_j^n\right)^2 + \left(\underline{\omega}_j^n\right)^2\right] \to c \in (0,\infty), \tag{3.3}$$

as $n \to \infty$, then the conditions of the CLT are satisfied. For example, suppose that $\overline{\omega}_j^n = v_n^{-1/2}\phi\left(j/v_n\right)$ and $\underline{\omega}_j^n = v_n^{-1/2}\mu\left(j/v_n\right)$ for some fixed decreasing functions $\phi(\cdot),\mu(\cdot)$ such that $\int\left(\phi(u)+\mu(u)\right)du = 0$. Then (3.3) is satisfied with $c = \frac{1}{2}\int\left(\phi^2(u)+\mu^2(u)\right)du$.

This says that the class of eigenstructures generated by the Kronecker parameterization can be quite general, and is determined by the mean and variance of the logarithms of the eigenvalues of the low dimensional matrices.

3.3 Portfolio Choice

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

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

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

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

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

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

which is very easy to compute.

4 Estimation of the Correlation Matrix Θ

We now 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 $\rho_1, \ldots, \rho_v$ in Θ^0 in (2.2) in the case where $n, T \to \infty$ simultaneously, i.e., joint asymptotics (Phillips and Moon (1999)). We achieve dimension reduction because originally Θ has $n(n-1)/2$ parameters whereas Θ^0 has only $v = O(\log n)$ parameters.

To study the theoretical properties of our model, we assume both μ and D are *known*. The case where D is unknown is considerably much more difficult. Not only it will affect the information bound for $\rho_1, \ldots, \rho_v$ in the maximum likelihood, but also has a non-trivial impact on the derivation of the asymptotic distribution of the minimum distance estimator due to its growing dimension.

Let $\rho := (\rho_1, \ldots, \rho_2)^\intercal \in \mathbb{R}^v$. Recall that Ω_1 in (2.4) has two distinct parameters a_1 and b_1. We denote similarly for $\Omega_2, \ldots, \Omega_v$. As mentioned before, we will not utilise the information about the signs of the entries of Ω_i for $i = 1, \ldots, v$. Define $\theta^\dagger := (a_1, b_1, a_2, b_2, \ldots, a_v, b_v)^\intercal \in \mathbb{R}^{2v}$. Note that

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

The same principle applies to $\Omega_2, \ldots, \Omega_v$. By (2.3) and Proposition 5 in Appendix A, we have

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

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

where E_i for $i = 1, \ldots, v$ are defined in (9.1). That is, the log correlation matrix Ω^0 obeys a linear model. We next give two examples.

Example 1 ($v = 2$)**.**

$$\Omega_1 = \log \Theta_1^0 = \begin{pmatrix} a_1 & b_1 \\ b_1 & a_1 \end{pmatrix} \qquad \Omega_2 = \log \Theta_2^0 = \begin{pmatrix} a_2 & b_2 \\ b_2 & a_2 \end{pmatrix}.$$

In this simple case, the matrix E_* takes the following form

$$\begin{aligned} vech(\Omega^0) &= vech(\Omega_1 \otimes I_2 + I_2 \otimes \Omega_2) \\ &= vech \begin{pmatrix} a_1 + a_2 & b_2 & b_1 & 0 \\ b_2 & a_1 + a_2 & 0 & b_1 \\ b_1 & 0 & a_1 + a_2 & 0 \\ 0 & b_1 & b_2 & a_1 + a_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ b_1 \\ a_2 \\ b_2 \end{pmatrix} \\ &=: E_* \begin{pmatrix} a_1 \\ b_1 \\ a_2 \\ b_2 \end{pmatrix}. \end{aligned}$$

$E_*^\intercal E_*$ is a 4×4 matrix:

$$E_*^\intercal E_* = \begin{pmatrix} 4 & 0 & 4 & 0 \\ 0 & 2 & 0 & 0 \\ 4 & 0 & 4 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

Example 2 (v=3)**.**

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

Now

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

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

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

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 summarise 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 in vech(Ω^0).

(iii) 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 also deduce that $\text{rank}(E_*^\intercal E_*) = \text{rank}(E_*) = v + 1 = 4$.

(iv) 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 recognise 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 summarise 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 in vech(Ω^0).

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

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

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

Proof. See Appendix A. □

Based on Example 1 or 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 leave E_* as it is and use Moore-Penrose generalised inverse to invert $E_*^\intercal E_*$ for estimation, but this creates unnecessary technicality in the proofs for the asymptotics. A better alternative is to re-parametrise

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

where $\theta := (\sum_{i=1}^{v} a_i, b_1, \ldots, b_v)^\intercal$ and E is the $n(n+1)/2 \times (v+1)$ submatrix of E_* after deleting the duplicate columns. Then we have the following proposition.

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

(i) $\text{rank}(E^\intercal E) = \text{rank}(E^\intercal) = \text{rank}(E)$ *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*

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

Proof. Follows trivially from Proposition 1. □

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

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

Any two of the preceding three allow us to recover ρ. We shall not address this over-identification issue in this paper.

4.1 Maximum Likelihood

The Gaussian QMLE is a natural starting point for estimation here. For maximum likelihood estimation to make sense, we impose the Kronecker product structure Θ^0 on the true correlation matrix Θ; that is, $\Theta = \Theta^0$. Then the log likelihood function for a sample $\{x_1, x_2, \ldots, x_T\} \subset \mathbb{R}^n$ is given by

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

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

$$\hat{\rho}_{MLE} = \arg\max_{\rho} \ell_T(\rho).$$

Substituting $\Theta = \Theta^0 = \exp(\Omega^0)$ (see (2.3)) into the log likelihood function, we have

$$\ell_T(\theta) = \\ -\frac{Tn}{2}\log(2\pi) - \frac{T}{2}\log\left|D^{1/2}\exp(\Omega^0(\theta))D^{1/2}\right| - \frac{1}{2}\sum_{t=1}^{T}(x_t-\mu)^\intercal D^{-1/2}[\exp(\Omega^0(\theta))]^{-1}D^{-1/2}(x_t-\mu), \tag{4.3}$$

where the parametrisation of Ω^0 in terms of θ is due to (4.2). We may define

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

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

To compute the MLE 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 MLE and so it is not so easy to analyse from a theoretical point of view.

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

4.2 The Minimum Distance Estimator

Define the sample second moment matrix

$$M_T := D^{-1/2}\left[\frac{1}{T}\sum_{t=1}^{T}(x_t-\mu)(x_t-\mu)^\intercal\right]D^{-1/2} =: D^{-1/2}\tilde{\Sigma}D^{-1/2}, \tag{4.4}$$

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

$$\hat{\theta}(W) := \arg\min_{\theta\in\mathbb{R}^{v+1}}[\text{vech}(\log M_T) - E\theta]^\intercal W[\text{vech}(\log M_T) - E\theta],$$

where the matrix E is defined in Proposition 2. This has a closed form solution

$$\hat{\theta}(W) = (E^\intercal WE)^{-1}E^\intercal W\text{vech}(\log M_T).$$

In the interest of space, we will only consider the special case of the identity weighting matrix ($W = I$)

$$\hat{\theta}_T := (E^\intercal E)^{-1}E^\intercal\text{vech}(\log M_T).$$

5 Asymptotic Properties

We derive the large sample properties of two estimators, the identity weighted minimum distance estimator $\hat{\theta}_T$ and the one-step QMLE which we define below, since the one-step QMLE depends on the properties of $\hat{\theta}_T$. We consider the case where $n, T \to \infty$ simultaneously.

5.1 The Identity-Weighted Minimum Distance Estimator

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 the above by an absolute constant. Further suppose*

$$\left\| [t(A - I) + I]^{-1} tB \right\|_{\ell_2} \leq C < 1 \tag{5.1}$$

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

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

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

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

for positive constants K and C.

Assumption 1 is standard in high-dimensional theoretical work. We are aware that financial data often exhibit heavy tails, hence violating Assumption 1. However Assumption 1 is not necessary as we mainly use it to invoke some concentration inequality such as Bernstein's inequality in Appendix B. Concentration inequalities corresponding to weaker versions of Assumption 1 do exist. Since the proofs of the asymptotics are already quite involved, we stick with Assumption 1.

Assumption 2.

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

(ii) $n, T \to \infty$ simultaneously.

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

Assumption 2(i) is for the rate of convergence of the minimum distance estimator $\hat{\theta}_T$ (Theorem 1). Assumption 2(ii) is *sufficient* for the asymptotic normality of both $\hat{\theta}_T$ (Theorem 2) and the one-step estimator $\tilde{\theta}_T$ (Theorem 4).

Assumption 3.

(i) Recall that $D := \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$*, 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 away from the 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 the 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), ensure 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 by Observation 7.1.6 in Horn and Johnson (1985) p399. Similarly, Assumptions 3(i)-(ii) ensure that Θ has maximum eigenvalue bounded away from the 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, 2(i), and 3 hold. We have*

(i)

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

(ii) Then (5.1) is satisfied with probability approaching 1 for $A = \Theta$ *and* $B = M_T - \Theta$*. That is,*

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

for some constant C*.*

Proof. See Appendix A. □

Assumption 4. *Suppose* $M_T := D^{-1/2}\tilde{\Sigma}D^{-1/2}$ *in (4.4) 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 the 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, 2(i), 3, and 4 be satisfied. Then*

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

where $\theta = (E^\intercal E)^{-1} E^\intercal \text{vech}(\log \Theta)$.

Proof. See Appendix A. □

Note that θ contains the unique parameters of the Kronecker product Θ^0 which we use to approximate the true correlation matrix Θ. The dimension of θ 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 $\sqrt{n^2/T}$ (square root of summing up n^2 terms each of which has rate $1/T$ via central limit theorem). We have some rate improvement in Theorem 1 as compared to this crude rate.

However, given the dimension of θ, one would conjecture that the optimal rate of convergence should be $(\log n/T)^{1/2}$. The reason for the rate difference lies in nonlinearity of matrix logarithm. Linearisation of matrix logarithm introduced a non-sparse Frechet derivative matrix, sandwiched by the sparse matrix $E^\intercal D_n^+$ and the vector $\text{vec}(M_T - \Theta)$. As a result, we were not able to use the sparse structure of $E^\intercal$ except the information about eigenvalues (Proposition 2(iii)). Had one made some assumption directly on the entries of matrix logarithm, we conjecture that one would achieve a better rate.

To derive the asymptotic distribution of the minimum distance estimator $\hat{\theta}_T$, we make the following assumption to simplify the derivation.

Assumption 5. $\{x_t\}_{t=1}^T$ *are normally distributed.*

Assumption 5 is rather innocuous given that we already have Assumption 1. We would like to stress that it is not necessary for the derivation of asymptotic normality of $\hat{\theta}_T$. All the arguments go through without normality assumption but will be more involved.

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

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

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

respectively.[3] Define the $n^2 \times n^2$ matrix

$$
\begin{aligned}
V &:= \operatorname{var}\left(\sqrt{T}\operatorname{vec}(\tilde{\Sigma} - \Sigma)\right) \\
&= \operatorname{var}\left(\sqrt{T}\operatorname{vec}\left(\frac{1}{T}\sum_{t=1}^{T}(x_t - \mu)(x_t - \mu)^\intercal - \mathbb{E}(x_t - \mu)(x_t - \mu)^\intercal\right)\right) \\
&= \frac{1}{T}\operatorname{var}\left(\operatorname{vec}\sum_{t=1}^{T}\left((x_t - \mu)(x_t - \mu)^\intercal - \mathbb{E}(x_t - \mu)(x_t - \mu)^\intercal\right)\right) = \operatorname{var}\left(\operatorname{vec}(x_t - \mu)(x_t - \mu)^\intercal\right) \\
&= \operatorname{var}\left((x_t - \mu) \otimes (x_t - \mu)\right) = 2D_n D_n^+(\Sigma \otimes \Sigma),
\end{aligned}
$$

where the second last equality is due to independence, and the last equality is due to Magnus and Neudecker (1986) Lemma 9.

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

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

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

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

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

$$
\hat{G}_T := 2c^\intercal(E^\intercal E)^{-1}E^\intercal D_n^+ \hat{H}_T(M_T \otimes M_T)\hat{H}_T D_n^{+\intercal}E(E^\intercal E)^{-1}c.
$$

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

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

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

Proof. See Appendix A. □

5.2 An Approximation to the Maximum Likelihood Estimator

We first define the score function and Hessian function of (4.3), which we give in the theorem below, since it is a non-trivial calculation.

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

Theorem 3. *The score function takes the following form*

$$\frac{\partial \ell_T(\theta)}{\partial \theta^\intercal} =$$
$$\frac{T}{2}E^\intercal D_n^\intercal \int_0^1 e^{t\Omega^0} \otimes e^{(1-t)\Omega^0} dt \left[vec\left([\exp(\Omega^0)]^{-1} D^{-1/2} \tilde{\Sigma} D^{-1/2} [\exp(\Omega^0)]^{-1} - \left[\exp(\Omega^0)\right]^{-1} \right) \right],$$

where $\tilde{\Sigma}$ is defined in (4.4). The Hessian matrix takes the following form

$$\mathcal{H}(\theta) = \frac{\partial^2 \ell_T(\theta)}{\partial \theta \partial \theta^\intercal} =$$
$$\begin{aligned}
&-\frac{T}{2}E^\intercal D_n^\intercal \Psi_1 \left([\exp \Omega^0]^{-1} D^{-1/2} \tilde{\Sigma} D^{-1/2} \otimes I_n + I_n \otimes [\exp \Omega^0]^{-1} D^{-1/2} \tilde{\Sigma} D^{-1/2} - I_{n^2} \right) \cdot \\
&\quad \left([\exp \Omega^0]^{-1} \otimes [\exp \Omega^0]^{-1} \right) \Psi_1 D_n E \\
&\quad + \frac{T}{2} (\Psi_2^\intercal \otimes E^\intercal D_n^\intercal) \int_0^1 P \left(I_{n^2} \otimes vec e^{(1-t)\Omega^0} \right) \int_0^1 e^{st\Omega^0} \otimes e^{(1-s)t\Omega^0} ds \cdot t dt D_n E \\
&\quad + \frac{T}{2} (\Psi_2^\intercal \otimes E^\intercal D_n^\intercal) \int_0^1 P \left(vec e^{t\Omega^0} \otimes I_{n^2} \right) \int_0^1 e^{s(1-t)\Omega^0} \otimes e^{(1-s)(1-t)\Omega^0} ds \cdot (1-t) dt D_n E.
\end{aligned}$$

where

$$\Psi_1 := \int_0^1 e^{t\Omega^0} \otimes e^{(1-t)\Omega^0} dt,$$
$$\Psi_2 := vec\left([\exp \Omega^0]^{-1} D^{-1/2} \tilde{\Sigma} D^{-1/2} [\exp \Omega^0]^{-1} - \left[\exp \Omega^0\right]^{-1} \right),$$
$$P := I_n \otimes K_{n,n} \otimes I_n.$$

Proof. See Appendix A. □

Note that $\mathbb{E}\Psi_2 = 0$, so the normalized Hessian matrix taken expectation at θ takes the following form

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

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

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

Using $\hat{\Psi}_{1,T}$ to estimate Ψ_1 does not utilise the information that Θ now has a Kronecker product structure (2.2). An alternative choice of an estimate for Θ in Ψ_1 could be the one formulated from the minimum distance estimator $\hat{\theta}_T$ taking into account of the Kronecker product structure. However, different choices should not matter for the asymptotic distribution.

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

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

We will 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 $\|b\|_2 = 1$,*

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

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 or 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, with derivative tending to Υ in probability, but this is not an assumption. The radius of the shrinking neighbourhood $\sqrt{n/T}$ 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 $o_p(n^{1/2})$ if one looks at the proof of Theorem 4.

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

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

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

Proof. See Appendix A. ☐

Theorem 4 says that $\sqrt{T} b^\intercal (\tilde{\theta}_T - \theta) \xrightarrow{d} N\left(0, b^\intercal \left(-\mathbb{E}\mathcal{H}(\theta)/T\right)^{-1} b\right)$. 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 (4.3) with respect to both θ and the diagonal elements of D. The working becomes considerably more involved and we leave it for the future work.

Remark. We may consider the choice of factorization in (2.1). Suppose that $n = 2^v$ for some positive v integer. Then there are several different Kronecker factorizations, which can be described by the dimensions of the square submatrices. That is,

$$\underbrace{2 \times 2 \times \cdots \times 2}_{v \text{ times}}, \quad \underbrace{2 \times 2 \times \cdots \times 2}_{v-2 \text{ times}} \times 4, \quad \ldots, \quad \underbrace{4 \times 4 \times \cdots \times 4}_{v/2 \text{ times}}, \quad \ldots,$$

$$2 \times 2^{v-1}, \quad \ldots, \quad 2^v,$$

have varying numbers of parameters. We might choose between these using some model choice criterion that penalizes the larger models. For example,

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

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

Remark. The Kronecker structure is not invariant with respect to permutations of the series in the system. One may want to find the permutation that is "closest" to a Kronecker structure in a certain sense. For example, one could estimate the model for many permutations and choose the one that maximizes the likelihood. In practice it may not be feasible to compute the likelihood for all possible permutations, as there are too many if the dimension is high ($n!$). One possibility is to use random permutations and choose the one that is closest to a Kronecker structure according to the likelihood criterion. 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.

6 Simulation Study

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

6.1 Kronecker Structure Is True

We simulate T random vectors y of dimension n according to

$$\begin{aligned} y_t &= \Sigma^{1/2}\xi_t, \quad \xi_t \sim N(0, I_n) \\ \Sigma &= \Sigma_1 \otimes \Sigma_2 \otimes \cdots \otimes \Sigma_v, \end{aligned}$$

where $n = 2^v$ and $v \in \mathbb{N}$. The matrices Σ_j are (2×2). These matrices 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 $n = 300$. 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 wrt the sample covariance estimator, S. It is defined as

$$PRIAL1 = 1 - \mathbb{E}||\hat{\Sigma} - \Sigma||^2 / \mathbb{E}||S - \Sigma||^2$$

where Σ is the true covariance matrix generated as above, $\hat{\Sigma}$ is the maximum likelihood estimator, and S is the sample covariance matrix. Often the estimator of the precision

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: *Median over 1000 replications of the PRIAL1 and PRIAL2 criteria for Kronecker estimates wrt to sample covariance matrix in the case of true non-Kronecker structure. VR is median of the ratio (multiplied by 100) of the variance of the MVP using Kronecker factorization to that using the sample covariance estimator. The sample size is fixed at* $T = 300$.

matrix, Σ^{-1}, is more important than that of Σ itself, so we also compute the PRIAL for the inverse covariance matrix, i.e.

$$PRIAL2 = 1 - \mathbb{E}||\hat{\Sigma}^{-1} - \Sigma^{-1}||^2 / \mathbb{E}||S^{-1} - \Sigma^{-1}||^2$$

Note that this requires invertibility of the sample covariance matrix S 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. These weights are applied to construct a portfolio from out-of-sample returns generated with the same distribution and the same covariance matrix as the in-sample returns. The first portfolio uses the sample covariance matrix, the second the Kronecker factorized matrix, and the ratio of the variances VR of the two portfolios is recorded.

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 RV for each dimension. 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.

6.2 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. The mean of the true eigenvalues is, w.l.o.g., fixed at one, while their dispersion varies and is given by α^2.

Figure 1 depicts the PRIAL1 of the Kronecker estimator for three different ratios n/T: 1/2, 1 and 2, as a function of the dispersion of eigenvalues of the true covariance matrix. Clearly, the PRIAL1 is higher for higher ratios n/T, while it decreases as the eigenvalue dispersion increases. The reason for the latter is that the distance from a Kronecker structure increases with the eigenvalue dispersion, so that the bias of the Kronecker estimator becomes more important. At the other extreme with no dispersion (α close to zero), the Kronecker structure is true, so that the PRIAL1 is high. Note that this behavior of the Kronecker estimator as a function n/T and α resembles that of the shrinkage estimator of Ledoit and Wolf (2004).

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: *Median over 1000 replications of the PRIAL1 and PRIAL2 criteria for Kronecker estimates wrt to sample covariance matrix 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 Kronecker factorization to that using the sample covariance estimator.* α^2 *is the dispersion of eigenvalues of the true covariance matrix.*

Figure 2 depicts the condition number of the estimated (Kronecker and sample) and true covariance matrices. Again as in the shrinkage case, the condition number of the Kronecker estimator is smaller than that of the true one, while the condition number of the sample covariance estimator increases strongly with the eigenvalue dispersion.

The precision matrix can not be estimated by the inverse of the sample covariance matrix in cases where $n \geq T$). We therefore consider the additional case $n/T = 0.8$ and report in Table 2 the results for the PRIALs and the variance ratios of MVP. First of all, the relative performance of the Kronecker precision matrix estimator is better than that of the covariance matrix itself, comparing PRIAL2 with PRIAL1. Thus, even in cases where the sample covariance matrix has a smaller average squared loss than the Kronecker estimator, it often occurs that the inverse of the Kronecker estimator has a smaller average squared loss for the precision matrix.

While the variance ratios (VR) indicate that for large eigenvalue dispersions and non-Kronecker structures it may be better to use the sample covariance matrix (i.e., VR is larger than 100%), this result diminishes as n/T approaches one.

7 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 maximum likelihood using the standardized returns, $D^{-1/2}x_t$. 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 prime factorizations of the numbers from 441 to 512, and selecting those yielding not more than 30 parameters. Table 2 gives a summary of these models including estimation results. The Schwarz information criterion favors the specification of model M6 with 27 parameters.

					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.*

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 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 2 summarize the relative performance of the Kronecker model with respect to SFM and the sample covariance matrix.

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 a non-sparse alternative to high dimensional covariance modelling.

8 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.

9 Appendix A

9.1 More Details about the Matrix E_*

Proposition 5. *If*

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

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

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

where

$$E_i := D_n^+(I_{2^i} \otimes K_{2^{v-i},2^i} \otimes I_{2^{v-i}}) \left(I_{2^{2i}} \otimes vecI_{2^{v-i}}\right) (I_{2^{i-1}} \otimes K_{2,2^{i-1}} \otimes I_2)(vecI_{2^{i-1}} \otimes I_4)D_2, \quad (9.1)$$

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 $\text{vec}(\Omega_1 \otimes I_2 \otimes \cdots \otimes I_2)$.

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

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

$$\text{vech}(\Omega_1 \otimes I_2 \otimes \cdots \otimes I_2) = D_n^+ \left(I_2 \otimes K_{2^{v-1},2} \otimes I_{2^{v-1}}\right) \left(I_4 \otimes \text{vec}I_{2^{v-1}}\right) D_2 \text{vech}\Omega_1, \quad (9.2)$$

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 $\text{vec}(I_2 \otimes \Omega_2 \otimes \cdots \otimes I_2)$.

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

Thus

$$\begin{aligned}&\mathrm{vech}(I_2 \otimes \Omega_2 \otimes \cdots \otimes I_2)\\&= D_n^+(I_4 \otimes K_{2^{v-2},4} \otimes I_{2^{v-2}}) \left(I_{2^4} \otimes \mathrm{vec} I_{2^{v-2}}\right) (I_2 \otimes K_{2,2} \otimes I_2)(\mathrm{vec} I_2 \otimes I_4) D_2 \mathrm{vech}\Omega_2.\end{aligned} \tag{9.3}$$

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

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

Thus

$$\begin{aligned}&\mathrm{vech}(I_2 \otimes I_2 \otimes \Omega_3 \otimes \cdots \otimes I_2)\\&= D_n^+(I_{2^3} \otimes K_{2^{v-3},2^3} \otimes I_{2^{v-3}}) \left(I_{2^6} \otimes \mathrm{vec} I_{2^{v-3}}\right) (I_4 \otimes K_{2,4} \otimes I_2)(\mathrm{vec} I_4 \otimes I_4) D_2 \mathrm{vech}\Omega_3.\end{aligned} \tag{9.4}$$

By observing (9.2), (9.3) and (9.4), we deduce the following general formula: for $i = 1, 2, \ldots, v$

$$\begin{aligned}&\mathrm{vech}(I_2 \otimes \cdots \otimes \Omega_i \otimes \cdots \otimes I_2)\\&= D_n^+(I_{2^i} \otimes K_{2^{v-i},2^i} \otimes I_{2^{v-i}}) \left(I_{2^{2i}} \otimes \mathrm{vec} I_{2^{v-i}}\right) (I_{2^{i-1}} \otimes K_{2,2^{i-1}} \otimes I_2)(\mathrm{vec} I_{2^{i-1}} \otimes I_4) D_2 \mathrm{vech}\Omega_i\\&=: E_i \mathrm{vech}\Omega_i,\end{aligned} \tag{9.5}$$

where E_i is a $n(n+1)/2 \times 3$ matrix. Using (9.5), we have

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

□

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) - (iv) are similar to those in the observations made in Example 2. □

9.2 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 5 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 (5.1) to invoke Proposition 13 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 that $\text{maxeval}(A) < C < \infty$ and $\text{mineval}[t(A-I)+I] > C' > 0$. □

9.3 Proof of Proposition 4

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

Proposition 6. *Suppose Assumptions 1, 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, $\{z_{a,t}\}_{t=1}^T$ are independent subgaussian random variables. For $\epsilon > 0$,

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

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.

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

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, $\{z_{a,t}\}_{t=1}^T$ 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

$$\max_{a\in\mathcal{N}_{1/4}}|(\hat{\mu}-\mu)^\intercal a|=O_p\left(\sqrt{\frac{n}{T}}\right).\tag{9.6}$$

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

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

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 (9.6) deliver the result. For part (iv), via Assumption 1, 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),

$$\begin{aligned}
&\|M_T - \Theta\|_{\ell_2} = \|D^{-1/2}\tilde{\Sigma}D^{-1/2} - D^{-1/2}\Sigma D^{-1/2}\|_{\ell_2} = \|D^{-1/2}(\tilde{\Sigma} - \Sigma)D^{-1/2}\|_{\ell_2} \\
&\leq \|D^{-1/2}\|_{\ell_2}^2 \|\tilde{\Sigma} - \Sigma\|_{\ell_2} = O(1)\|\tilde{\Sigma} - \Sigma\|_{\ell_2} \\
&= O(1)\left\| \frac{1}{T}\sum_{t=1}^{T} x_t x_t^\intercal - \mathbb{E}x_t x_t^\intercal + 2\mu\mu^\intercal - \hat{\mu}\mu^\intercal - \mu\hat{\mu}^\intercal \right\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right),
\end{aligned} \tag{9.7}$$

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 that mineval$(t(\Theta - I) + I) > C > 0$ for some absolute constant C and Assumption 2(i). □

9.4 Proof of Theorem 1

Proof.

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

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 $B = M_T - \Theta$:

$$\log M_T - \log \Theta = \int_0^1 [t(\Theta - I) + I]^{-1}(M_T - \Theta)[t(\Theta - I) + I]^{-1} dt + O_p(\|M_T - \Theta\|_{\ell_2}^2). \tag{9.8}$$

(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

$$\begin{aligned}
&\|\text{vec}(\log M_T - \log \Theta)\|_2 \\
&\leq \left\| \int_0^1 [t(\Theta - I) + I]^{-1} \otimes [t(\Theta - I) + I]^{-1} dt \text{vec}(M_T - \Theta) \right\|_2 + \|\text{vec} O_p(\|M_T - \Theta\|_{\ell_2}^2)\|_2 \\
&\leq \left\| \int_0^1 [t(\Theta - I) + I]^{-1} \otimes [t(\Theta - I) + I]^{-1} dt \right\|_{\ell_2} \|M_T - \Theta\|_F + \|O_p(\|M_T - \Theta\|_{\ell_2}^2)\|_F \\
&\leq C\sqrt{n}\|M_T - \Theta\|_{\ell_2} + \sqrt{n}\|O_p(\|M_T - \Theta\|_{\ell_2}^2)\|_{\ell_2} \\
&\leq C\sqrt{n}\|M_T - \Theta\|_{\ell_2} + \sqrt{n} O_p(\|M_T - \Theta\|_{\ell_2}^2) = O_p(\sqrt{n^2/T}),
\end{aligned} \tag{9.9}$$

where the third inequality is due to (9.12), and the last inequality is due to Proposition 4. Finally,

$$\|(E^\intercal E)^{-1}E^\intercal\|_{\ell_2} = \sqrt{\text{maxeval}\left(\left[(E^\intercal E)^{-1}E^\intercal\right]^\intercal (E^\intercal E)^{-1}E^\intercal\right)}$$
$$= \sqrt{\text{maxeval}\left((E^\intercal E)^{-1}E^\intercal \left[(E^\intercal E)^{-1}E^\intercal\right]^\intercal\right)} = \sqrt{\text{maxeval}\left((E^\intercal E)^{-1}E^\intercal E(E^\intercal E)^{-1}\right)}$$
$$= \sqrt{\text{maxeval}\left((E^\intercal E)^{-1}\right)} = \sqrt{2/n}, \tag{9.10}$$

where the second equality is due to 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

$$\|D_n^+\|_{\ell_2} = \|D_n^{+\intercal}\|_{\ell_2} = O(1), \qquad \|D_n\|_{\ell_2} = \|D_n^\intercal\|_{\ell_2} = O(1). \tag{9.11}$$

The result follows after assembling the rates. □

9.5 Proof of Theorem 2

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

$$\|H\|_{\ell_2} = O(1), \qquad \|\hat{H}_T\|_{\ell_2} = O_p(1), \qquad \|\hat{H}_T - H\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right). \tag{9.12}$$

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 $B_t := [t(\Theta - I) + I]^{-1}$.

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

where the second equality is to Proposition 14 in Appendix B, and the last equality is due to Assumption 4. Now,

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

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 14 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. $\square$

Proof of Theorem 2.

$$\begin{aligned}\frac{\sqrt{T}c^\intercal(\hat{\theta}_T - \theta)}{\sqrt{\hat{G}_T}} &= \frac{\sqrt{T}c^\intercal(E^\intercal E)^{-1}E^\intercal D_n^+ H(D^{-1/2}\otimes D^{-1/2})\text{vec}(\tilde{\Sigma}-\Sigma)}{\sqrt{\hat{G}_T}} \\ &\quad + \frac{\sqrt{T}c^\intercal(E^\intercal E)^{-1}E^\intercal D_n^+ \text{vec} O_p(\|M_T - \Theta\|_{\ell_2}^2)}{\sqrt{\hat{G}_T}} \\ &=: t_1 + t_3.\end{aligned}$$

Define

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

To prove Theorem 2, it suffices to show $t_1' \xrightarrow{d} N(0,1)$, $t_1' - t_1 = o_p(1)$, and $t_3 = o_p(1)$.

9.5.1 $t_1' \xrightarrow{d} N(0,1)$

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

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

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 $\gamma > 2$,

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

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

$$\begin{aligned}G &= 2c^\intercal(E^\intercal E)^{-1}E^\intercal D_n^+ H(\Theta\otimes\Theta)HD_n^{+\intercal}E(E^\intercal E)^{-1}c\\ &\geq 2\text{mineval}(\Theta\otimes\Theta)\text{mineval}(H^2)\text{mineval}(D_n^+D_n^{+\intercal})\text{mineval}((E^\intercal E)^{-1})\\ &\geq 2\text{mineval}(\Theta\otimes\Theta)\text{mineval}(H^2)\text{mineval}(D_n^+D_n^{+\intercal})\frac{1}{n}\end{aligned}$$

where the first inequality is true by repeatedly invoking Rayleigh-Ritz theorem. Since the minimum eigenvalue of Θ is bounded away from zero by an absolute constant by Assumption 3(iii)-(iv), and the minimum eigenvalue of H is bounded away from zero by an absolute constant by Assumption 3(i)-(ii), and $\text{mineval}(D_n^+D_n^{+\intercal})=\text{mineval}[(D_n^\intercal D_n)^{-1}]>C>0$ for some absolute constant C, we have

$$\frac{1}{\sqrt{G}}=O(\sqrt{n}). \tag{9.13}$$

Then a sufficient condition for the Lindeberg condition is:

$$T^{1-\frac{\gamma}{2}}n^{\gamma/2}\mathbb{E}\left|c^\intercal(E^\intercal E)^{-1}E^\intercal D_n^+H(D^{-1/2}\otimes D^{-1/2})\text{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal-\mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right|^\gamma$$
$$=o(1), \tag{9.14}$$

for some $\gamma>2$. We now verify (9.14).

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

where the first equality is because of (9.10), (9.12), and Proposition 6(ii), the third inequality is due to the decoupling inequality $|f+g|^p\leq 2^{p-1}(|f|^p+|g|^p)$ for $p\geq 1$, the fourth inequality is due to Jensen's inequality, the fourth equality is due to the definition of L_p norm. By Assumption 1, 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

$$\left\|\max_{1\leq i,j\leq n}(x_t-\mu)_i(x_t-\mu)_j\right\|_{L_\gamma}^{\gamma} \leq (\gamma!)^\gamma \left\|\max_{1\leq i,j\leq n}(x_t-\mu)_i(x_t-\mu)_j\right\|_{\psi_1}^{\gamma} = O(\log^\gamma n). \quad (9.15)$$

Summing up the rates, we have

$$T^{1-\frac{\gamma}{2}}n^{\gamma/2}\mathbb{E}\left|c^\intercal(E^\intercal E)^{-1}E^\intercal D_n^+ H(D^{-1/2}\otimes D^{-1/2})\text{vec}\left[x_t-\mu)(x_t-\mu)^\intercal-\mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right|^\gamma$$
$$= T^{1-\frac{\gamma}{2}}n^{\gamma/2}O(n^{\gamma/2})O(\log^\gamma n) = O\left(\frac{n\log n}{T^{\frac{1}{2}-\frac{1}{\gamma}}}\right)^\gamma = o(1)$$

by Assumption 2(ii). Thus, we have verified (9.14).

9.5.2 $t_1' - t_1 = o_p(1)$

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

$$t_1' - t_1 = \frac{A}{\sqrt{G}} - \frac{A}{\sqrt{\hat{G}_T}} = \frac{A}{\sqrt{G}}\left(\frac{\sqrt{n\hat{G}_T}-\sqrt{nG}}{\sqrt{n\hat{G}_T}}\right)$$
$$= \frac{A}{\sqrt{G}}\frac{1}{\sqrt{n\hat{G}_T}}\left(\frac{n\hat{G}_T-nG}{\sqrt{n\hat{G}_T}+\sqrt{nG}}\right).$$

Since we have already shown in (9.13) that nG is bounded away from zero by an absolute constant and $A/\sqrt{G} = O_p(1)$, if in addition we show that $n\hat{G}_T - nG = o_p(1)$, then the right hand side of the preceding display is $o_p(1)$ by repeatedly invoking continuous mapping theorem. Now we show that $n\hat{G}_T - nG = o_p(1)$. Define

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

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

$$n|\hat{G}_T-\tilde{G}_T|$$
$$= 2n|c^\intercal(E^\intercal E)^{-1}E^\intercal D_n^+\hat{H}_T(M_T\otimes M_T)\hat{H}_T D_n^{+\intercal}E(E^\intercal E)^{-1}c$$
$$\quad - c^\intercal(E^\intercal E)^{-1}E^\intercal D_n^+\hat{H}_T(\Theta\otimes\Theta)\hat{H}_T D_n^{+\intercal}E(E^\intercal E)^{-1}c|$$
$$= 2n|c^\intercal(E^\intercal E)^{-1}E^\intercal D_n^+\hat{H}_T(M_T\otimes M_T-\Theta\otimes\Theta)\hat{H}_T D_n^{+\intercal}E(E^\intercal E)^{-1}c|$$
$$\leq 2n|\text{maxeval}(M_T\otimes M_T-\Theta\otimes\Theta)|\|\hat{H}_T D_n^{+\intercal}E(E^\intercal E)^{-1}c\|_2^2$$
$$= 2n\|M_T\otimes M_T-\Theta\otimes\Theta\|_{\ell_2}\|\hat{H}_T D_n^{+\intercal}E(E^\intercal E)^{-1}c\|_2^2 \leq 2nO_p\left(\sqrt{\frac{n}{T}}\right)\|\hat{H}_T\|_{\ell_2}^2\|D_n^{+\intercal}\|_{\ell_2}^2\|(E^\intercal E)^{-1}\|_{\ell_2}$$
$$= O_p\left(\sqrt{\frac{n}{T}}\right) = o_p(1),$$

where the second inequality is due to Proposition 4, Assumptions 3(i)-(ii) and 4, the second last equality is due to (9.12), (9.10), and (9.11, and the last equality is due to

Assumption 2(ii). We now prove $n|\tilde{G}_T - G| = o_p(1)$.

$$\begin{aligned}
&n|\tilde{G}_T - G| \\
&= 2n|c^\intercal(E^\intercal E)^{-1}E^\intercal D_n^+ \hat{H}_T(\Theta\otimes\Theta)\hat{H}_T D_n^{+\intercal} E(E^\intercal E)^{-1}c \\
&\quad - c^\intercal(E^\intercal E)^{-1}E^\intercal D_n^+ H(\Theta\otimes\Theta)HD_n^{+\intercal}E(E^\intercal E)^{-1}c| \\
&\leq 2n|\text{maxeval}(\Theta\otimes\Theta)|\|(\hat{H}_T - H)D_n^{+\intercal}E(E^\intercal E)^{-1}c\|_2^2 \\
&\quad + 2n\|(\Theta\otimes\Theta)HD_n^{+\intercal}E(E^\intercal E)^{-1}c\|_2\|(\hat{H}_T - H)D_n^{+\intercal}E(E^\intercal E)^{-1}c\|_2 \qquad (9.16)
\end{aligned}$$

where the inequality is due to Lemma 5 in Appendix B. We consider the first term of (9.16) first.

$$\begin{aligned}
&2n|\text{maxeval}(\Theta\otimes\Theta)|\|(\hat{H}_T - H)D_n^{+\intercal}E(E^\intercal E)^{-1}c\|_2^2 = O(n)\|\hat{H}_T - H\|_{\ell_2}^2\|D_n^{+\intercal}\|_{\ell_2}^2\|(E^\intercal E)^{-1}\|_{\ell_2} \\
&= O_p(n/T) = o_p(1),
\end{aligned}$$

where the second last equality is due to (9.12), (9.11), and (9.10), and the last equality is due to Assumption 2(ii). We now consider the second term of (9.16).

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

where the first equality is due to (9.12), (9.11), and (9.10), and the last equality is due to Assumption 2(ii). We have proved $|n\tilde{G}_T - nG| = o_p(1)$ and hence $|n\hat{G}_T - nG| = o_p(1)$.

9.5.3 $t_3 = o_p(1)$

Last, we prove that $t_3 = o_p(1)$. Write

$$t_3 = \frac{\sqrt{T}\sqrt{n}c^\intercal(E^\intercal E)^{-1}E^\intercal D_n^+ \text{vec}O_p(\|M_T - \Theta\|_{\ell_2}^2)}{\sqrt{n\hat{G}_T}}.$$

Since the denominator of the preceding equation is bounded away from zero by an absolute constant with probability approaching one by (9.13) and that $|n\hat{G}_T - nG| = o_p(1)$, it suffices to show

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

This is straightforward:

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

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

9.6 Proof of Theorem 3

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

$$
\begin{aligned}
&d\ell_T(\theta)\\
&= -\frac{T}{2}d\log\left|D^{1/2}\exp(\Omega^0)D^{1/2}\right| - \frac{T}{2}d\mathrm{tr}\left(\frac{1}{T}\sum_{t=1}^{T}(x_t-\mu)^\intercal D^{-1/2}[\exp(\Omega^0)]^{-1}D^{-1/2}(x_t-\mu)\right)\\
&= -\frac{T}{2}d\log\left|D^{1/2}\exp(\Omega^0)D^{1/2}\right| - \frac{T}{2}d\mathrm{tr}\left(D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1}\right)\\
&= -\frac{T}{2}\mathrm{tr}\left(\left[D^{1/2}\exp(\Omega^0)D^{1/2}\right]^{-1}D^{1/2}d\exp(\Omega^0)D^{1/2}\right) - \frac{T}{2}d\mathrm{tr}\left(D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1}\right)\\
&= -\frac{T}{2}\mathrm{tr}\left(\left[\exp(\Omega^0)\right]^{-1}d\exp(\Omega^0)\right) - \frac{T}{2}\mathrm{tr}\left(D^{-1/2}\tilde{\Sigma}D^{-1/2}d[\exp(\Omega^0)]^{-1}\right)\\
&= -\frac{T}{2}\mathrm{tr}\left(\left[\exp(\Omega^0)\right]^{-1}d\exp(\Omega^0)\right) + \frac{T}{2}\mathrm{tr}\left(D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1}d\exp(\Omega^0)[\exp(\Omega^0)]^{-1}\right)\\
&= \frac{T}{2}\mathrm{tr}\left(\left\{[\exp(\Omega^0)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1} - \left[\exp(\Omega^0)\right]^{-1}\right\}d\exp(\Omega^0)\right)\\
&= \frac{T}{2}\left[\mathrm{vec}\left(\left\{[\exp(\Omega^0)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1} - \left[\exp(\Omega^0)\right]^{-1}\right\}^\intercal\right)\right]^\intercal \mathrm{vec}d\exp(\Omega^0)\\
&= \frac{T}{2}\left[\mathrm{vec}\left([\exp(\Omega^0)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1} - \left[\exp(\Omega^0)\right]^{-1}\right)\right]^\intercal \mathrm{vec}d\exp(\Omega^0),
\end{aligned}
$$

where in the second equality we used the definition of $\tilde{\Sigma}$ (4.4), 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:

$$
d\exp(\Omega^0) = \int_0^1 e^{(1-t)\Omega^0}(d\Omega^0)e^{t\Omega^0}dt.
$$

Therefore,

$$
\begin{aligned}
\mathrm{vec}d\exp(\Omega^0) &= \int_0^1 e^{t\Omega^0}\otimes e^{(1-t)\Omega^0}dt\mathrm{vec}(d\Omega^0) = \int_0^1 e^{t\Omega^0}\otimes e^{(1-t)\Omega^0}dtD_n\mathrm{vech}(d\Omega^0)\\
&= \int_0^1 e^{t\Omega^0}\otimes e^{(1-t)\Omega^0}dtD_nEd\theta,
\end{aligned}
$$

where the last equality is due to (4.1). Hence,

$$
\begin{aligned}
&d\ell_T(\theta)\\
&= \frac{T}{2}\left[\mathrm{vec}\left([\exp(\Omega^0)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1} - \left[\exp(\Omega^0)\right]^{-1}\right)\right]^\intercal \int_0^1 e^{t\Omega^0}\otimes e^{(1-t)\Omega^0}dtD_nEd\theta
\end{aligned}
$$

and

$$
\begin{aligned}
y &:= \frac{\partial \ell_T(\theta)}{\partial \theta^\intercal} \\
&= \frac{T}{2} E^\intercal D_n^\intercal \int_0^1 e^{t\Omega^0} \otimes e^{(1-t)\Omega^0} dt \left[\mathrm{vec} \left([\exp(\Omega^0)]^{-1} D^{-1/2} \tilde{\Sigma} D^{-1/2} [\exp(\Omega^0)]^{-1} - \left[\exp(\Omega^0)\right]^{-1} \right) \right] \\
&=: \frac{T}{2} E^\intercal D_n^\intercal \Psi_1 \Psi_2.
\end{aligned}
$$

Now we derive the Hessian matrix.

$$
dy = \frac{T}{2} E^\intercal D_n^\intercal (d\Psi_1)\Psi_2 + \frac{T}{2} E^\intercal D_n^\intercal \Psi_1 d\Psi_2 = \frac{T}{2} (\Psi_2^\intercal \otimes E^\intercal D_n^\intercal) \mathrm{vec} d\Psi_1 + \frac{T}{2} E^\intercal D_n^\intercal \Psi_1 d\Psi_2. \quad (9.17)
$$

Consider $d\Psi_1$ first.

$$
\begin{aligned}
d\Psi_1 &= d \int_0^1 e^{t\Omega^0} \otimes e^{(1-t)\Omega^0} dt = \int_0^1 de^{t\Omega^0} \otimes e^{(1-t)\Omega^0} dt + \int_0^1 e^{t\Omega^0} \otimes de^{(1-t)\Omega^0} dt \\
&=: \int_0^1 A \otimes e^{(1-t)\Omega^0} dt + \int_0^1 e^{t\Omega^0} \otimes B dt,
\end{aligned}
$$

where

$$
A := \int_0^1 e^{(1-s)t\Omega^0} d(t\Omega^0) e^{st\Omega^0} ds, \quad B := \int_0^1 e^{(1-s)(1-t)\Omega^0} d((1-t)\Omega^0) e^{s(1-t)\Omega^0} ds.
$$

Therefore,

$$
\begin{aligned}
\mathrm{vec} d\Psi_1 &= \int_0^1 \mathrm{vec} \left(A \otimes e^{(1-t)\Omega^0} \right) dt + \int_0^1 \mathrm{vec} \left(e^{t\Omega^0} \otimes B \right) dt \\
&= \int_0^1 P \left(\mathrm{vec} A \otimes \mathrm{vec} e^{(1-t)\Omega^0} \right) dt + \int_0^1 P \left(\mathrm{vec} e^{t\Omega^0} \otimes \mathrm{vec} B \right) dt \\
&= \int_0^1 P \left(I_{n^2} \otimes \mathrm{vec} e^{(1-t)\Omega^0} \right) \mathrm{vec} A dt + \int_0^1 P \left(\mathrm{vec} e^{t\Omega^0} \otimes I_{n^2} \right) \mathrm{vec} B dt \\
&= \int_0^1 P \left(I_{n^2} \otimes \mathrm{vec} e^{(1-t)\Omega^0} \right) \int_0^1 e^{st\Omega^0} \otimes e^{(1-s)t\Omega^0} ds \cdot \mathrm{vec} d(t\Omega^0) dt \\
&\quad + \int_0^1 P \left(\mathrm{vec} e^{t\Omega^0} \otimes I_{n^2} \right) \int_0^1 e^{s(1-t)\Omega^0} \otimes e^{(1-s)(1-t)\Omega^0} ds \cdot \mathrm{vec} d((1-t)\Omega^0) dt \\
&= \int_0^1 P \left(I_{n^2} \otimes \mathrm{vec} e^{(1-t)\Omega^0} \right) \int_0^1 e^{st\Omega^0} \otimes e^{(1-s)t\Omega^0} ds \cdot t dt D_n E d\theta \\
&\quad + \int_0^1 P \left(\mathrm{vec} e^{t\Omega^0} \otimes I_{n^2} \right) \int_0^1 e^{s(1-t)\Omega} \otimes e^{(1-s)(1-t)\Omega^0} ds \cdot (1-t) dt D_n E d\theta \qquad (9.18)
\end{aligned}
$$

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 $d\Psi_2$.

$$\begin{aligned}
d\Psi_2 &= d\text{vec}\left([\exp(\Omega^0)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1} - \left[\exp(\Omega^0)\right]^{-1}\right) \\
&= \text{vec}\left(d[\exp(\Omega^0)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1}\right) \\
&\quad + \left([\exp(\Omega^0)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}d[\exp(\Omega^0)]^{-1}\right) - \text{vec}\left(d\left[\exp(\Omega^0)\right]^{-1}\right) \\
&= \text{vec}\left(-[\exp(\Omega^0)]^{-1}d\exp(\Omega^0)[\exp(\Omega^0)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1}\right) \\
&\quad + \text{vec}\left(-[\exp(\Omega^0)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1}d\exp(\Omega^0)[\exp(\Omega^0)]^{-1}\right) \\
&\quad + \text{vec}\left([\exp(\Omega^0)]^{-1}d\exp(\Omega^0)[\exp(\Omega^0)]^{-1}\right) \\
&= \left([\exp(\Omega^0)]^{-1} \otimes [\exp(\Omega^0)]^{-1}\right)\text{vec}d\exp(\Omega^0) \\
&\quad - \left([\exp(\Omega^0)]^{-1} \otimes [\exp(\Omega^0)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1}\right)\text{vec}d\exp(\Omega^0) \\
&\quad - \left([\exp(\Omega^0)]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2}[\exp(\Omega^0)]^{-1} \otimes [\exp(\Omega^0)]^{-1}\right)\text{vec}d\exp(\Omega^0) \qquad (9.19)
\end{aligned}$$

Substituting (9.18) and (9.19) into (9.17) yields the result:

$$\begin{aligned}
&\frac{\partial^2 \ell_T(\theta)}{\partial\theta\partial\theta^\intercal} = \\
&-\frac{T}{2}E^\intercal D_n^\intercal \Psi_1\left([\exp\Omega^0]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2} \otimes I_n + I_n \otimes [\exp\Omega^0]^{-1}D^{-1/2}\tilde{\Sigma}D^{-1/2} - I_{n^2}\right)\cdot \\
&\quad \left([\exp\Omega^0]^{-1} \otimes [\exp\Omega^0]^{-1}\right)\Psi_1 D_n E \\
&\quad + \frac{T}{2}(\Psi_2^\intercal \otimes E^\intercal D_n^\intercal)\int_0^1 P\left(I_{n^2} \otimes \text{vec}e^{(1-t)\Omega^0}\right)\int_0^1 e^{st\Omega^0} \otimes e^{(1-s)t\Omega^0}ds \cdot tdtD_nE \\
&\quad + \frac{T}{2}(\Psi_2^\intercal \otimes E^\intercal D_n^\intercal)\int_0^1 P\left(\text{vec}e^{t\Omega^0} \otimes I_{n^2}\right)\int_0^1 e^{s(1-t)\Omega^0} \otimes e^{(1-s)(1-t)\Omega^0}ds \cdot (1-t)dtD_nE.
\end{aligned}$$

□

9.7 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 the 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 the above by absolute constants. Define

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

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 8. *Suppose Assumptions 1, 2(i), 3 and 4 hold. Then Ψ_1 is positive definite for all n with its minimum eigenvalue bounded away from zero by an absolute constant and maximum eigenvalue bounded from the 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 the above by an absolute constant with probability approaching 1.*

Proof. For part (i), 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 9. *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 $\|a\|=1$

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

for some positive constant C. □

Proposition 10. *Let Assumptions 1, 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 14 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 11. *Let Assumptions 1, 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^0, 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 7 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 (9.8):

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

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 12. *Let Assumptions 1, 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 $\|\hat{\mathcal{X}}_T\|_{\ell_2}$.

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

where the last equality is due to Propositions 2(iii) and 8 and (9.11). Now

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

The proposition result (ii) follows after invoking Propositions 10 and 11. For part (iii),

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

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 $\|a\|_2 = 1$,

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

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

$$\|(-\hat{\Upsilon}_T)^{-1}\|_{\ell_2} = \frac{1}{\text{mineval}(-\hat{\Upsilon}_T)} = O_p(n^{-1}). \tag{9.20}$$

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

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

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

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

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

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

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

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

where the first equality is due to (9.20), Proposition 12 (iii) and Theorem 1, and the last equation is due to Assumption 2(ii). Thus

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

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

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

Define

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

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

9.7.1 $t'_{2,1} \xrightarrow{d} N(0,1)$

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

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

The proof is very similar to that of $t_1' \xrightarrow{d} N(0,1)$ in Section 9.5.1. 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

$$\frac{1}{\sqrt{b^\intercal(-\Upsilon)^{-1}b}} = O(\sqrt{n}). \tag{9.21}$$

Then a sufficient condition for the Lindeberg condition is:

$$T^{1-\frac{\gamma}{2}} n^{\gamma/2} \cdot$$
$$\mathbb{E}\left|b^\intercal(\mathcal{X}^\intercal\mathcal{X})^{-1}\mathcal{X}^\intercal(\Theta^{-1/2}\otimes\Theta^{-1/2})(D^{-1/2}\otimes D^{-1/2})\text{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal - \mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right|^{\gamma}$$
$$= o(1), \tag{9.22}$$

for some $\gamma > 2$. We now verify (9.22). We shall be concise as the proof is very similar to that in Section 9.5.1.

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

where the second last equality is due to Proposition 9 and the last equality is due to (9.15). 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 (9.22).

9.7.2 $t'_{2,1} - t_{2,1} = o_p(1)$

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 (9.21) 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 9. Then to show that $t'_{2,1} - t_{2,1} = o_p(1)$, it suffices to show

$$\sqrt{n}A - \sqrt{n}\hat{A} = o_p(1) \tag{9.23}$$

$$n\hat{G} - nG = o_p(1). \tag{9.24}$$

9.7.3 Proof of (9.23)

We now show that

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

This is straightforward.

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

where the second equality is due to Proposition 12(iv) and (9.7), and the last equality is due to Assumption 2(ii).

9.7.4 Proof of (9.24)

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 12(iv) and the last equality is due to Assumption 2(i).

9.7.5 $t_{2,2} = o_p(1)$

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 (9.21) and (9.24). □

10 Appendix B

10.1 Minimum Distance Estimator

Proposition 13. *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

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

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 5 (Higham (2008) p269; Dieci et al. (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 6 (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*

$$\|A - B\|_{\ell_2} \leq \frac{\|A^2 - B^2\|_{\ell_2}}{\mathit{mineval}(A) + \mathit{mineval}(B)}.$$

Proof. See Horn and Johnson (1985) p410. □

Proposition 14. *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 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*

$$|\hat{v}^\intercal A \hat{v} - v^\intercal A v| \leq |\mathit{maxeval}(A)|\|\hat{v} - v\|_2^2 + 2(\|Av\|_2\|\hat{v} - v\|_2).$$

Proof. See Lemma 3.1 in the supplementary material of van de Geer et al. (2014). □

10.2 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 7. *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. □

References

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.

Bühlmann, P. and S. van de Geer (2011): *Statistics for High-Dimensional Data*, Springer.

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.

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.

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.

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.

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.

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.

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.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, IV.

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.

Saikkonen, P. and H. Lutkepohl (1996): "Infinite-order Cointegrated Vector Autoregressive Processes," *Econometric Theory*, 12, 814–844.

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.

van der Vaart, A. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

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.*

Wansbeek, T. and E. Meijer (2000): *Measurement Error and Latent Variables in Econometrics*, Elsevier.