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

Christian M. Hafner† Université catholique de Louvain

Oliver B. Linton‡ University of Cambridge

Haihan Tang§ Fudan University

February 28, 2018

Abstract

We propose a Kronecker product model for correlation or covariance matrices in the large dimension case. The number of parameters of the model increases logarithmically with the dimension of the matrix. We propose a minimum distance (MD) estimator based on a log-linear property of the model, as well as a one-step estimator, which is a one-step approximation to the quasi-maximum likelihood estimator (QMLE). We establish the rate of convergence and a central limit theorem (CLT) for our estimators in the large dimensional case. A specification test and tools for Kronecker product model selection and inference are provided. In an empirical application to portfolio choice for S&P500 daily returns, we show that our model outperforms the sample covariance matrix and a linear shrinkage estimator.

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

JEL subject classification: C55, C58, G11

*We would like to thank Heather Battey, Michael I. Gil', Liudas Giraitis, Bowei Guo, Xumin He, Chen Huang, Anders Bredahl Kock, Alexei Onatskiy, Qi-Man Shao, Richard Smith, Chen Wang, Tengyao Wang, Tom Wansbeek, Jianbin Wu, Qiwei Yao for useful comments.

†Institut de statistique, biostatistique et sciences actuarielles, and CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium. Email: `christian.hafner@uclouvain.be`

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

§Corresponding author. Fanhai International School of Finance and School of Economics, Fudan University. Email: `hhtang@fudan.edu.cn`.

1 Introduction

Covariance and correlation matrices are of great importance in many fields. In finance, they are a key element in portfolio choice and risk management. In psychology, scholars have long been assuming that the observed variables are related to the unobserved traits such that factor models for the covariance matrix of the observed variables are appropriate. Anderson (1984) is a classic reference on multivariate analysis that treats estimation of covariance matrices and hypotheses testing on them.

More recently empirical work has considered the case where the dimension of the covariance matrix, n, relative to the sample size T, is large. This is because, in the era of big data, many datasets now used are large. For instance, as finance theory suggests that one should choose a well-diversified portfolio that perforce includes a large number of assets with non-zero weights, investors now consider many securities when forming a portfolio. The listed company Knight Capital Group claims to make markets in thousands of securities worldwide, and is constantly updating its inventories/portfolio weights to optimize its positions. If n/T is not negligible, we call this the *large dimensional* case.[1] The correct theoretical framework to study the large dimensional case is to use the joint asymptotics (i.e., both n and T diverge to infinity simultaneously albeit subject to some restriction on their relative growth rate), not the usual asymptotics (i.e., n fixed, T tends to infinity). Thus, standard statistical methods under the usual asymptotic framework, such as principal component analysis (PCA) and canonical-correlation analysis (CCA), do not directly generalise to the large dimension case; applications to, say, portfolio choice, face considerable difficulties (see Wang and Fan (2016)).

There are many new methodological approaches for the large dimensional case, for example Ledoit and Wolf (2003), Bickel and Levina (2008), Onatski (2009), Fan, Fan, and Lv (2008), Ledoit and Wolf (2012) Fan, Liao, and Mincheva (2013), and Ledoit and Wolf (2015). Yao, Zheng, and Bai (2015) gave an excellent account of the recent developments in the theory and practice of estimating large dimensional covariance matrices. Generally speaking, the approach is either to impose some sparsity on the covariance matrix, meaning that many elements of the covariance matrix are assumed to be zero or small, thereby reducing the number of parameters of a model for the covariance matrix to be estimated, or to use some device, such as shrinkage or a factor model, to reduce dimension.

We consider a parametric model for the covariance or correlation matrix - the Kronecker product model. For a real symmetric positive definite $n \times n$ correlation matrix Θ, a *Kronecker product model* is a family of $n \times n$ matrices $\{\Theta^*\}$, each of which has the following structure:

$$\Theta^* = \Theta_1^* \otimes \Theta_2^* \otimes \cdots \otimes \Theta_v^*, \tag{1.1}$$

where Θ_j^* is a $n_j \times n_j$ dimensional real symmetric, positive definite *sub-matrix* such that $n = n_1 \cdot n_2 \cdots n_v$. We require $n_j \in \mathbb{Z}$ and $n_j \geq 2$ for all j; n_j need not be distinct. The Kronecker product model, per se, is not new as it has been previously considered by Swain (1975) and Verhees and Wansbeek (1990) under the title of multimode analysis. Verhees and Wansbeek (1990) defined several estimation methods based on the least squares and maximum likelihood principles, and provided large sample variances under assumptions that the data are normal and fixed n. There is also a growing Bayesianist and frequentist literature on multiway array or tensor datasets, where a Kronecker product model is commonly employed. See for example Akdemir and Gupta (2011), Allen (2012), Browne, MacCallum, Kim, Andersen, and Glaser (2002), Cohen, Usevich, and Comon (2016), Constantinou, Kokoszka, and Reimherr (2015), Dobra (2014), Fosdick and Hoff (2014), Gerard and Hoff (2015), Hoff (2011), Hoff (2015), Hoff (2016), Krijnen (2004), Leiva and Roy (2014), Leng and Tang (2012), Li and Zhang (2016), Manceura and Dutilleul (2013), Ning and Liu (2013), Ohlson, Ahmada, and von Rosen (2013), Singull, Ahmad, and von Rosen (2012), Volfovsky and Hoff (2014), Volfovsky and Hoff (2015),

[1] We reserve the phrase "the high dimensional case" particularly for $n > T$.

and Yin and Li (2012). However, in both these (apparently separate) literatures, (i) n is fixed, (ii) the number v of sub-matrices of a Kronecker product is fixed and typically small, and (iii) each n_j is also fixed but perhaps of moderate size.

We consider the Kronecker product model in the large dimensional case where v is allowed to increase with n according to the factorization of n (each n_j is fixed though). In this model, the number of parameters of a Kronecker product model grows *logarithmically* with n. In particular, we will show that a Kronecker product model induces a type of sparsity on the covariance or correlation matrix: The logarithm of a Kronecker product model has many zero elements, so that sparsity is explicitly imposed on the logarithm of the covariance or correlation matrix - we call this *log sparsity*. Our work is among the first dealing with log sparsity; the other is Battey and Fan (2017), although there are a few differences. First, their log sparsity is an assumption from the onset, in a similar spirit as Bickel and Levina (2008), whereas our log sparsity is induced by a Kronecker product model. Second, they work with covariance matrices while we shall focus on correlation matrices. Although a Kronecker product model could also be applied to covariance matrices, log sparsity on a correlation matrix does not necessarily imply that its corresponding covariance matrix has log sparsity. In other words, if a Kronecker product model is correctly specified for a correlation matrix, its corresponding covariance matrix need not have a Kronecker product structure. Even if we look at covariance matrices only, for the purpose of comparison, a Kronecker product model imposes different sparsity restrictions - compared to those imposed by Battey and Fan (2017) - on the elements of the logarithm of the covariance matrix. Third and perhaps most important, we are looking at completely different estimators.

What kind of data give rise to a Kronecker product model? In other words, when is a Kronecker product model correctly specified? This question has been answered by Verhees and Wansbeek (1990) and Cudeck (1988): When covariance or correlation has some *multiplicative* structure. For example, suppose that $u_{j,k}$ are error terms in a panel regression model with $j = 1, \ldots, n_J$ and $k = 1, \ldots, n_K$. The interactive effects model of Bai (2009) is that $u_{j,k} = \gamma_j f_k$, which implies that $u = \gamma \otimes f$, where u is the $n_J n_K \times 1$ vector containing all the elements of $u_{j,k}$, $\gamma = (\gamma_1, \ldots, \gamma_{n_J})^\intercal$, and $f = (f_1, \ldots, f_{n_K})^\intercal$. If we assume that γ, f are random, γ is independent of f, and both vectors have mean zero, this implies that

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

We hence see that the covariance matrix is a Kronecker product of two sub-matrices.

We can think of our more general model (1.1) arising from multi-index data with v multiplicative factors. Multiway arrays are one such example as each observation has v different indices (see Hoff (2015)). Suppose that

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

or in vector form

$$u = (u_{1,1,\ldots,1}, \ldots, u_{n_1,n_2,\ldots,n_v})^\intercal = \varepsilon_1 \otimes \varepsilon_2 \otimes \cdots \otimes \varepsilon_v,$$

where the factor $\varepsilon_j = (\varepsilon_{j,1}, \ldots, \varepsilon_{j,n_j})^\intercal$ is a mean zero random vector of length n_j with covariance matrix Σ_j for $j = 1, \ldots, v$, and in addition the factors $\varepsilon_1, \ldots, \varepsilon_v$ are mutually independent. Then

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

We hence see that the covariance matrix is a Kronecker product of v sub-matrices. Indeed, such multiplicative effects may be a valid description of a covariance or correlation structure. In psychometrics, multi-trait multi-method (MTMM) context has this multiplicative structure (e.g., Campbell and O'Connell (1967) and Cudeck (1988)). In portfolio choice, one might consider, say, 250 equity portfolios constructed by intersections of 5 size groups (quintiles), 5 book-to-market equity ratio groups (quintiles) and 10 industry groups, in the spirit of Fama and French (1993). For example, one equity portfolio might consist of stocks which are in the smallest size quintile, largest book-to-market equity ratio quintile, and construction industry simultaneously. Then a Kronecker product model is applicable either directly to the covariance matrix of returns of these 250 equity portfolios or to the covariance matrix of the residuals after purging other common risk factors such as momentum.

Often a covariance or correlation matrix might not exactly correspond to a Kronecker product; that is, a Kronecker product model is misspecified. The previous literature on Kronecker product models did not touch this aspect, but we shall demonstrate in this article that a Kronecker product model is a very good approximating device to general covariance or correlation matrices, by trading off variance with bias. Indeed we show that there always exists a member in a Kronecker product model which is closest to the covariance or correlation matrix in some sense to be made precise shortly.

The Kronecker product model has a number of intrinsic advantages for applications. The eigenvalues of a Kronecker product are products of the eigenvalues of its sub-matrices. Its inverse, determinant, and other key quantities are easily obtained from the corresponding quantities of its sub-matrices, which facilitates computation and analysis. In addition, a Kronecker product model could be used as one component of a super model consisting of several models.

For instance, the idea of the decomposition in (1.1) could be applied to components of *dynamic* models such as multivariate GARCH, an area in which Luc Bauwens has contributed significantly over the recent years, see also his highly cited review paper Bauwens, Laurent, and Rombouts (2006). For example, the dynamic conditional correlation (DCC) model of Engle (2002), or the BEKK model of Engle and Kroner (1995) both have intercept matrices that are required to be positive definite and suffer from the curse of dimensionality, for which model (1.1) would be helpful. Also, parameter matrices associated with the dynamic terms in the model could be equipped with a Kronecker product, similar to a suggestion by Hoff (2015) for vector autoregressions.

In this article, we shall focus on correlation matrices rather than covariance matrices. This is partly because the asymptotic theories of a Kronecker product model for correlation matrices nest those for covariance matrices, and partly because this will allow us to adopt a more flexible approach to approximating a general covariance matrix, since we can estimate the variances consistently by other well-understood methods. In practice, fitting a correlation matrix with a Kronecker product model tends to perform better than doing so for its corresponding covariance matrix.

We show that the logarithm of a Kronecker product model is linear in its unknown parameters, and use this as a basis to propose a minimum distance (MD) estimator. We establish a rather "crude" rate of convergence for the MD estimator under joint asymptotics. So far endeavours to obtain a better rate have proven to be unfruitful and this question remains open. There is a large literature on the optimal rates of convergence for estimation of high-dimensional covariance and inverse (i.e., *precision*) matrices (see Cai, Zhang, and Zhou (2010) and Cai and Zhou (2012)). Cai, Ren, and Zhou (2014) gave a nice review on those recent results. However their optimal rates are not applicable to our setting because here sparsity is not imposed on the covariance or correlation matrix, but on its logarithm.

Although the MD estimator allows direct theoretical analysis, this method is likely to be computationally intensive and in practice we recommend to use quasi-maximum likelihood estimation. Hence we also discuss a quasi-maximum likelihood estimator (QMLE) and a one-step estimator, which is an approximate QMLE. Under the joint asymptotics, we provide feasible central limit theorems (CLT) for the MD and one-step estimators, the latter of which is shown to achieve the parametric efficiency bound (Cramer-Rao lower bound) in the fixed n case. When choosing the weighting matrix optimally, we also show that the optimally-weighted MD and one-step estimators have the same asymptotic distribution. These CLTs are of independent interest and contribute to the literature on the large dimensional CLTs (see Huber (1973), Yohai and Maronna (1979), Portnoy (1985), Mammen (1989), Welsh (1989), Bai and Wu (1994), Saikkonen and Lutkepohl (1996) and He and Shao (2000)). Last, we give a specification test which allows us to test whether a Kronecker product model is correctly specified.

We provide some evidence that the Kronecker product model works well numerically. We also apply the Kronecker product model to portfolio selection and compare the model with the sample covariance matrix and a linear shrinkage estimator (Ledoit and Wolf (2004)).

The rest of the paper is structured as follows. In Section 2 we lay out the Kronecker product model in detail. Section 3 introduces the MD estimator, gives its asymptotic properties, and includes a specification test, while Section 4 discusses the QMLE and one-step estimator, and provides the asymptotic properties of the one-step estimator. Section 5 examines the issue of model selection. Section 6 provides numerical evidence for the performance of the Kronecker product model in a simulation study and an empirical application. Section 7 concludes. Primary proofs are to be found in Appendix; the remaining proofs are put in Supplementary Material (SM in what follows).

2 The Kronecker Product Model

2.1 Notation

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

For $x \in \mathbb{R}^n$, let $\|x\|_2 := \sqrt{\sum_{i=1}^n x_i^2}$ and $\|x\|_\infty := \max_{1 \leq i \leq n} |x_i|$ denote the Euclidean norm and the element-wise maximum norm, respectively. $\operatorname{diag}(x)$ gives an $n \times n$ diagonal matrix with the diagonal being the elements of x. Let $\operatorname{maxeval}(\cdot)$ and $\operatorname{mineval}(\cdot)$ denote the maximum and minimum eigenvalues of some real symmetric matrix, respectively. For any real $m \times n$ matrix $A = (a_{i,j})_{1 \leq i \leq m, 1 \leq j \leq n}$, let $\|A\|_F := [\operatorname{tr}(A^\intercal A)]^{1/2} \equiv [\operatorname{tr}(AA^\intercal)]^{1/2} \equiv \|\operatorname{vec} A\|_2$, $\|A\|_{\ell_2} := \max_{\|x\|_2 = 1} \|Ax\|_2 \equiv \sqrt{\operatorname{maxeval}(A^\intercal A)}$, and $\|A\|_{\ell_\infty} := \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{i,j}|$ denote the Frobenius norm, spectral norm (ℓ_2 operator norm) and maximum row sum matrix norm (ℓ_∞ operator norm) of A, respectively. Note that $\| \cdot \|_\infty$ can also be applied to matrix A, i.e., $\|A\|_\infty = \max_{1 \leq i \leq m, 1 \leq j \leq n} |a_{i,j}|$; however $\| \cdot \|_\infty$ is not a matrix norm so it does not have the submultiplicative property of a matrix norm.

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

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

For matrix calculus, what we adopt is called the *numerator layout* or *Jacobian formulation*; that is, the derivative of a scalar with respect to a column vector is a row vector.

2.2 The Model and Identification

In this section we provide more details on the specific model we consider for the large correlation matrix. We first give a definition of the *principal matrix logarithm* for real symmetric, positive definite matrices. More generally, the principal matrix logarithm could be defined for any square complex matrix having no eigenvalues lying on the closed real axis $(-\infty, 0]$, but we do not need this level of generality in this article. We shall drop the qualifier "principal" for simplicity.

Definition 2.1 (Matrix logarithm)**.** *Suppose that a real, positive definite matrix A $(n \times n)$ has the orthogonal diagonalization $A = U^\intercal \mathit{diag}(\lambda_1, \ldots, \lambda_n)U$. Then its matrix logarithm, denoted* $\log A$*, is defined as*

$$\log A := U^\intercal \mathit{diag}(\log \lambda_1, \ldots, \log \lambda_n)U.$$

Consider an n-dimensional vector time series $\{x_t\}_{t=1}^T$ that is i.i.d. with $\mu := \mathbb{E}x_t$ and covariance matrix $\Sigma := \mathbb{E}[(x_t - \mu)(x_t - \mu)^\intercal]$. Let D be the diagonal matrix containing diagonal entries of Σ. Its correlation matrix Θ is

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

A Kronecker product model for Θ is given by (1.1).[2][3] That Θ is a correlation matrix implies that the diagonal entries of Θ_j^* must be the same, although this diagonal entry could differ as j varies. Without loss of generality, we shall impose a normalisation constraint that all these v diagonal entries of $\{\Theta_j^*\}_{j=1}^v$ are 1.

A Kronecker product model substantially reduces the number of parameters to estimate for a correlation matrix. In an unrestricted correlation matrix, there are $n(n-1)/2$ parameters, while a Kronecker product model has only $\sum_{j=1}^v n_j(n_j-1)/2$ parameters. As an extreme illustration, when $n = 256$, the unrestricted correlation matrix has 32,640 parameters while a Kronecker product model of factorization $256 = 2^8$ has only 8 parameters! Although Θ^* is not sparse, $\log \Theta^*$ is sparse. This is due to a property of Kronecker products (see Proposition 8.1 in SM 8.1 for derivation):

$$\log \Theta^* = \log \Theta_1^* \otimes I_{n_2} \otimes \cdots \otimes I_{n_v} + I_{n_1} \otimes \log \Theta_2^* \otimes I_{n_3} \otimes \cdots \otimes I_{n_v} + \cdots + I_{n_1} \otimes I_{n_2} \otimes \cdots \otimes \log \Theta_v^*,$$

whence we see that $\log \Theta^*$ has many zero elements, generated by identity sub-matrices.

After the normalisation of diagonal entries of Θ_j to be 1 for all j, parameters in Θ_j^* still warrants some discussion. As an illustration, suppose

$$\Theta_1^* = \begin{pmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 1 & 0.2 \\ 0.5 & 0.2 & 1 \end{pmatrix},$$

and then one can compute that

$$\log \Theta_1^* = \begin{pmatrix} -0.75 & 1.18 & 0.64 \\ 1.18 & -0.55 & -0.07 \\ 0.64 & -0.07 & -0.17 \end{pmatrix}.$$

[2]Note that if n is not composite, one can add a vector of pseudo variables to the system until the final dimension is composite. It is recommended to add a vector of independent variables $u_t \sim N(0, I_k)$ such that $z_t := (x_t^\intercal, u_t^\intercal)^\intercal$ is an $n \times 1$ random vector with $n \times n$ correlation matrix

$$\Theta = \begin{bmatrix} \Theta_x & 0 \\ 0 & I_k \end{bmatrix}.$$

[3]The Kronecker product model is invariant under the Lie group of transformations $\mathcal{G}$ generated by $A_1 \otimes A_2 \otimes \cdots \otimes A_v$, where A_j are $n_j \times n_j$ nonsingular matrices (see Browne and Shapiro (1991)). This structure can be used to characterise the tangent space $\mathcal{T}$ of $\mathcal{G}$ and to define a relevant equivariance concept for restricting the class of estimators for optimality considerations.

Thus there are $n_j(n_j+1)/2$ parameters in $\log\Theta_j^*$ for $j=1,\ldots,v$; we call these *log parameters.* On the other hand, there are only $n_j(n_j-1)/2$ parameters in Θ_j^* for $j=1,\ldots,v$; we call these *original parameters.* These $n_j(n_j-1)/2$ original parameters completely pin down those $n_j(n_j+1)/2$ log parameters. In other words, there exists a function $f:\mathbb{R}^{n_j(n_j-1)/2}\to\mathbb{R}^{n_j(n_j+1)/2}$ which maps original parameters to log parameters. However, when $n_j>4$, f does *not* have a closed form because when $n_j>4$ the continuous functions which map elements of a matrix to its eigenvalues have no closed form. When $n_j=2$, we can solve f by hand (see Example 2.1). When $n_j=3$, one could use, say, Matlab, to perform symbolic computation, but the expressions will be extremely complicated.

Example 2.1. *Suppose*

$$\Theta_1^*=\begin{pmatrix}1 & \rho_1^*\\ \rho_1^* & 1\end{pmatrix}.$$

The eigenvalues of Θ_1^* are $1+\rho_1^*$ and $1-\rho_1^*$, respectively. The corresponding eigenvectors are $(1,1)^\intercal/\sqrt{2}$ and $(1,-1)^\intercal/\sqrt{2}$, respectively. Therefore

$$\begin{aligned}\log\Theta_1^* &= \begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}\log(1+\rho_1^*) & 0\\ 0 & \log(1-\rho_1^*)\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\frac{1}{2}\\ &= \begin{pmatrix}\frac{1}{2}\log(1-[\rho_1^*]^2) & \frac{1}{2}\log\left(\frac{1+\rho_1^*}{1-\rho_1^*}\right)\\ \frac{1}{2}\log\left(\frac{1+\rho_1^*}{1-\rho_1^*}\right) & \frac{1}{2}\log(1-[\rho_1^*]^2)\end{pmatrix}.\end{aligned}$$

Thus

$$f(\rho)=\left(\frac{1}{2}\log(1-\rho^2),\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right),\frac{1}{2}\log(1-\rho^2)\right)^\intercal.$$

To separately identify log parameters in $\Theta_1^*,\ldots,\Theta_v^*$ from the onset, we need to set the first diagonal entry of $\log\Theta_j^*$ to be 0 for $j=1,\ldots,v-1$. In total there are

$$s:=\sum_{j=1}^{v}\frac{n_j(n_j+1)}{2}-(v-1)=O(\log n)$$

(identifiable) log parameters in $\Theta_1^*,\ldots,\Theta_v^*$; let $\theta^*\in\mathbb{R}^s$ denote these. On the other hand, to separately identify original parameters in $\Theta_1^*,\ldots,\Theta_v^*$ from the onset, no additional identification restriction is needed.

To estimate a Kronecker product model, there are two approaches. First, one can estimate original parameters directly using Gaussian quasi-maximum likelihood estimation (see Section 4.1). Second, one can estimate log parameters θ^* using the principle of minimum distance or Gaussian quasi-maximum likelihood estimation (see Section 3 and Section 4.1); then recover the estimates of original parameters via the matrix exponential. When one adopts the second approach, the diagonal of the estimated Θ_j^* cannot have exact ones. In this case, one can replace these diagonal estimates with 1. To study the theoretical properties of a Kronecker product model, we feel that the second approach is more appealing as log parameters are additive in nature while original parameters are multiplicative in nature; additive objects are easier to analyse theoretically than multiplicative objects. To use Kronecker product models in practice, the first approach is far easier to implement.

3 The Minimum Distance Estimator

In this section, we study how to estimate log parameters θ^* of the Kronecker product model (1.1).

3.1 Estimation

We first give the main useful model property that delivers a simple estimation strategy. Proposition A.1 in Appendix A.1 proves that there exists an $n(n+1)/2 \times s$ full column rank, deterministic matrix E such that

$$\mathrm{vech}(\log \Theta^*) = E\theta^*.$$

(The R code for computing this matrix E is available.) Given a factorization $n = n_1 \cdot n_2 \cdots n_v$, if there exists an $\Theta^\dagger \in \{\Theta^*\}$ such that $\Theta = \Theta^\dagger$, we say that the Kronecker product model $\{\Theta^*\}$ is *correctly specified* (i.e., $\mathrm{vech}(\log \Theta) = E\theta$). Otherwise the Kronecker product model $\{\Theta^*\}$ is *misspecified*.

Define the sample covariance matrix and sample correlation matrix

$$\hat{\Sigma}_T := \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})(x_t - \bar{x})^\intercal, \qquad \hat{\Theta}_T := \hat{D}_T^{-1/2}\hat{\Sigma}_T\hat{D}_T^{-1/2},$$

where $\bar{x} := (1/T)\sum_{t=1}^{T} x_t$ and $\hat{D}_T$ is a diagonal matrix whose diagonal elements are diagonal elements of $\hat{\Sigma}_T$.

We show in Appendix A.2 that in the Kronecker product model $\{\Theta^*\}$ there exists a unique member, denoted Θ^0, which is closest to the correlation matrix Θ in the following sense:

$$\theta^0 = \theta^0(W) := \arg\min_{\theta^* \in \mathbb{R}^s} [\mathrm{vech}(\log \Theta) - E\theta^*]^\intercal W [\mathrm{vech}(\log \Theta) - E\theta^*], \tag{3.1}$$

where W is a $n(n+1)/2 \times n(n+1)/2$ positive definite weighting matrix which is free to choose. Clearly, θ^0 has the closed form solution $\theta^0 = (E^\intercal W E)^{-1} E^\intercal W \,\mathrm{vech}(\log \Theta)$. The population objective function (3.1) allows us to define a minimum distance (MD) estimator:

$$\hat{\theta}_T = \hat{\theta}_T(W) := \arg\min_{b \in \mathbb{R}^s} [\mathrm{vech}(\log \hat{\Theta}_T) - Eb]^\intercal W [\mathrm{vech}(\log \hat{\Theta}_T) - Eb], \tag{3.2}$$

whence we can solve

$$\hat{\theta}_T = (E^\intercal W E)^{-1} E^\intercal W \,\mathrm{vech}(\log \hat{\Theta}_T). \tag{3.3}$$

Thus we have

$$\hat{\theta}_T - \theta^0 = (E^\intercal W E)^{-1} E^\intercal W \,\mathrm{vech}(\log \hat{\Theta}_T - \log \Theta).$$

Note that θ^0 is the quantity which one should expect $\hat{\theta}_T$ to converge to in some probabilistic sense regardless of whether the Kronecker product model $\{\Theta^*\}$ is correctly specified or not. When $\{\Theta^*\}$ is correctly specified, i.e., there exists a θ such that $\mathrm{vech}(\log \Theta) = E\theta$, we have $\theta^0 = (E^\intercal W E)^{-1} E^\intercal W \,\mathrm{vech}(\log \Theta) = (E^\intercal W E)^{-1} E^\intercal W E\theta = \theta$. In this case, $\hat{\theta}_T$ is indeed estimating the elements of the correlation matrix Θ.

3.2 Rate of Convergence

We shall now introduce some assumptions for our theoretical analysis.

Assumption 3.1.

(i) $\{x_t\}_{t=1}^T$ *are subgaussian random vectors. That is, for all* t, *for every* $a \in \mathbb{R}^n$ *with* $\|a\|_2 = 1$, *and every* $\epsilon > 0$

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

for positive absolute constants K *and* C.

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

Assumption 3.2.

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

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

$$\frac{n^2\kappa^3(W)\varpi^2\log^2 n}{T}\left(T^{2/\gamma}\log^2 n \vee n^2\kappa^3(W)\varpi^2\log^2 n\cdot\log^5 n^4\right) = o(1), \quad \textit{for some } \gamma > 2,$$

where $\kappa(W)$ *is the condition number of* W *for matrix inversion with respect to the spectral norm, i.e.,* $\kappa(W) := \|W^{-1}\|_{\ell_2}\|W\|_{\ell_2}$ *and* ϖ *is defined in Assumption 3.3(ii).*

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

(a) for some $\gamma > 2$,

$$\frac{n^2\varpi^2\log^3 n}{T}\left(n\varpi\kappa(W) \vee T^{\frac{2}{\gamma}}\log n\right) = o(1),$$

(b)

$$\frac{\varpi^2\log n}{n} = o(1).$$

Assumption 3.3.

(i) The minimum eigenvalue of Σ *is bounded away from zero by an absolute constant.*

(ii)

$$\textit{mineval}\left(\frac{1}{n}E^{\intercal}E\right) \geq \frac{1}{\varpi} > 0.$$

(At most $\varpi = o(n)$*.)*

Assumption 3.1(i) is standard in high-dimensional theoretical work. In essence it assumes that a random vector has exponential tail probabilities, which allows us to invoke some concentration inequality such as the Bernstein's inequality in Appendix A.5. Note that Assumption 3.1(i) could be replaced by a finite moment assumption and this will only result a rate slightly worse than $\sqrt{n/T}$ in Proposition 3.1(i) (c.f. Vershynin (2012)). Assumption 3.1(ii), which will only be used in Section 4 for quasi-maximum likelihood or one-step estimation, implies Assumption 3.1(i); we stress that Assumption 3.1(ii) is not needed for the minimum distance estimation (Theorem 3.1 or 3.2).

Assumption 3.2(i) is for the derivation of the rate of convergence of spectral norm of $\hat{\Theta}_T - \Theta$. To establish the *same* rate of convergence of spectral norm of $\hat{\Sigma}_T - \Sigma$, one only needs $n/T \to c \in [0, 1]$. However for correlation matrices, we need $n/T \to 0$. This is because a correlation matrix involves inverses of standard deviations. Assumptions 3.2(ii) and (iii) are *sufficient* conditions for the asymptotic normality of the minimum distance estimators (Theorems 3.1 and 3.2) and of the one-step estimator (Theorem 4.1), respectively. If Assumption 3.1(i) holds, γ in Assumption 3.2(ii) and (iii) could be made arbitrarily large, which makes Assumptions 3.2(ii) and (iii) much less restrictive. Assumption 3.2(ii) necessarily requires $n^4/T \to 0$. At first glance, it looks restrictive, but we would like to remark that this is only a sufficient condition. More importantly, we are trying to establish a CLT for elements of the second moment of x_t in the large dimensional case. If one is familiar with the literature on the large-dimensional CLT (e.g., Lewis and Reinsel (1985), Saikkonen and Lutkepohl (1996), Chang, Chen, and Chen (2015)), they usually require $n^3/T \to 0$ for establishment of CLTs for elements of the first moment of the data, so our assumption is nothing bold. The same reasoning applies to Assumption 3.2(iii).

Assumption 3.3(i) is also standard. This ensures that Θ is positive definite with the minimum eigenvalue bounded away from 0 by an absolute positive constant (see Proposition A.5(i) in

Appendix A.4) and its logarithm is well-defined. Assumption 3.3(ii) postulates a lower bound for the minimum eigenvalue of $E^\intercal E/n$; that is

$$\frac{1}{\sqrt{\text{mineval}\left(\frac{1}{n}E^\intercal E\right)}} = O(\sqrt{\varpi}).$$

We divide $E^\intercal E$ by n because all the non-zero elements of $E^\intercal E$ are a multiple of n (see Proposition A.2 in Appendix A.1). In words, Assumption 3.3(ii) says that the minimum eigenvalue of $E^\intercal E/n$ slowly drifts to zero.

Proposition 3.1.

(i) Suppose Assumptions 3.1(i), 3.2(i) and 3.3(i) hold. Then

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

(ii) Suppose that $\|\hat{\Theta}_T - \Theta\|_{\ell_2} < a$ with probability approaching 1 for some absolute constant $a > 1$, then we have

$$\|\log \hat{\Theta}_T - \log \Theta\|_{\ell_2} = O_p(\|\hat{\Theta}_T - \Theta\|_{\ell_2}).$$

(iii) Suppose Assumptions 3.1(i), 3.2(i) and 3.3 hold. Then

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

where $\|\cdot\|_2$ is the Euclidean norm, $\kappa(W)$ is the condition number of W for matrix inversion with respect to the spectral norm, i.e., $\kappa(W) := \|W^{-1}\|_{\ell_2}\|W\|_{\ell_2}$, and ϖ is defined in Assumption 3.3(ii).

Proof. See Appendix A.3. □

Proposition 3.1(i) provides the rate of convergence of the spectral norm of $\hat{\Theta}_T - \Theta$, which is a stepping stone for the rest of theoretical results. Strictly speaking, the rate should be $n/T \vee \sqrt{n/T}$, which collapses to $\sqrt{n/T}$ under Assumption 3.2(i). This rate is the same as that of $\|\hat{\Sigma}_T - \Sigma\|_{\ell_2}$. Proposition 3.1(ii) is also of independent interest as it relates $\|\log \hat{\Theta}_T - \log \Theta\|_{\ell_2}$ to $\|\hat{\Theta}_T - \Theta\|_{\ell_2}$.

Proposition 3.1(iii) gives the rate of convergence of the minimum distance estimator $\hat{\theta}_T$. θ^0 are log parameters of the member in the Kronecker product model, which is closest to Θ in the sense discussed earlier. For sample correlation matrix $\hat{\Theta}_T$, the rate of convergence of $\|\operatorname{vec}(\hat{\Theta}_T - \Theta)\|_2$ is $\sqrt{n^2/T}$ (square root of summing up $O(n^2)$ terms each of which has a rate $1/T$). Thus the minimum distance estimator $\hat{\theta}_T$ of the Kronecker product model converges faster provided $\varpi\kappa(W)$ is not too large, in line with principle of dimension reduction. However, given that the dimension of θ^0 is $s = O(\log n)$, one would conjecture that the optimal rate of convergence should be $\sqrt{\log n/T}$. In this sense, Proposition 3.1(iii) does not demonstrate the full advantages of a Kronecker product model. Because of the severe non-linearity introduced by matrix logarithm as it is defined through spectrum, it is beyond the scope of this article to prove a faster rate of convergence of $\|\hat{\theta}_T - \theta^0\|_2$.

3.3 Asymptotic Normality

To derive the asymptotic normality of the minimum distance estimator, we consider two cases

(i) μ is unknown but D is known;

(ii) both μ and D are unknown.

We will derive the asymptotic normality of the minimum distance estimator for both cases.

Define the following $n^2 \times n^2$ dimensional matrix H:

$$H := \int_0^1 [t(\Theta - I) + I]^{-1} \otimes [t(\Theta - I) + I]^{-1} dt. \tag{3.4}$$

Define also the $n \times n$ and $n^2 \times n^2$ matrices:

$$\tilde{\Sigma}_T := \frac{1}{T}\sum_{t=1}^{T}(x_t - \mu)(x_t - \mu)^\intercal \qquad V := \operatorname{var}(\sqrt{T}\operatorname{vec}(\tilde{\Sigma}_T - \Sigma)).$$

Since $x \mapsto (\lceil \frac{x}{n} \rceil, x - \lfloor \frac{x}{n} \rfloor n)$ is a bijection from $\{1, \ldots, n^2\}$ to $\{1, \ldots, n\} \times \{1, \ldots, n\}$, it is easy to show that the (x, y)th entry of V is

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

where $\mu_i = \mathbb{E}x_{t,i}$ (similarly for μ_j, μ_k, μ_ℓ), $x, y \in \{1, \ldots, n^2\}$ and $i, j, k, \ell \in \{1, \ldots, n\}$. In the special case of normality, $V = 2D_n D_n^+(\Sigma \otimes \Sigma)$ (Magnus and Neudecker (1986) Lemma 9).

Assumption 3.4. *V is positive definite for all n, with its minimum eigenvalue bounded away from zero by an absolute constant and maximum eigenvalue bounded from above by an absolute constant.*

Assumption 3.4 is also a standard regularity condition. It is automatically satisfied under normality given Assumptions 3.2(i) and 3.3(i) (via Proposition A.3(vi) in Appendix A.3). Assumption 3.4 could be relaxed to the case where the minimum (maximum) eigenvalue of V is slowly drifting towards zero (infinity) at certain rate. The proofs for Theorem 3.1 and Theorem 3.2 remain unchanged, but this rate will need to be incorporated in Assumption 3.2(ii).

3.3.1 When μ Is Unknown But D Is Known

In this case, $\hat{\Theta}_T$ simplifies into $\hat{\Theta}_{T,D} := D^{-1/2}\hat{\Sigma}_T D^{-1/2}$. Similarly, the minimum distance estimator $\hat{\theta}_T$ simplifies into $\hat{\theta}_{T,D} := (E^\intercal W E)^{-1} E^\intercal W \operatorname{vech}(\log \hat{\Theta}_{T,D})$. Let $\hat{H}_{T,D}$ denote the $n^2 \times n^2$ matrix

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

Define V's sample analogue $\hat{V}_T$ whose (x, y)th entry is

$$\begin{aligned}\hat{V}_{T,x,y} \equiv \hat{V}_{T,i,j,k,\ell} := &\frac{1}{T}\sum_{t=1}^{T}(x_{t,i} - \bar{x}_i)(x_{t,j} - \bar{x}_j)(x_{t,k} - \bar{x}_k)(x_{t,\ell} - \bar{x}_\ell) \\ &- \left(\frac{1}{T}\sum_{t=1}^{T}(x_{t,i} - \bar{x}_i)(x_{t,j} - \bar{x}_j)\right)\left(\frac{1}{T}\sum_{t=1}^{T}(x_{t,k} - \bar{x}_k)(x_{t,\ell} - \bar{x}_\ell)\right),\end{aligned}$$

where $\bar{x}_i := \frac{1}{T}\sum_{t=1}^{T} x_{t,i}$ (similarly for $\bar{x}_j, \bar{x}_k$ and $\bar{x}_\ell$), $x, y \in \{1, \ldots, n^2\}$ and $i, j, k, \ell \in \{1, \ldots, n\}$.

For any $c \in \mathbb{R}^s$ define the scalar

$$G_D := c^\intercal J_D c := c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H(D^{-1/2} \otimes D^{-1/2}) V (D^{-1/2} \otimes D^{-1/2}) H D_n^{+\intercal} W E (E^\intercal W E)^{-1} c.$$

In the special case of normality, G_D could be simplified into (see Example 8.1 in SM 8.6 for details): $2c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H(\Theta \otimes \Theta) H D_n^{+\intercal} W E (E^\intercal W E)^{-1} c$. We also define the estimate $\hat{G}_{T,D}$:

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

Theorem 3.1. *Let Assumptions 3.1(i), 3.2(ii), 3.3 and 3.4 be satisfied. Then*

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

for any $s \times 1$ non-zero vector c with $\|c\|_2 = 1$.

Proof. See Appendix A.4. □

Theorem 3.1 is a version of the large-dimensional CLT, whose proof is mathematically non-trivial. Because the dimension of θ^0 is growing with the sample size, for a CLT to make sense, we need to transform $\hat{\theta}_{T,D} - \theta^0$ to a univariate quantity by pre-multiplying $c^\intercal$. The magnitudes of the elements of c are not important, so we normalize it to have unit Euclidean norm. What is important is whether the elements of c are zero or not. The components of $\hat{\theta}_{T,D} - \theta^0$ whose positions correspond to the non-zero elements of c are effectively entering the CLT.

We contribute to the literature on the large-dimensional CLT (see Huber (1973), Yohai and Maronna (1979), Portnoy (1985), Mammen (1989), Welsh (1989), Bai and Wu (1994), Saikkonen and Lutkepohl (1996) and He and Shao (2000)). In this strand of literature, a distinct feature is that the dimension of parameter, say, θ^0, is growing with the sample size, and at the same time we do not impose sparsity on θ^0. As a result, the rate of growth of dimension of parameter has to be restricted by an assumption like Assumption 3.2(ii); in particular, the dimension of parameter cannot exceed the sample size. This approach is different from the recent literature on high-dimensional statistics such as Lasso, where one imposes sparsity on parameter to allow its dimension to exceed the sample size.

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

Corollary 3.1. *Let Assumptions 3.1(i), 3.2(ii), 3.3 and 3.4 be satisfied. Given a full-column-rank $s \times k$ matrix A where k is finite with $\|A\|_{\ell_2} = O_p(\sqrt{\log n \cdot n\kappa(W)})$, we have*

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

Proof. See SM 8.6. □

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

If one chooses the weighting matrix W optimally, albeit infeasibly,

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

the scalar G_D reduces to

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

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

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

by Lemma 14 of Magnus and Neudecker (1986). We shall compare the preceding display with the variance of the asymptotic distribution of the one-step estimator in Section 4.

3.3.2 When Both μ and D Are Unknown

The case where both μ and D are unknown is considerably more difficult. If one simply recycles the proof for the case where only μ is unknown and replaces D with its plug-in estimator $\hat{D}_T$, it will not work.

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

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

Define the $n^2 \times n^2$ matrix P:

$$P := I_{n^2} - D_nD_n^+(I_n \otimes \Theta)M_d, \qquad M_d := \sum_{i=1}^{n}(F_{ii} \otimes F_{ii}),$$

where F_{ii} is an $n \times n$ matrix with one in its (i,i)th position and zeros elsewhere. M_d is a $n^2 \times n^2$ diagonal matrix with diagonal elements equal to 0 or 1; the positions of 1 in the diagonal of M_d correspond to the positions of diagonal entries of an arbitrary matrix A in $\operatorname{vec} A$. Note that matrix P is an idempotent matrix of rank $n^2 - n$ and first appeared in (4.6) of Neudecker and Wesselman (1990). In particular, given any correlation matrix Θ, P has $n^2 - n$ rows of zeros. Neudecker and Wesselman (1990) proved that

$$\frac{\partial \operatorname{vec} \Theta}{\partial \operatorname{vec} \Sigma} = P(D^{-1/2} \otimes D^{-1/2});$$

the derivative is a function of Σ.

For any $c \in \mathbb{R}^s$ define the scalar G and its estimate $\hat{G}_T$:

$$G := c^\intercal Jc := c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+HP(D^{-1/2} \otimes D^{-1/2})V(D^{-1/2} \otimes D^{-1/2})P^\intercal HD_n^{+\intercal}WE(E^\intercal WE)^{-1}c.$$

$$\hat{G}_T := c^\intercal \hat{J}_Tc := c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T\hat{P}_T(\hat{D}_T^{-1/2} \otimes \hat{D}_T^{-1/2})\hat{V}_T(\hat{D}_T^{-1/2} \otimes \hat{D}_T^{-1/2})\hat{P}_T^\intercal \hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c,$$

where $\hat{P}_T := I_{n^2} - D_nD_n^+(I_n \otimes \hat{\Theta}_T)M_d$.

Assumption 3.5.

(i) For every positive constant C

$$\sup_{\Sigma^*: \|\Sigma^* - \Sigma\|_F \le C\sqrt{\frac{n^2}{T}}} \left\| \frac{\partial \operatorname{vec} \Theta}{\partial \operatorname{vec} \Sigma}\bigg|_{\Sigma = \Sigma^*} - P(D^{-1/2} \otimes D^{-1/2}) \right\|_{\ell_2} = O\left(\sqrt{\frac{n}{T}}\right),$$

where $\cdot|_{\Sigma=\Sigma^*}$ means "evaluate the argument Σ at Σ^*".

(ii) The $s \times s$ matrix

$$E^\intercal WD_n^+HP(D^{-1/2} \otimes D^{-1/2})V(D^{-1/2} \otimes D^{-1/2})P^\intercal HD_n^{+\intercal}WE$$

has full rank s (i.e, being positive definite). Moreover,

$$\text{mineval}\Big(E^\intercal WD_n^+HP(D^{-1/2} \otimes D^{-1/2})V(D^{-1/2} \otimes D^{-1/2})P^\intercal HD_n^{+\intercal}WE\Big) \ge \frac{n}{\varpi}\text{mineval}^2(W).$$

Assumption 3.5(i) characterises some sort of uniform rate of convergence in terms of spectral norm of the Jacobian matrix $\frac{\partial \operatorname{vec} \Theta}{\partial \operatorname{vec} \Sigma}$. This type of assumption is usually needed when one wants to stop Taylor expansion, say, of $\operatorname{vec} \hat{\Theta}_T$, at first order. If one goes into the second-order expansion (a tedious route), Assumption 3.5(i) can be completely dropped at some expense of further restricting the relative growth rate between n and T. The radius of the shrinking neighbourhood $\sqrt{n^2/T}$ is determined by the rate of convergence in terms of the Frobenius norm of the sample covariance matrix $\hat{\Sigma}_T$. The rate on the right side of Assumption 3.5(i) is chosen to be $\sqrt{n/T}$ because it is the rate of convergence of

$$\left\| \left. \frac{\partial \operatorname{vec} \Theta}{\partial \operatorname{vec} \Sigma} \right|_{\Sigma=\hat{\Sigma}_T} - P(D^{-1/2} \otimes D^{-1/2}) \right\|_{\ell_2}$$

which could be easily deduced from the proof of Theorem 3.2. This rate $\sqrt{n/T}$ could even be relaxed to $\sqrt{n^2/T}$ as the part of the proof of Theorem 3.2 which requires Assumption 3.5(i) is not the "binding" part of the whole proof.

We now examine Assumption 3.5(ii). The $s \times s$ matrix

$$E^\intercal W D_n^+ H P (D^{-1/2} \otimes D^{-1/2}) V (D^{-1/2} \otimes D^{-1/2}) P^\intercal H D_n^{+\intercal} W E$$

is symmetric and positive semidefinite. By Observation 7.1.8 of Horn and Johnson (2013), its rank is equal to $\operatorname{rank}(E^\intercal W D_n^+ H P)$, if $(D^{-1/2} \otimes D^{-1/2}) V (D^{-1/2} \otimes D^{-1/2})$ is positive definite. In other words, Assumption 3.5(ii) is assuming $\operatorname{rank}(E^\intercal W D_n^+ H P) = s$, provided $(D^{-1/2} \otimes D^{-1/2}) V (D^{-1/2} \otimes D^{-1/2})$ is positive definite. Even though P has only rank $n^2 - n$, in general the rank condition does hold except in a special case. The special case is $\Theta = I_n \& W = I_{n(n+1)/2}$. In this special case

$$\operatorname{rank}(E^\intercal W D_n^+ H P) = \operatorname{rank}(E^\intercal D_n^+ P) = \sum_{j=1}^{v} \frac{n_j(n_j - 1)}{2} < s.$$

The second part of Assumption 3.5(ii) postulates a lower bound for its minimum eigenvalue. The rate $\operatorname{mineval}^2(W) n / \varpi$ is specified as such because of Assumption 3.3(ii). Other magnitudes of the rate are also possible as long as the proof of Theorem 3.2 goes through.

Theorem 3.2. *Let Assumptions 3.1(i), 3.2(ii), 3.3, 3.4 and 3.5 be satisfied. Then*

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

for any $s \times 1$ non-zero vector c with $\|c\|_2 = 1$.

Proof. See SM 8.3. □

Again Theorem 3.2 is a version of the large-dimensional CLT, whose proof is mathematically non-trivial. It has the same structure as that of Theorem 3.1. However $\hat{G}_T$ differs from $\hat{G}_{T,D}$ reflecting the difference between G and G_D. That is, the asymptotic distribution of the minimum distance estimator depends on whether D is known or not.

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

Corollary 3.2. *Let Assumptions 3.1(i), 3.2(ii), 3.3, 3.4 and 3.5 be satisfied. Given a full-column-rank $s \times k$ matrix A where k is finite with $\|A\|_{\ell_2} = O_p(\sqrt{\log^2 n \cdot n \kappa^2(W) \varpi})$, we have*

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

Proof. Essentially the same as that of Corollary 3.1. $\square$

The condition $\|A\|_{\ell_2} = O_p(\sqrt{\log^2 n \cdot n\kappa^2(W)\varpi})$ is trivial because the dimension of A is only of order $O(\log n) \times O(1)$. Moreover we can always rescale A when carrying out hypothesis testing.

3.4 A Specification Test

We give a specification test (also known as an over-identification test) based on the minimum distance objective function in (3.2). Suppose we want to test whether the Kronecker product model $\{\Theta^*\}$ is correctly specified given the factorization $n = n_1 \cdot n_2 \cdots n_v$. That is,

$$H_0 : \Theta \in \{\Theta^*\} \quad (i.e., \mathrm{vech}(\log \Theta) = E\theta), \qquad H_1 : \Theta \notin \{\Theta^*\}.$$

We first *fix* n (and hence v and s). Recall (3.2):

$$\hat{\theta}_T = \hat{\theta}_T(W) := \arg\min_{b\in\mathbb{R}^s}[\mathrm{vech}(\log \hat{\Theta}_T) - Eb]^\intercal W[\mathrm{vech}(\log \hat{\Theta}_T) - Eb] =: \arg\min_{b\in\mathbb{R}^s} g_T(b)^\intercal W g_T(b).$$

Proposition 3.2. *Fix n (and hence v and s).*

(i) Suppose μ is unknown but D is known. Let Assumptions 3.1(i), 3.3, and 3.4 be satisfied. Thus, under H_0,

$$Tg_{T,D}(\hat{\theta}_{T,D})^\intercal \hat{S}_{T,D}^{-1} g_{T,D}(\hat{\theta}_{T,D}) \xrightarrow{d} \chi^2_{n(n+1)/2-s}, \tag{3.5}$$

where

$$g_{T,D}(b) := \mathrm{vech}(\log \hat{\Theta}_{T,D}) - Eb$$
$$\hat{S}_{T,D} := D_n^+ \hat{H}_{T,D}(D^{-1/2} \otimes D^{-1/2})\hat{V}_T(D^{-1/2} \otimes D^{-1/2})\hat{H}_{T,D} D_n^{+\intercal}.$$

(ii) Suppose both μ and D are unknown. Let Assumptions 3.1(i), 3.3, 3.4 and 3.5 be satisfied. Thus, under H_0,

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

where

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

Proof. See SM 8.6. $\square$

From Proposition 3.2, we can easily get the following result of the diagonal path asymptotics, which is more general than the sequential asymptotics but less general than the joint asymptotics (see Phillips and Moon (1999)).

Corollary 3.3.

(i) Suppose μ is unknown but D is known. Let Assumptions 3.1(i), 3.3, and 3.4 be satisfied. Under H_0,

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

where $n = n_T$ as $T \to \infty$.

(ii) Suppose both μ and D are unknown. Let Assumptions 3.1(i), 3.3, 3.4 and 3.5 be satisfied. Under H_0, as $T \to \infty$,

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

where $n = n_T$ as $T \to \infty$.

Proof. See SM 8.6. □

4 The QMLE and One-Step Estimator

4.1 The QMLE

In the context of Gaussian quasi-maximum likelihood estimation (QMLE), given a factorization $n = n_1 \cdot n_2 \cdots n_v$, we shall additionally assume that the Kronecker product model $\{\Theta^*\}$ is correctly specified (i.e. $\mathrm{vech}(\log \Theta) = E\theta$). Let $\rho \in [-1,1]^{s_\rho}$ be original parameters of the Kronecker product model; we have mentioned that

$$s_\rho = \sum_{j=1}^{v} n_j(n_j - 1)/2.$$

The log likelihood function in terms of original parameters ρ for a sample $\{x_1, x_2, \ldots, x_T\}$ is given by

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

Write $\Omega = \Omega(\theta) := \log \Theta = \log \Theta(\rho)$. The log likelihood function in terms of log parameters θ for a sample $\{x_1, x_2, \ldots, x_T\}$ is given by

$$\begin{aligned}&\ell_T(\mu, D, \theta)\\ &= -\frac{Tn}{2}\log(2\pi) - \frac{T}{2}\log\left|D^{1/2}\exp(\Omega(\theta))D^{1/2}\right| - \frac{1}{2}\sum_{t=1}^{T}(x_t - \mu)^\intercal D^{-1/2}[\exp(\Omega(\theta))]^{-1}D^{-1/2}(x_t - \mu).\end{aligned} \tag{4.1}$$

In practice, conditional on some estimates of μ and D, we use an iterative algorithm based on the derivatives of ℓ_T with respect to either ρ or θ to compute the QMLE of either ρ or θ. Proposition 4.1 below provides formulas for the derivatives of ℓ_T with respect to θ. The computations required are typically not too onerous, since for example the Hessian matrix is of an order $\log n$ by $\log n$. See Singull et al. (2012) and Ohlson et al. (2013) for a discussion of estimation algorithms in the case where the data are multiway array and v is of low dimension. Nevertheless since there is quite complicated non-linearity involved in the definition of the QMLE, it is not so easy to directly analyse QMLE.

Instead we shall consider a one-step estimator that uses the minimum distance estimator in Section 3 to provide a starting value and then takes a Newton-Raphson step towards the QMLE of θ. In the fixed n it is known that the one-step estimator is equivalent to the QMLE in the sense that it shares its asymptotic distribution (Bickel (1975)).

Below, for slightly abuse of notation, we shall use μ, D, θ to denote the true parameter (i.e., characterising the data generating process) as well as the generic parameter of the likelihood function; we will be more specific whenever any confusion is likely to arise.

4.2 The One-Step Estimator

Here we only examine the one-step estimator when μ is unknown but D is known. When neither μ nor D is known, one has to differentiate (4.1) with respect to both θ and D. The analysis becomes considerably more involved and we leave it for future work. Suppose D is known, the likelihood function (4.1) reduces to

$$\ell_{T,D}(\theta,\mu) =$$
$$-\frac{Tn}{2}\log(2\pi) - \frac{T}{2}\log\left|D^{1/2}\exp(\Omega(\theta))D^{1/2}\right| - \frac{1}{2}\sum_{t=1}^{T}(x_t-\mu)^\intercal D^{-1/2}[\exp(\Omega(\theta))]^{-1}D^{-1/2}(x_t-\mu). \tag{4.2}$$

It is a well-known result that for any choice of Σ (i.e., D and θ), the QMLE for μ is $\bar{x}$. Hence we may define

$$\hat{\theta}_{QMLE,D} = \arg\max_{\theta} \ell_{T,D}(\theta,\bar{x}).$$

Proposition 4.1. *The $s\times 1$ score function of (4.2) with respect to θ takes the following form*

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

The $s\times s$ block corresponding to θ of the Hessian matrix of (4.2) takes the following form

$$\frac{\partial^2 \ell_{T,D}(\theta,\mu)}{\partial\theta\partial\theta^\intercal} =$$
$$-\frac{T}{2}E^\intercal D_n^\intercal \Psi_1\left(e^{-\Omega}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}\otimes I_n + I_n\otimes e^{-\Omega}D^{-1/2}\tilde{\Sigma}_T D^{-1/2} - I_{n^2}\right)\left(e^{-\Omega}\otimes e^{-\Omega}\right)\Psi_1 D_n E$$
$$+\frac{T}{2}(\Psi_2^\intercal\otimes E^\intercal D_n^\intercal)\int_0^1 P_K\left(I_{n^2}\otimes \operatorname{vec} e^{(1-t)\Omega}\right)\int_0^1 e^{st\Omega}\otimes e^{(1-s)t\Omega}ds\cdot t dt D_n E$$
$$+\frac{T}{2}(\Psi_2^\intercal\otimes E^\intercal D_n^\intercal)\int_0^1 P_K\left(\operatorname{vec} e^{t\Omega}\otimes I_{n^2}\right)\int_0^1 e^{s(1-t)\Omega}\otimes e^{(1-s)(1-t)\Omega}ds\cdot(1-t)dt D_n E,$$

where $P_K := I_n\otimes K_{n,n}\otimes I_n$ and

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

$$\Psi_2 = \Psi_2(\theta) := \operatorname{vec}\left(e^{-\Omega(\theta)}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}e^{-\Omega(\theta)} - e^{-\Omega(\theta)}\right).$$

Proof. See SM 8.4. □

Since $\mathbb{E}\Psi_2(\theta)=0$, where θ denotes the true parameter, so the negative normalized expected Hessian matrix evaluated at the true parameter θ takes the following form

$$\Upsilon_D := \mathbb{E}\left[-\frac{1}{T}\frac{\partial^2\ell_{T,D}(\theta,\mu)}{\partial\theta\partial\theta^\intercal}\right] = \frac{1}{2}E^\intercal D_n^\intercal\Psi_1(\theta)\left(e^{-\Omega(\theta)}\otimes e^{-\Omega(\theta)}\right)\Psi_1(\theta)D_n E$$
$$= \frac{1}{2}E^\intercal D_n^\intercal\int_0^1\int_0^1 e^{(t+s-1)\Omega}\otimes e^{(1-t-s)\Omega}dtds D_n E$$
$$= \frac{1}{2}E^\intercal D_n^\intercal\left[\int_0^1\int_0^1 \Theta^{t+s-1}\otimes\Theta^{1-t-s}dtds\right]D_n E =: \frac{1}{2}E^\intercal D_n^\intercal \Xi D_n E.$$

Under normality (i.e., $V = 2D_nD_n^+(\Sigma\otimes\Sigma)$), one can verify that $\Upsilon_D = \mathbb{E}\left[\frac{1}{T}\frac{\partial \ell_{T,D}(\theta,\mu)}{\partial\theta^\intercal}\frac{\partial \ell_{T,D}(\theta,\mu)}{\partial\theta}\right]$. We then propose the following one-step estimator in the spirit of van der Vaart (1998) p72 or Newey and McFadden (1994) p2150:

$$\tilde{\theta}_{T,D} := \hat{\theta}_{T,D} - \hat{\Upsilon}_{T,D}^{-1}\frac{\partial \ell_{T,D}(\hat{\theta}_{T,D},\bar{x})}{\partial\theta^\intercal}/T, \tag{4.3}$$

where $\hat{\Upsilon}_{T,D} := \frac{1}{2}E^\intercal D_n^\intercal\left[\int_0^1\int_0^1 \hat{\Theta}_T^{t+s-1}\otimes\hat{\Theta}_T^{1-t-s}dtds\right]D_nE =: \frac{1}{2}E^\intercal D_n^\intercal\hat{\Xi}_TD_nE$. (We show in SM 8.5 that $\hat{\Upsilon}_{T,D}$ is invertible with probability approaching 1.) We did not use the plain vanilla one-step estimator because the Hessian matrix $\frac{\partial^2\ell_{T,D}(\theta,\mu)}{\partial\theta\partial\theta^\intercal}$ is rather complicated to analyse.

4.3 Large Sample Properties

Assumption 4.1. *For every positive constant M and uniformly in $b\in\mathbb{R}^s$ with $\|b\|_2=1$,*

$$\sup_{\theta^*:\|\theta^*-\theta\|_2\leq M\sqrt{\frac{n\varpi\kappa(W)}{T}}}\left|\sqrt{T}b^\intercal\left[\frac{1}{T}\frac{\partial\ell_{T,D}(\theta^*,\bar{x})}{\partial\theta^\intercal}-\frac{1}{T}\frac{\partial\ell_{T,D}(\theta,\bar{x})}{\partial\theta^\intercal}-\Upsilon_D(\theta^*-\theta)\right]\right| = o_p(1).$$

Assumption 4.1 is one of the sufficient conditions needed for the asymptotic normality of $\tilde{\theta}_{T,D}$ (Theorem 4.1). This kind of assumption is standard in the asymptotics of one-step estimators (see (5.44) of van der Vaart (1998) p71) or of M-estimation (see (C3) of He and Shao (2000)). Assumption 4.1 implies that $\frac{1}{T}\frac{\partial\ell_{T,D}(\theta,\bar{x})}{\partial\theta^\intercal}$ is differentiable at the true parameter θ, with derivative tending to Υ_D in probability. The radius of the shrinking neighbourhood $\sqrt{n\varpi\kappa(W)/T}$ is determined by the rate of convergence of any preliminary estimator, say, $\hat{\theta}_{T,D}$ in our case. It is possible to relax the $o_p(1)$ on the right side of Assumption 4.1 to $o_p(\sqrt{n/(\varpi^2\log n)})$ by examining the proof of Theorem 4.1.

We next provide the large sample theory for $\tilde{\theta}_{T,D}$.

Theorem 4.1. *Suppose that the Kronecker product model $\{\Theta^*\}$ is correctly specified. Let Assumptions 3.1(ii), 3.2(iii), 3.3, and 4.1 be satisfied. Then*

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

for any $s\times 1$ vector b with $\|b\|_2=1$.

Proof. See SM 8.5. □

Theorem 4.1 is a version of the large-dimensional CLT, which is difficult to derive. It has the same structure as that of Theorem 3.1 or Theorem 3.2. Note that if we replace normality (Assumption 3.1(ii)) with the subgaussian assumption (Assumption 3.1(i)) - that is Gaussian likelihood is not correctly specified - although the norm consistency of $\tilde{\theta}_{T,D}$ should still hold, the asymptotic variance in Theorem 4.1 needs to be changed to have a sandwich formula. Theorem 4.1 says that $\sqrt{T}b^\intercal(\tilde{\theta}_{T,D}-\theta)\xrightarrow{d}N\left(0,b^\intercal\left(\mathbb{E}\left[-\frac{1}{T}\frac{\partial^2\ell_{T,D}(\theta,\mu)}{\partial\theta\partial\theta^\intercal}\right]\right)^{-1}b\right)$. In the fixed n case, this estimator achieves the parametric efficiency bound by a well-known result $\frac{\partial^2\ell_{T,D}(\theta,\mu)}{\partial\mu\partial\theta^\intercal}=0$. This shows that our one-step estimator $\tilde{\theta}_{T,D}$ is efficient when D (the variances) is known.

By recognising that $H^{-1}=\int_0^1 e^{t\log\Theta}\otimes e^{(1-t)\log\Theta}dt=\Psi_1$, (see Proposition 8.6 in SM 8.6), we see that, when D is known, under normality and correct specification of the Kronecker product model, $\tilde{\theta}_{T,D}$ and the optimal minimum distance estimator $\hat{\theta}_{T,D}(W_{op})$ have the same asymptotic variance, i.e., $\left(\frac{1}{2}E^\intercal D_n^\intercal H^{-1}(\Theta^{-1}\otimes\Theta^{-1})H^{-1}D_nE\right)^{-1}$.

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

Corollary 4.1. *Suppose the Kronecker product model $\{\Theta^*\}$ is correctly specified. Let Assumptions 3.1(ii), 3.2(iii), 3.3, and 4.1 be satisfied. Given a full-column-rank $s \times k$ matrix A where k is finite with $\|A\|_{\ell_2} = O_p(\sqrt{\log n \cdot n})$, we have*

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

Proof. Essentially the same as that of Corollary 3.1. □

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

5 Model Selection

We briefly discuss the issue of model selection here. One shall not worry about this if the data are in the multi-index format with v multiplicative factors. This is because in this setting the Kronecker product model is pinned down by the structure of multiway arrays - there is no model uncertainty. This issue will pop up when one uses a Kronecker product model to approximate a general covariance or correlation matrix.

First, note that for a given Kronecker product model, if one permutes the data, the performance of this Kronecker product model is likely to vary. Thus in practice one needs to investigate sensitivity of performance of a Kronecker product model when permuting the data.

Second, if one fixes the ordering of the data as well as factorization $n = n_1 \cdots n_v$, but simply permutes Θ_j^*s, one obtains a different Θ^* (i.e., a different Kronecker product model). Although the eigenvalues of these two Kronecker product models are the same, the eigenvectors of them are not.

Third, if one fixes the ordering of the data, but uses a different factorization of n, one then obtains a completely different Kronecker product model. Suppose that n has the prime factorization $n = p_1 p_2 \cdots p_v$ for some positive integer v ($v \geq 2$) and primes p_j for $j = 1, \ldots, v$. Then there exist several different Kronecker product models, each of which is indexed by the dimensions of the sub-matrices. The base model has dimensions $(p_1, p_2, \ldots, p_v)$, but there are many possible aggregations of this, for example, $\big((p_1 + p_2), \ldots, (p_{v-1} + p_v)\big)$.

To address the second and third issues, we might choose between those Kronecker product models using some model selection criterion that penalizes the larger models. For example,

$$BIC(\rho) = -2\ell_T(\mu, D, \rho) + s_\rho \log T,$$

where $\ell_T(\cdot, \cdot, \cdot)$ is the log likelihood function defined in Section 4, ρ is original parameters associated with a Kronecker product model, and s_ρ is dimension of ρ. Typically there are not so many factorizations to consider, so this is not too computationally burdensome.

6 Numerical Studies and an Application

6.1 Numerical Studies

We first provide a small simulation study that evaluates the performance of the QMLE, and then apply our model to daily stock returns.

We simulate T random vectors x_t of dimension n according to

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

where $n = 2^v$ and $v \in \mathbb{N}$. The sub-matrices Σ_j are 2×2. These sub-matrices Σ_j are generated with unit variances and off-diagonal elements drawn from a uniform distribution on $(0, 1)$. This ensures positive definiteness of Σ. Due to the unit variances, Σ is both the covariance and correlation matrix of x_t, but the econometrician is unaware of this and applies a Kronecker product model to the covariance matrix of x_t. We shall consider the correctly specified case, i.e., the Kronecker product model has a factorization $n = 2^v$. The sample size is set to $T = 300$. We shall adopt the first approach of estimation to estimate original parameters directly. For identification, the upper diagonal elements of $\Sigma_j, j \geq 2$, are set to 1; altogether, there are $2v+1$ original parameters to estimate by quasi-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 product model $\hat{\Sigma}_{\text{Kron}}$ with respect to the sample covariance estimator $\hat{\Sigma}_T$. It is defined as

$$\text{PRIAL1} = 1 - \frac{\mathbb{E}\|\hat{\Sigma}_{\text{Kron}} - \Sigma\|_F^2}{\mathbb{E}\|\hat{\Sigma}_T - \Sigma\|_F^2}.$$

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

$$\text{PRIAL2} = 1 - \frac{\mathbb{E}\|\hat{\Sigma}_{\text{Kron}}^{-1} - \Sigma^{-1}\|_F^2}{\mathbb{E}\|\hat{\Sigma}_T^{-1} - \Sigma^{-1}\|_F^2}.$$

Note that this requires invertibility of the sample covariance matrix $\hat{\Sigma}_T$ 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. 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 $\iota_n = (1, 1, \ldots, 1)^\intercal$ of dimension n, see e.g., Ledoit and Wolf (2003) and Chan, Karceski, and Lakonishok (1999). The inverse of a Kronecker product model is easily found by inverting the sub-matrices Σ_j^*, which can be done analytically, since

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

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

$$\begin{aligned}
w_{MV} &= \frac{\left((\Sigma_1^*)^{-1} \otimes (\Sigma_2^*)^{-1} \otimes \cdots \otimes (\Sigma_v^*)^{-1}\right)\iota_n}{\iota_n^\intercal \left((\Sigma_1^*)^{-1} \otimes (\Sigma_2^*)^{-1} \otimes \cdots \otimes (\Sigma_v^*)^{-1}\right)\iota_n} \\
&= \frac{(\Sigma_1^*)^{-1}\iota_{n_1}}{\iota_{n_1}^\intercal(\Sigma_1^*)^{-1}\iota_{n_1}} \otimes \frac{(\Sigma_2^*)^{-1}\iota_{n_2}}{\iota_{n_2}^\intercal(\Sigma_2^*)^{-1}\iota_{n_2}} \otimes \cdots \otimes \frac{(\Sigma_v^*)^{-1}\iota_{n_v}}{\iota_{n_v}^\intercal(\Sigma_v^*)^{-1}\iota_{n_v}}, \\
\text{var}(w_{MV}^\intercal x_t) &= \frac{1}{\iota_{n_1}^\intercal(\Sigma_1^*)^{-1}\iota_{n_1} \times \cdots \times \iota_{n_v}^\intercal(\Sigma_v^*)^{-1}\iota_{n_v}}.
\end{aligned}$$

In cases where n is large, this structure is very convenient computationally. The first portfolio weights are constructed using the sample covariance matrix $\hat{\Sigma}_T$ and the second portfolio weights are constructed using the Kronecker product model $\hat{\Sigma}_{\text{Kron}}$. These two portfolios are then evaluated (by calculating the variance) using the out-of-sample returns generated using the same data generating mechanism. The ratio of the variance of the latter portfolio over that of the former (VR) is recorded. See Fan, Liao, and Shi (2015) for a discussion of risk estimation for large dimensional portfolio choice problems.

We repeat the simulation 1000 times and obtain for each simulation PRIAL1, PRIAL2 and VR. Table 1 reports the median of the obtained PRIALs and VR for various dimensions.

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

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

Clearly, as the dimension increases, the Kronecker product model 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 product model 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 An Application

We now 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 $T = 2732$.

Kronecker product models are fitted to the correlation matrix $\Theta = D^{-1/2}\Sigma D^{-1/2}$, where D is the diagonal matrix containing the variances of x_t. The first Kronecker 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.

Again we adopt the first approach of estimation to estimate original parameters directly. The estimation is done in two steps: First, D is estimated using the sample variances, and then the original parameters of Θ^* are estimated by quasi-maximum likelihood estimation using the standardized returns $\hat{D}^{-1/2}x_t$ and pseudo variables u_t. Re-ordering the data x_t according to variance in a descending way prior to adding the pseudo variables u_t did not improve the final outcomes, so we keep the original order of the data. We experiment with more generous decompositions by looking at all factorizations of numbers from 441 to 512, and selecting some yielding not more than 30 parameters. Table 2 gives a summary of these models.

Next, we follow the approach of Fan et al. (2013) and estimate the Kronecker product model on windows of size 504 days (equal to two years' trading days) that are shifted from the beginning to the end of the sample. The estimated Kronecker product model yields an estimator of the covariance matrix that is used to construct the minimum variance portfolio (MVP) weights. The same is done for two competing devices: the sample covariance matrix and the linear shrinkage estimator of Ledoit and Wolf (2004). After each estimation, the minimum variance portfolios constructed by these three models are compared in terms of standard error using the next 21 days (equal to one month's trading days) out-of-sample. Then the estimation window of 504 days is shifted by 21 days, etc. The total number of out-of-sample evaluations is 106.

The last four columns of Table 2 summarize the relative performance of the Kronecker MVP with respect to those of the sample covariance matrix and the linear shrinkage estimator of Ledoit and Wolf (2004). We consider two criteria: *Impr* and *Prop*. *Impr* is the average of standard error improvements (in percentage) and *Prop* is the proportion of the times (out of 106) that the Kronecker MVP outperforms a competing MVP. All models outperform the

			Sample Cov		Ledoit-Wolf	
Model	p	Decomp	Impr	Prop	Impr	Prop
M1	9	$512 = 2^9$	30%	0.92	8%	0.84
M2	16	$486 = 2 \times 3^5$	32%	0.92	0%	0.50
M3	17	$512 = 2^5 \times 4^2$	32%	0.92	11%	0.94
M4	18	$480 = 2^5 \times 3 \times 5$	33%	0.92	-2%	0.39
M5	25	$512 = 4^4 \times 2$	34%	0.92	13%	0.94
M6	27	$448 = 2^6 \times 7$	35%	0.92	-30%	0.09

Table 2: *Summary of Kronecker product models for the correlation matrix of $(x_t^\intercal, u_t^\intercal)^\intercal$. p is the number of original parameters in a Kronecker product model. Decomp is the factorization used for the full system including the additional pseudo variables. Prop is the proportion of the times that the Kronecker MVP outperforms a competing MVP (generated by the sample covariance matrix, or the Ledoit-Wolf estimator), and Impr is the average of standard error improvements (in percentage).*

sample covariance matrix, while models with smaller dimensional sub-matrices (i.e., M1, M3 and M5) tend to outperform the shrinkage estimator. The reason could be that it is more difficult to ensure positive definiteness of a bigger sub-matrix in the constrained maximum likelihood optimisation.

7 Conclusions

We have established the large sample properties of our estimation methods of Kronecker product models in the large dimensional case. In particular, we obtained norm consistency and the large dimensional CLTs. The Kronecker product model outperforms the sample covariance matrix theoretically, in a simulation study, and in an application to portfolio choice. It is possible to extend the framework in various directions to improve performance. One may also consider the case where both n_j and v increase with the sample size.

A Appendix

A.1

Proposition A.1. *Suppose that*

$$\Theta^* = \Theta_1^* \otimes \Theta_2^* \otimes \cdots \otimes \Theta_v^*,$$

where Θ_j^* is $n_j \times n_j$ dimensional such that $n = n_1 \cdot n_2 \cdots n_v$. Taking the logarithm on both sides gives

$$\log \Theta^* = \log \Theta_1^* \otimes I_{n_2} \otimes \cdots \otimes I_{n_v} + I_{n_1} \otimes \log \Theta_2^* \otimes I_{n_3} \otimes \cdots \otimes I_{n_v} + \cdots$$
$$+ I_{n_1} \otimes I_{n_2} \otimes \cdots \otimes \log \Theta_v^*.$$

For identification we set the first diagonal entry of $\log \Theta_j^*$ to be 0 for $j = 1, \ldots, v-1$. In total there are

$$s := \sum_{j=1}^{n} \frac{n_j(n_j+1)}{2} - (v-1)$$

(identifiable) parameters in $\log \Theta_1^*, \ldots, \log \Theta_v^*$; let θ^* denote these. Then there exists a $n(n+1)/2 \times s$ full column rank matrix E such that

$$\mathrm{vech}(\log \Theta^*) = E\theta^*.$$

Proof. Note that

$$\log \Theta^* = \log \Theta_1^* \otimes I_{n_2} \otimes \cdots \otimes I_{n_v} + I_{n_1} \otimes \log \Theta_2^* \otimes I_{n_3} \otimes \cdots \otimes I_{n_v} + \cdots \\ + I_{n_1} \otimes I_{n_2} \otimes \cdots \otimes \log \Theta_v^*.$$

Then

$$\operatorname{vech}(\log \Theta^*) = \begin{bmatrix} E_1 & E_2 & \cdots & E_v \end{bmatrix} \begin{bmatrix} \operatorname{vech}(\log \Theta_1^*) \\ \operatorname{vech}(\log \Theta_2^*) \\ \vdots \\ \operatorname{vech}(\log \Theta_v^*) \end{bmatrix},$$

where for $i = 1, \ldots, v$

$$E_i := D_n^+ \left[I_{n_1 \cdot n_2 \cdots n_i} \otimes K_{n/(n_1 \cdot n_2 \cdots n_i), n_1 \cdot n_2 \cdots n_i} \otimes I_{n/(n_1 \cdot n_2 \cdots n_i)} \right] \left[I_{(n_1 \cdot n_2 \cdots n_i)^2} \otimes \operatorname{vec} I_{n/(n_1 \cdot n_2 \cdots n_i)} \right] \cdot \\ (I_{n_1 \cdot n_2 \cdots n_{i-1}} \otimes K_{n_i, n_1 \cdot n_2 \cdots n_{i-1}} \otimes I_{n_i})(\operatorname{vec} I_{n_1 \cdot n_2 \cdots n_{i-1}} \otimes I_{n_i^2}) D_{n_i}, \tag{A.1}$$

where D_n^+ is the Moore-Penrose generalised inverse of D_n, i.e. $D_n^+ = (D_n^\intercal D_n)^{-1} D_n^\intercal$, D_n and D_{n_i} are the $n^2 \times n(n+1)/2$ and $n_i^2 \times n_i(n_i+1)/2$ duplication matrices, respectively, and $K_{n/(n_1 \cdot n_2 \cdots n_i), n_1 \cdot n_2 \cdots n_i}$ and $K_{n_i, n_1 \cdot n_2 \cdots n_{i-1}}$ are commutation matrices of various dimensions. When $i = 1$, the term $(I_{n_1 \cdot n_2 \cdots n_{i-1}} \otimes K_{n_i, n_1 \cdot n_2 \cdots n_{i-1}} \otimes I_{n_i})(\operatorname{vec} I_{n_1 \cdot n_2 \cdots n_{i-1}} \otimes I_{n_i^2})$ in (A.1) is set to be 1. To see this, we first consider $\operatorname{vec}(\log \Theta_1^* \otimes I_{n_2} \otimes \cdots \otimes I_{n_v})$.

$$\begin{aligned} &\operatorname{vec}(\log \Theta_1^* \otimes I_{n_2} \otimes \cdots \otimes I_{n_v}) = \operatorname{vec}(\log \Theta_1^* \otimes I_{n/n_1}) \\ &= \left(I_{n_1} \otimes K_{n/n_1, n_1} \otimes I_{n/n_1} \right) \left(\operatorname{vec}(\log \Theta_1^*) \otimes \operatorname{vec} I_{n/n_1} \right) \\ &= \left(I_{n_1} \otimes K_{n/n_1, n_1} \otimes I_{n/n_1} \right) \left(I_{n_1^2} \operatorname{vec}(\log \Theta_1^*) \otimes \operatorname{vec} I_{n/n_1} \cdot 1 \right) \\ &= \left(I_{n_1} \otimes K_{n/n_1, n_1} \otimes I_{n/n_1} \right) \left(I_{n_1^2} \otimes \operatorname{vec} I_{n/n_1} \right) \operatorname{vec}(\log \Theta_1^*), \end{aligned}$$

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

$$\operatorname{vech}(\log \Theta_1^* \otimes I_{n_2} \otimes \cdots \otimes I_{n_v}) = D_n^+ \left(I_{n_1} \otimes K_{n/n_1, n_1} \otimes I_{n/n_1} \right) \left(I_{n_1^2} \otimes \operatorname{vec} I_{n/n_1} \right) D_{n_1} \operatorname{vech}(\log \Theta_1^*). \tag{A.2}$$

We now consider $\operatorname{vec}(I_{n_1} \otimes \log \Theta_2^* \otimes \cdots \otimes I_{n_v})$.

$$\begin{aligned} &\operatorname{vec}(I_{n_1} \otimes \log \Theta_2^* \otimes \cdots \otimes I_{n_v}) = \operatorname{vec}(I_{n_1} \otimes \log \Theta_2^* \otimes I_{n/(n1 \cdot n_2)}) \\ &= (I_{n_1 \cdot n_2} \otimes K_{n/(n_1 \cdot n_2), n_1 \cdot n_2} \otimes I_{n/(n_1 \cdot n_2)}) \left(\operatorname{vec}(I_{n_1} \otimes \log \Theta_2^*) \otimes \operatorname{vec} I_{n/(n_1 \cdot n_2)} \right) \\ &= (I_{n_1 \cdot n_2} \otimes K_{n/(n_1 \cdot n_2), n_1 \cdot n_2} \otimes I_{n/(n_1 \cdot n_2)}) \left(I_{(n_1 \cdot n_2)^2} \otimes \operatorname{vec} I_{n/(n_1 \cdot n_2)} \right) \operatorname{vec}(I_{n_1} \otimes \log \Theta_2^*) \\ &= (I_{n_1 \cdot n_2} \otimes K_{n/(n_1 \cdot n_2), n_1 \cdot n_2} \otimes I_{n/(n_1 \cdot n_2)}) \left(I_{(n_1 \cdot n_2)^2} \otimes \operatorname{vec} I_{n/(n_1 \cdot n_2)} \right) \cdot \\ &\quad (I_{n_1} \otimes K_{n_2, n_1} \otimes I_{n_2})(\operatorname{vec} I_{n_1} \otimes \operatorname{vec} \log(\Theta_2^*)) \\ &= (I_{n_1 \cdot n_2} \otimes K_{n/(n_1 \cdot n_2), n_1 \cdot n_2} \otimes I_{n/(n_1 \cdot n_2)}) \left(I_{(n_1 \cdot n_2)^2} \otimes \operatorname{vec} I_{n/(n_1 \cdot n_2)} \right) \cdot \\ &\quad (I_{n_1} \otimes K_{n_2, n_1} \otimes I_{n_2})(\operatorname{vec} I_{n_1} \otimes I_{n_2^2}) \operatorname{vec} \log(\Theta_2^*). \end{aligned}$$

Thus

$$\begin{aligned} &\operatorname{vech}(I_{n_1} \otimes \log \Theta_2^* \otimes \cdots \otimes I_{n_v}) \\ &= D_n^+ (I_{n_1 \cdot n_2} \otimes K_{n/(n_1 \cdot n_2), n_1 \cdot n_2} \otimes I_{n/(n_1 \cdot n_2)}) \left(I_{(n_1 \cdot n_2)^2} \otimes \operatorname{vec} I_{n/(n_1 \cdot n_2)} \right) \cdot \\ &\quad (I_{n_1} \otimes K_{n_2, n_1} \otimes I_{n_2})(\operatorname{vec} I_{n_1} \otimes I_{n_2^2}) D_{n_2} \operatorname{vech} \log(\Theta_2^*). \end{aligned} \tag{A.3}$$

Next we consider $\operatorname{vec}(I_{n_1} \otimes I_{n_2} \otimes \log \Theta_3^* \otimes \cdots \otimes I_{n_v})$.

$$
\begin{aligned}
&\operatorname{vec}(I_{n_1} \otimes I_{n_2} \otimes \log \Theta_3^* \otimes \cdots \otimes I_{n_v}) = \operatorname{vec}(I_{n_1 \cdot n_2} \otimes \log \Theta_3^* \otimes I_{n/(n_1 \cdot n_2 \cdot n_3)}) \\
&= (I_{n_1 \cdot n_2 \cdot n_3} \otimes K_{n/(n_1 \cdot n_2 \cdot n_3), n_1 \cdot n_2 \cdot n_3} \otimes I_{n/(n_1 \cdot n_2 \cdot n_3)}) \left(\operatorname{vec}(I_{n_1 \cdot n_2} \otimes \log \Theta_3^*) \otimes \operatorname{vec} I_{n/(n_1 \cdot n_2 \cdot n_3)}\right) \\
&= (I_{n_1 \cdot n_2 \cdot n_3} \otimes K_{n/(n_1 \cdot n_2 \cdot n_3), n_1 \cdot n_2 \cdot n_3} \otimes I_{n/(n_1 \cdot n_2 \cdot n_3)}) \left(I_{(n_1 \cdot n_2 \cdot n_3)^2} \otimes \operatorname{vec} I_{n/(n_1 \cdot n_2 \cdot n_3)}\right) \operatorname{vec}(I_{n_1 \cdot n_2} \otimes \log \Theta_3^*) \\
&= (I_{n_1 \cdot n_2 \cdot n_3} \otimes K_{n/(n_1 \cdot n_2 \cdot n_3), n_1 \cdot n_2 \cdot n_3} \otimes I_{n/(n_1 \cdot n_2 \cdot n_3)}) \left(I_{(n_1 \cdot n_2 \cdot n_3)^2} \otimes \operatorname{vec} I_{n/(n_1 \cdot n_2 \cdot n_3)}\right) \cdot \\
&\qquad (I_{n_1 \cdot n_2} \otimes K_{n_3, n_1 \cdot n_2} \otimes I_{n_3})(\operatorname{vec} I_{n_1 \cdot n_2} \otimes \operatorname{vec}(\log \Theta_3^*)) \\
&= (I_{n_1 \cdot n_2 \cdot n_3} \otimes K_{n/(n_1 \cdot n_2 \cdot n_3), n_1 \cdot n_2 \cdot n_3} \otimes I_{n/(n_1 \cdot n_2 \cdot n_3)}) \left(I_{(n_1 \cdot n_2 \cdot n_3)^2} \otimes \operatorname{vec} I_{n/(n_1 \cdot n_2 \cdot n_3)}\right) \cdot \\
&\qquad (I_{n_1 \cdot n_2} \otimes K_{n_3, n_1 \cdot n_2} \otimes I_{n_3})(\operatorname{vec} I_{n_1 \cdot n_2} \otimes I_{n_3^2}) \operatorname{vec}(\log \Theta_3^*).
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\operatorname{vech}(I_{n_1} \otimes I_{n_2} \otimes \log \Theta_3^* \otimes \cdots \otimes I_{n_v}) \\
&= D_n^+ (I_{n_1 \cdot n_2 \cdot n_3} \otimes K_{n/(n_1 \cdot n_2 \cdot n_3), n_1 \cdot n_2 \cdot n_3} \otimes I_{n/(n_1 \cdot n_2 \cdot n_3)}) \left(I_{(n_1 \cdot n_2 \cdot n_3)^2} \otimes \operatorname{vec} I_{n/(n_1 \cdot n_2 \cdot n_3)}\right) \cdot \\
&\qquad (I_{n_1 \cdot n_2} \otimes K_{n_3, n_1 \cdot n_2} \otimes I_{n_3})(\operatorname{vec} I_{n_1 \cdot n_2} \otimes I_{n_3^2}) D_{n_3} \operatorname{vech}(\log \Theta_3^*).
\end{aligned} \tag{A.4}
$$

By observing (A.2), (A.3) and (A.4), we deduce the following general formula: for $i = 1, 2, \ldots, v$

$$
\begin{aligned}
&\operatorname{vech}(I_{n_1} \otimes \cdots \otimes \log \Theta_i^* \otimes \cdots \otimes I_{n_v}) \\
&= D_n^+ \left[I_{n_1 \cdot n_2 \cdots n_i} \otimes K_{n/(n_1 \cdot n_2 \cdots n_i), n_1 \cdot n_2 \cdots n_i} \otimes I_{n/(n_1 \cdot n_2 \cdots n_i)}\right] \left[I_{(n_1 \cdot n_2 \cdots n_i)^2} \otimes \operatorname{vec} I_{n/(n_1 \cdot n_2 \cdots n_i)}\right] \cdot \\
&\qquad (I_{n_1 \cdot n_2 \cdots n_{i-1}} \otimes K_{n_i, n_1 \cdot n_2 \cdots n_{i-1}} \otimes I_{n_i})(\operatorname{vec} I_{n_1 \cdot n_2 \cdots n_{i-1}} \otimes I_{n_i^2}) D_{n_i} \operatorname{vech}(\log \Theta_i^*) \\
&=: E_i \operatorname{vech}(\log \Theta_i^*),
\end{aligned} \tag{A.5}
$$

where E_i is a $n(n+1)/2 \times n_i(n_i+1)/2$ matrix. When $i = 1$, the term $(I_{n_1 \cdot n_2 \cdots n_{i-1}} \otimes K_{n_i, n_1 \cdot n_2 \cdots n_{i-1}} \otimes I_{n_i})(\operatorname{vec} I_{n_1 \cdot n_2 \cdots n_{i-1}} \otimes I_{n_i^2})$ in E_i is set to be 1. Using (A.5), we have

$$
\begin{aligned}
\operatorname{vech}(\log \Theta^*) &= E_1 \operatorname{vech}(\log \Theta_1^*) + E_2 \operatorname{vech}(\log \Theta_2^*) + \cdots + E_v \operatorname{vech}(\log \Theta_v^*) \\
&= \begin{bmatrix} E_1 & E_2 & \cdots & E_v \end{bmatrix} \begin{bmatrix} \operatorname{vech}(\log \Theta_1^*) \\ \operatorname{vech}(\log \Theta_2^*) \\ \vdots \\ \operatorname{vech}(\log \Theta_v^*) \end{bmatrix}
\end{aligned}
$$

For identification we set the first diagonal entry of $\log \Theta_j^*$ to be 0 for $j = 1, \ldots, v-1$. In total there are

$$
s := \sum_{j=1}^{n} \frac{n_j(n_j+1)}{2} - (v-1)
$$

(identifiable) parameters in $\log \Theta_1^*, \ldots, \log \Theta_v^*$; let θ^* denote these. Then there exists a $n(n+1)/2 \times s$ full column rank matrix E such that

$$
\operatorname{vech}(\log \Theta^*) = E\theta^*,
$$

where

$$
E = \begin{bmatrix} E_{1,(-1)} & E_{2,(-1)} & \cdots & E_{v-1,(-1)} & E_v \end{bmatrix}
$$

and $E_{i,(-1)}$ stands for matrix E_i with its first column removed. □

Proposition A.2. *Given that $n = n_1 \cdot n_2 \cdots n_v$, the $s \times s$ matrix $E^\intercal E$ takes the following form:*

(i) For $i = 1, \ldots, s$, the ith diagonal entry of $E^\intercal E$ records how many times the ith parameter in θ^0 has appeared in $\mathrm{vech}(\log \Theta^0)$. *The value depends on to which $\log \Theta_j^0$ the ith parameter in θ^0, θ_i^0, belongs to. For instance, suppose θ_i^0 is a parameter belonging to $\log \Theta_3^0$, then*

$$(E^\intercal E)_{i,i} = n/n_3.$$

(ii) For $i, k = 1, \ldots, s$ ($i \neq k$), the (i, k) entry of $E^\intercal E$ (or the (k, i) entry of $E^\intercal E$ by symmetry) records how many times the ith parameter in θ^0, θ_i^0, and kth parameter in θ^0, θ_k^0, have appeared together (as summands) in an entry of $\mathrm{vech}(\log \Theta^0)$. *The value depends on to which $\log \Theta_j^0$ the ith parameter in θ^0, θ_i^0, and kth parameter in θ^0, θ_k^0, belong to. For instance, suppose θ_i^0 is a parameter belonging to $\log \Theta_3^0$ and θ_k^0 is a parameter belonging to $\log \Theta_5^0$, then*

$$(E^\intercal E)_{i,k} = (E^\intercal E)_{k,i} = n/(n_3 \cdot n_5).$$

Note that if both θ_i^0 and θ_k^0 belong to the same $\log \Theta_j^0$, then $(E^\intercal E)_{i,k} = (E^\intercal E)_{k,i} = 0$. Also note that when θ_i^0 is an off-diagonal entry of some $\log \Theta_j^0$, then

$$(E^\intercal E)_{i,k} = (E^\intercal E)_{k,i} = 0$$

for any $k = 1, \ldots, s$ ($i \neq k$).

Proof. Proof by spotting the pattern. □

We here give a concrete example to illustrate Proposition A.2.

Example A.1. *Suppose that $n_1 = 3, n_2 = 2, n_3 = 2$. We have*

$$\log \Theta_1^0 = \begin{pmatrix} 0 & a_{1,2} & a_{1,3} \\ a_{1,2} & a_{2,2} & a_{2,3} \\ a_{1,3} & a_{2,3} & a_{3,3} \end{pmatrix} \qquad \log \Theta_2^0 = \begin{pmatrix} 0 & b_{1,2} \\ b_{1,2} & b_{2,2} \end{pmatrix} \qquad \log \Theta_3^0 = \begin{pmatrix} c_{1,1} & c_{1,2} \\ c_{1,2} & c_{2,2} \end{pmatrix}$$

The leading diagonals of $\log \Theta_1^0$ and $\log \Theta_2^0$ are set to zero for identification as explained before. Thus

$$\theta^0 = (a_{1,2}, a_{1,3}, a_{2,2}, a_{2,3}, a_{3,3}, b_{1,2}, b_{2,2}, c_{1,1}, c_{1,2}, c_{2,2})^\intercal.$$

Then we can invoke Proposition A.2 to write down $E^\intercal E$ without even using R code to compute E; that is,

$$E^\intercal E = \begin{pmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 2 & 2 & 0 & 2 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 2 & 2 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 2 & 0 & 6 & 3 & 0 & 3 \\ 0 & 0 & 2 & 0 & 2 & 0 & 3 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 \\ 0 & 0 & 2 & 0 & 2 & 0 & 3 & 0 & 0 & 6 \end{pmatrix}$$

A.2

In this section of appendix, we show that for any given $n \times n$ real symmetric, positive definite covariance matrix (or correlation matrix), there is a uniquely defined member of the Kronecker product model that is closest to the covariance matrix (or correlation matrix) in some sense in terms of the *log parameter* space, once a factorization $n = n_1 \cdots n_v$ is determined.

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

$$\langle A, B \rangle_W := (\operatorname{vech} A)^\intercal W \operatorname{vech} B \qquad A, B \in \mathcal{M}_n.$$

It is easy to show that $\langle \cdot, \cdot \rangle_W$ is an inner product. $\mathcal{M}_n$ with inner product $\langle \cdot, \cdot \rangle_W$ can be identified by $\mathbb{R}^{n(n+1)/2}$ with the usual Euclidean inner product. Since $\mathbb{R}^{n(n+1)/2}$ with the usual Euclidean inner product is a Hilbert space (for finite n), so is $\mathcal{M}_n$. The inner product $\langle \cdot, \cdot \rangle_W$ induces the following norm

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

Let $\mathcal{D}_n$ denote the set of matrices of the form

$$\Omega_1 \otimes I_{n_1} \otimes \cdots \otimes I_{n_v} + I_{n_1} \otimes \Omega_2 \otimes \cdots \otimes I_{n_v} + \cdots + I_{n_1} \otimes \cdots \otimes \Omega_v,$$

where Ω_j are $n_j \times n_j$ real symmetric matrices for $j = 1, \ldots, v$. $\mathcal{D}_n$ is a (linear) subspace of $\mathcal{M}_n$ as, for $\alpha, \beta \in \mathbb{R}$,

$$\begin{aligned}
&\alpha \left(\Omega_1 \otimes I_{n_1} \otimes \cdots \otimes I_{n_v} + I_{n_1} \otimes \Omega_2 \otimes \cdots \otimes I_{n_v} + \cdots + I_{n_1} \otimes \cdots \otimes \Omega_v\right) + \\
&\beta \left(\Xi_1 \otimes I_{n_1} \otimes \cdots \otimes I_{n_v} + I_{n_1} \otimes \Xi_2 \otimes \cdots \otimes I_{n_v} + \cdots + I_{n_1} \otimes \cdots \otimes \Xi_v\right) \\
&= (\alpha\Omega_1 + \beta\Xi_1) \otimes I_{n_1} \otimes \cdots \otimes I_{n_v} + I_{n_1} \otimes (\alpha\Omega_2 + \beta\Xi_2) \otimes \cdots \otimes I_{n_v} + \cdots + I_{n_1} \otimes \cdots \otimes (\alpha\Omega_v + \beta\Xi_v) \\
&\in \mathcal{D}_n.
\end{aligned}$$

For finite n, $\mathcal{D}_n$ is also closed.

Consider a real symmetric, positive definite covariance matrix Σ. $\log \Sigma \in \mathcal{M}_n$. By the projection theorem of the Hilbert space, there exists a unique matrix $L^0 \in \mathcal{D}_n$ such that

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

Note also that $\log \Sigma^{-1} = -\log \Sigma$, so that $-L^0$ simultaneously approximates the precision matrix Σ^{-1} in the same norm.

This says that any real symmetric, positive definite covariance matrix Σ has a closest approximating matrix Σ^0 in a sense that

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

That is, $\log \Sigma^0 = L^0$. Since $L^0 \in \mathcal{D}_n$, we can write

$$L^0 = L_1^0 \otimes I_{n_1} \otimes \cdots \otimes I_{n_v} + I_{n_1} \otimes L_2^0 \otimes \cdots \otimes I_{n_v} + \cdots + I_{n_1} \otimes \cdots \otimes L_v^0,$$

where L_j^0 are $n_j \times n_j$ real symmetric matrices for $j = 1, \ldots, v$. Then

$$\begin{aligned}
\Sigma^0 &= \exp L^0 \\
&= \exp \left[L_1^0 \otimes I_{n_1} \otimes \cdots \otimes I_{n_v} + I_{n_1} \otimes L_2^0 \otimes \cdots \otimes I_{n_v} + \cdots + I_{n_1} \otimes \cdots \otimes L_v^0\right] \\
&= \exp \left[L_1^0 \otimes I_{n_1} \otimes \cdots \otimes I_{n_v}\right] \times \exp \left[I_{n_1} \otimes L_2^0 \otimes \cdots \otimes I_{n_v}\right] \times \cdots \times \exp \left[I_{n_1} \otimes \cdots \otimes L_v^0\right] \\
&= \left[\exp(L_1^0) \otimes I_{n_1} \otimes \cdots \otimes I_{n_v}\right] \times \left[I_{n_1} \otimes \exp(L_2^0) \otimes \cdots \otimes I_{n_v}\right] \times \cdots \times \left[I_{n_1} \otimes \cdots \otimes \exp(L_v^0)\right] \\
&= \exp(L_1^0) \otimes \exp(L_2^0) \otimes \cdots \otimes \exp(L_v^0) \\
&=: \Sigma_1^0 \otimes \cdots \otimes \Sigma_v^0,
\end{aligned}$$

where the third equality is due to Theorem 10.2 in Higham (2008) p235 and the fact that $L_1^0 \otimes I_{n_1} \otimes \cdots \otimes I_{n_v}$ and $I_{n_1} \otimes L_2^0 \otimes \cdots \otimes I_{n_v}$ commute, the fourth equality is due to $f(A) \otimes I = f(A \otimes I)$ for any matrix function f (e.g., Theorem 1.13 in Higham (2008) p10), the fifth equality is due to a property of Kronecker product, and Σ_j^0 is real symmetric, positive definite $n_j \times n_j$ matrix for $j = 1, \ldots, v$.

We thus see that Σ^0 is of the Kronecker product form, and that its precision matrix Σ^{-1} has a closest approximating matrix $(\Sigma^0)^{-1}$. This reasoning provides a justification (i.e., interpretation) for using Σ^0 even when the Kronecker product model is misspecified for the covariance matrix. The same reasoning applies to any real symmetric, positive definite correlation matrix Θ.

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

A.3

Lemma A.1. *Suppose Assumptions 3.1(i) and 3.2(i) hold. Then*

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

Proof. Write $\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^T x_t x_t^\intercal - \bar{x}\bar{x}^\intercal$. We have

$$\|\hat{\Sigma}_T - \Sigma\|_{\ell_2} \leq \left\|\frac{1}{T}\sum_{t=1}^T x_t x_t^\intercal - \mathbb{E}x_t x_t^\intercal\right\|_{\ell_2} + \|\bar{x}\bar{x}^\intercal - \mu\mu^\intercal\|_{\ell_2}. \tag{A.6}$$

We consider the first term on the right hand side of (A.6) first. Invoke Lemma A.4 in Appendix A.5 with $\varepsilon = 1/4$:

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

where $z_{a,t} := x_t^\intercal a$. By Assumption 3.1(i), $\{z_{a,t}\}_{t=1}^T$ 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 $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ be a non-decreasing, convex function with $\psi(0) = 0$ and $\lim_{x\to\infty}\psi(x) = \infty$, where $\mathbb{R}^+$ denotes the set of nonnegative real numbers. 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 A.5 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 Proposition A.8 in Appendix A.5, we have

$$2\max_{a\in\mathcal{N}_{1/4}}\left|\frac{1}{T}\sum_{t=1}^{T}(z_{a,t}^2 - \mathbb{E}z_{a,t}^2)\right| = O_p\left(\frac{\log|\mathcal{N}_{1/4}|}{T} \vee \sqrt{\frac{\log|\mathcal{N}_{1/4}|}{T}}\right).$$

Invoking Lemma A.3 in Appendix A.5, we have $|\mathcal{N}_{1/4}| \leq 9^n$. Thus we have

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

where the last equality is due to Assumption 3.2(i). We now consider the second term on the right hand side of (A.6).

$$\begin{aligned}&\|\bar{x}\bar{x}^\intercal - \mu\mu^\intercal\|_{\ell_2} = \|\bar{x}\bar{x}^\intercal - \mu\bar{x}^\intercal + \mu\bar{x}^\intercal - \mu\mu^\intercal\|_{\ell_2} \leq 2\max_{a\in\mathcal{N}_{1/4}}\left|a^\intercal\left(\bar{x}\bar{x}^\intercal - \mu\bar{x}^\intercal + \mu\bar{x}^\intercal - \mu\mu^\intercal\right)a\right|\\ &= 2\max_{a\in\mathcal{N}_{1/4}}\left|a^\intercal\left((\bar{x}-\mu)\bar{x}^\intercal + \mu(\bar{x}-\mu)^\intercal\right)a\right| \leq 2\max_{a\in\mathcal{N}_{1/4}}\left|a^\intercal(\bar{x}-\mu)\bar{x}^\intercal a\right| + 2\max_{a\in\mathcal{N}_{1/4}}\left|a^\intercal\mu(\bar{x}-\mu)^\intercal a\right|\\ &\leq 2\max_{a\in\mathcal{N}_{1/4}}\left|a^\intercal(\bar{x}-\mu)\right|\max_{a\in\mathcal{N}_{1/4}}\left|\bar{x}^\intercal a\right| + 2\max_{a\in\mathcal{N}_{1/4}}\left|a^\intercal\mu\right|\max_{a\in\mathcal{N}_{1/4}}\left|(\bar{x}-\mu)^\intercal a\right|.\end{aligned}$$

We consider $\max_{a\in\mathcal{N}_{1/4}}\left|(\bar{x}-\mu)^\intercal a\right|$ first.

$$\max_{a\in\mathcal{N}_{1/4}}\left|(\bar{x}-\mu)^\intercal a\right| = \max_{a\in\mathcal{N}_{1/4}}\left|\frac{1}{T}\sum_{t=1}^{T}(x_t^\intercal a - \mathbb{E}[x_t^\intercal a])\right| =: \max_{a\in\mathcal{N}_{1/4}}\left|\frac{1}{T}\sum_{t=1}^{T}(z_{a,t} - \mathbb{E}z_{a,t})\right|.$$

By Assumption 3.1(i), $\{z_{a,t}\}_{t=1}^T$ are independent subgaussian random variables. For $\epsilon > 0$, $\mathbb{P}(|z_{a,t}| \geq \epsilon) \leq Ke^{-C\epsilon^2}$. It follows from Lemma 2.2.1 in van der Vaart and Wellner (1996) that $\|z_{a,t}\|_{\psi_2} \leq (1+K)^{1/2}/C^{1/2}$. Then $\|z_{a,t} - \mathbb{E}z_{a,t}\|_{\psi_2} \leq \|z_{a,t}\|_{\psi_2} + \mathbb{E}\|z_{a,t}\|_{\psi_2} \leq \frac{2(1+K)^{1/2}}{C^{1/2}}$. Next, using the second last inequality in van der Vaart and Wellner (1996) p95, we have

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

Then, by the definition of the Orlicz norm, $\mathbb{E}\left[e^{W|z_{a,t}-\mathbb{E}z_{a,t}|}\right] \leq 2$. Use Fubini's theorem to expand out the exponential moment. It is easy to see that $z_{a,t} - \mathbb{E}z_{a,t}$ satisfies the moment conditions of Bernstein's inequality in Appendix A.5 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 Proposition A.8 in Appendix A.5, we have

$$\max_{a\in\mathcal{N}_{1/4}}\left|\frac{1}{T}\sum_{t=1}^{T}(z_{a,t} - \mathbb{E}z_{a,t})\right| = O_p\left(\frac{\log|\mathcal{N}_{1/4}|}{T} \vee \sqrt{\frac{\log|\mathcal{N}_{1/4}|}{T}}\right).$$

Invoking Lemma A.3 in Appendix A.5, we have $|\mathcal{N}_{1/4}| \leq 9^n$. Thus we have

$$\max_{a\in\mathcal{N}_{1/4}} \left|(\bar{x}-\mu)^\intercal a\right| = O_p\left(\frac{n}{T} \vee \sqrt{\frac{n}{T}}\right) = O_p\left(\sqrt{\frac{n}{T}}\right), \tag{A.7}$$

where the last equality is due to Assumption 3.2(i). Now Let's consider $\max_{a\in\mathcal{N}_{1/4}}\left|a^\intercal\mu\right|$.

$$\begin{aligned}\max_{a\in\mathcal{N}_{1/4}} \left|a^\intercal\mu\right| &=: \max_{a\in\mathcal{N}_{1/4}} \left|\mathbb{E}a^\intercal x_t\right| = \max_{a\in\mathcal{N}_{1/4}} \left|\mathbb{E}z_{a,t}\right| \leq \max_{a\in\mathcal{N}_{1/4}} \mathbb{E}|z_{a,t}| = \max_{a\in\mathcal{N}_{1/4}} \|z_{a,t}\|_{L_1} \\ &\leq \max_{a\in\mathcal{N}_{1/4}} \|z_{a,t}\|_{\psi_1} \leq \max_{a\in\mathcal{N}_{1/4}} \|z_{a,t}\|_{\psi_2}(\log 2)^{-1/2} \leq \frac{(1+K)^{1/2}}{C^{1/2}}(\log 2)^{-1/2},\end{aligned}$$

where $\|\cdot\|_{L_1}$ is the L_1 norm, the second and third inequalities are from van der Vaart and Wellner (1996) p95. Thus we have

$$\max_{a\in\mathcal{N}_{1/4}} |a^\intercal\mu| = O(1). \tag{A.8}$$

Next we consider $\max_{a\in\mathcal{N}_{1/4}}\left|a^\intercal\bar{x}\right|$.

$$\begin{aligned}\max_{a\in\mathcal{N}_{1/4}} \left|a^\intercal\bar{x}\right| &= \max_{a\in\mathcal{N}_{1/4}} \left|a^\intercal(\bar{x}-\mu+\mu)\right| \leq \max_{a\in\mathcal{N}_{1/4}} \left|a^\intercal(\bar{x}-\mu)\right| + \max_{a\in\mathcal{N}_{1/4}} \left|a^\intercal\mu\right| = O_p\left(\sqrt{\frac{n}{T}}\right) + O(1) \\ &= O_p(1),\end{aligned} \tag{A.9}$$

where the last equality is due to Assumption 3.2(i). Combining (A.7), (A.8) and (A.9), we have

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

□

Proposition A.3. *Suppose Assumptions 3.1(i), 3.2(i) and 3.3(i) hold. Then*

(i)

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

(ii) The minimum eigenvalue of D is bounded away from zero by an absolute positive constant (i.e., $\|D^{-1}\|_{\ell_2} = O(1)$), so is the minimum eigenvalue of $D^{1/2}$ (i.e., $\|D^{-1/2}\|_{\ell_2} = O(1)$).

(iii)

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

(iv)

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

(v)

$$\|\hat{D}_T^{-1/2}\|_{\ell_2} = O_p(1).$$

(vi) The maximum eigenvalue of Σ is bounded from the above by an absolute constant (i.e., $\|\Sigma\|_{\ell_2} = O(1)$). The maximum eigenvalue of D is bounded from the above by an absolute constant (i.e., $\|D\|_{\ell_2} = O(1)$).

(vii)

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

Proof. Define $\sigma_i^2 := \mathbb{E}(x_{t,i} - \sigma_i)^2$ and $\hat{\sigma}_i^2 := \frac{1}{T}\sum_{t=1}^T (x_{t,i} - \bar{x}_i)^2$, where the subscript i denotes the ith component of the corresponding vector. For part (i),

$$\begin{aligned}\|\hat{D}_T - D\|_{\ell_2} &= \max_{1\leq i\leq n} |\hat{\sigma}_i^2 - \sigma_i^2| = \max_{1\leq i\leq n} |e_i^\intercal(\hat{\Sigma}_T - \Sigma)e_i| \leq \max_{\|a\|_2=1} |a^\intercal(\hat{\Sigma}_T - \Sigma)a| \\ &= \|\hat{\Sigma}_T - \Sigma\|_{\ell_2},\end{aligned}$$

where e_i denotes a unit vector whose ith component is 1. Now invoke Lemma A.1 to get the result. For part (ii),

$$\text{mineval}(D) = \min_{1\leq i\leq n} \sigma_i^2 = \min_{1\leq i\leq n} e_i^\intercal \Sigma e_i \geq \min_{\|a\|_2=1} a^\intercal \Sigma a = \text{mineval}(\Sigma) > 0$$

where the last inequality is due to Assumption 3.3. For part (iii), invoking Lemma A.5 in Appendix A.5 gives

$$\|\hat{D}_T^{1/2} - D^{1/2}\|_{\ell_2} \leq \frac{\|\hat{D}_T - D\|_{\ell_2}}{\text{mineval}(\hat{D}_T^{1/2}) + \text{mineval}(D^{1/2})} = O_p(1)\|\hat{D}_T - D\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right),$$

where the first and second equalities are due to parts (ii) and (i), respectively. Part (iv) follows from Lemma A.6 in Appendix A.5 via parts (ii) and (iii). For part (v),

$$\begin{aligned}\|\hat{D}_T^{-1/2}\|_{\ell_2} &= \|\hat{D}_T^{-1/2} - D^{-1/2} + D^{-1/2}\|_{\ell_2} \leq \|\hat{D}_T^{-1/2} - D^{-1/2}\|_{\ell_2} + \|D^{-1/2}\|_{\ell_2} \\ &= O_p\left(\sqrt{\frac{n}{T}}\right) + O(1) = O_p(1).\end{aligned}$$

For part (vi), we have

$$\|\Sigma\|_{\ell_2} = \max_{\|a\|_2=1} \left|a^\intercal \left(\mathbb{E}[x_t x_t^\intercal] - \mu\mu^\intercal\right) a\right| \leq \max_{\|a\|_2=1} \mathbb{E}z_{a,t}^2 + \max_{\|a\|_2=1} (\mathbb{E}z_{a,t})^2 \leq 2\max_{\|a\|_2=1} \mathbb{E}z_{a,t}^2.$$

We have shown that in the proof of Lemma A.1 that $\|z_{a,t}^2\|_{\psi_1} \leq \frac{1+K}{C}$ for any $\|a\|_2 = 1$. This says that $z_{a,t}^2$ has bounded exponential moments, so the result follows. Next we consider

$$\|D\|_{\ell_2} = \max_{1\leq i\leq n} \sigma_i^2 = \max_{1\leq i\leq n} e_i^\intercal \Sigma e_i \leq \max_{\|a\|_2=1} a^\intercal \Sigma a = \text{maxeval}(\Sigma) < \infty.$$

For part (vii),

$$\begin{aligned}&\|\hat{D}_T^{-1/2} \otimes \hat{D}_T^{-1/2} - D^{-1/2} \otimes D^{-1/2}\|_{\ell_2} \\ &= \|\hat{D}_T^{-1/2} \otimes \hat{D}_T^{-1/2} - \hat{D}_T^{-1/2} \otimes D^{-1/2} + \hat{D}_T^{-1/2} \otimes D^{-1/2} - D^{-1/2} \otimes D^{-1/2}\|_{\ell_2} \\ &\leq \|\hat{D}_T^{-1/2} \otimes (\hat{D}_T^{-1/2} - D^{-1/2})\|_{\ell_2} + \|(\hat{D}_T^{-1/2} - D^{-1/2}) \otimes D^{-1/2}\|_{\ell_2} \\ &= \left(\|\hat{D}_T^{-1/2}\|_{\ell_2} + \|D^{-1/2}\|_{\ell_2}\right) \|\hat{D}_T^{-1/2} - D^{-1/2}\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right),\end{aligned}$$

where the second equality is due to Proposition A.10 in Appendix A.5. □

Proof of Proposition 3.1(i). Recall that

$$\hat{\Theta}_T = \hat{D}_T^{-1/2}\hat{\Sigma}_T\hat{D}_T^{-1/2}, \qquad \Theta = D^{-1/2}\Sigma D^{-1/2}.$$

Then we have

$$\begin{aligned}\|\hat{\Theta}_T - \Theta\|_{\ell_2} &= \|\hat{D}_T^{-1/2}\hat{\Sigma}_T\hat{D}_T^{-1/2} - \hat{D}_T^{-1/2}\Sigma\hat{D}_T^{-1/2} + \hat{D}_T^{-1/2}\Sigma\hat{D}_T^{-1/2} - D^{-1/2}\Sigma D^{-1/2}\|_{\ell_2} \\ &\leq \|\hat{D}_T^{-1/2}\|_{\ell_2}^2\|\hat{\Sigma}_T - \Sigma\|_{\ell_2} + \|\hat{D}_T^{-1/2}\Sigma\hat{D}_T^{-1/2} - D^{-1/2}\Sigma D^{-1/2}\|_{\ell_2}. \end{aligned} \tag{A.10}$$

Invoking Lemma A.1 and Proposition A.3(v), we conclude that the first term of (A.10) is $O_p(\sqrt{n/T})$. Let's consider the second term of (A.10). Write

$$\begin{aligned}&\|\hat{D}_T^{-1/2}\Sigma\hat{D}_T^{-1/2} - D^{-1/2}\Sigma\hat{D}_T^{-1/2} + D^{-1/2}\Sigma\hat{D}_T^{-1/2} - D^{-1/2}\Sigma D^{-1/2}\|_{\ell_2} \\ &\leq \|(\hat{D}_T^{-1/2} - D^{-1/2})\Sigma\hat{D}_T^{-1/2}\|_{\ell_2} + \|D^{-1/2}\Sigma(\hat{D}_T^{-1/2} - D^{-1/2})\|_{\ell_2} \\ &\leq \|\hat{D}_T^{-1/2}\|_{\ell_2}\|\Sigma\|_{\ell_2}\|\hat{D}_T^{-1/2} - D^{-1/2}\|_{\ell_2} + \|D^{-1/2}\|_{\ell_2}\|\Sigma\|_{\ell_2}\|\hat{D}_T^{-1/2} - D^{-1/2}\|_{\ell_2}.\end{aligned}$$

Invoking Proposition A.3(ii), (iv), (v) and (vi), we conclude that the second term of (A.10) is $O_p(\sqrt{n/T})$. □

To prove part (ii) of Proposition 3.1, we shall use Lemma 4.1 of Gil' (2012). That lemma will further simplify when we consider real symmetric, positive definite matrices. For the ease of reference, we state this simplified version of Lemma 4.1 of Gil' (2012) here.

Lemma A.2 (Simplified from Lemma 4.1 of Gil' (2012))**.** *For $n \times n$ real symmetric, positive definite matrices A, B, if*

$$\|A - B\|_{\ell_2} < a,$$

for some absolute constant $a > 1$, then

$$\|\log A - \log B\|_{\ell_2} \leq C\|A - B\|_{\ell_2},$$

for some positive absolute constant C.

Proof. First note that for any real symmetric, positive definite matrix A, $p(A, x) = x$ for any $x > 0$ in Lemma 4.1 of Gil' (2012). Since A is real symmetric and positive definite, all its eigenvalues lie in the region $|\arg(z - a)| \leq \pi/2$. Then according to Gil' (2012) p11, we have for any $t \geq 0$

$$\begin{aligned}\rho(A, -t) &\geq (a + t)\sin(\pi/2) = a + t \\ \rho(A, -t) - \delta &\geq a + t - \delta,\end{aligned}$$

where

$$\delta := \begin{cases} \|A - B\|_{\ell_2}^{1/n} & \text{if } \|A - B\|_{\ell_2} \leq 1 \\ \|A - B\|_{\ell_2} & \text{if } \|A - B\|_{\ell_2} \geq 1 \end{cases}$$

and $\rho(A, -t)$ is defined in Gil' (2012). Then the condition of Lemma A.2 allows one to invoke Lemma 4.1 of Gil' (2012) as

$$\rho(A, -t) \geq a + t \geq a > \delta.$$

Lemma 4.1 of Gil' (2012) says

$$\begin{aligned}&\|\log A - \log B\|_{\ell_2} \leq \|A - B\|_{\ell_2}\int_0^\infty p\left(A, \frac{1}{\rho(A, -t)}\right) p\left(B, \frac{1}{\rho(A, -t) - \delta}\right) dt \\ &= \|A - B\|_{\ell_2}\int_0^\infty \frac{1}{\rho(A, -t)}\frac{1}{\rho(A, -t) - \delta}dt \leq \|A - B\|_{\ell_2}\int_0^\infty \frac{1}{(a + t)(a + t - \delta)}dt \\ &\leq \|A - B\|_{\ell_2}\int_0^\infty \frac{1}{(a + t - \delta)^2}dt = \|A - B\|_{\ell_2}\frac{1}{a - \delta} =: C\|A - B\|_{\ell_2}.\end{aligned}$$

□

Proof of Proposition 3.1(ii). It follows trivially from Lemma A.2. □

Proof of Proposition 3.1(iii). We have

$$\begin{aligned}
\|\hat{\theta}_T - \theta^0\|_2 &= \|(E^\intercal W E)^{-1} E^\intercal W\|_{\ell_2} \|D_n^+\|_{\ell_2} \|\log \hat{\Theta}_T - \log \Theta\|_F \\
&\leq \|(E^\intercal W E)^{-1} E^\intercal W\|_{\ell_2} \sqrt{n} \|\log \hat{\Theta}_T - \log \Theta\|_{\ell_2} = O(\sqrt{\varpi\kappa(W)/n})\sqrt{n} O_p(\sqrt{n/T}) \\
&= O_p\left(\sqrt{\frac{n\varpi\kappa(W)}{T}}\right),
\end{aligned}$$

where the first inequality is due to (A.13), and the second equality is due to (A.19) and parts (i)-(ii) of this proposition. □

A.4

The following proposition linearizes the matrix logarithm.

Proposition A.4. *Suppose both $n \times n$ matrices $A+B$ and A are real, symmetric, and positive definite for all n with the minimum eigenvalues bounded away from zero by absolute constants. Suppose the maximum eigenvalue of A is bounded from above by an absolute constant. Further suppose*

$$\left\|[t(A-I)+I]^{-1} tB\right\|_{\ell_2} \leq C < 1 \tag{A.11}$$

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

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

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 A.4 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.

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 A.2 in Appendix A.5, 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 (A.11) to invoke Proposition A.9 in Appendix A.5 to expand $[t(A-I)+I+tB]^{-1}$ to get

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

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

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

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

Proposition A.5. *Suppose Assumptions 3.1(i), 3.2(i) and 3.3 hold.*

(i) *Θ has minimum eigenvalue bounded away from zero by an absolute constant and maximum eigenvalue bounded from above by an absolute constant.*

(ii) *$\hat{\Theta}_T$ has minimum eigenvalue bounded away from zero by an absolute constant and maximum eigenvalue bounded from above by an absolute constant with probability approaching 1.*

Proof. For part (i), the maximum eigenvalue of Θ is its spectral norm, i.e., $\|\Theta\|_{\ell_2}$.

$$\|\Theta\|_{\ell_2} = \|D^{-1/2}\Sigma D^{-1/2}\|_{\ell_2} \leq \|D^{-1/2}\|_{\ell_2}^2 \|\Sigma\|_{\ell_2} < C,$$

where the last inequality is due to Proposition A.3(ii) and (vi). Now let's consider the minimum eigenvalue of Θ.

$$\text{mineval}(\Theta) = \text{mineval}(D^{-1/2}\Sigma D^{-1/2}) = \min_{\|a\|_2=1} a^\intercal D^{-1/2}\Sigma D^{-1/2} a \geq \min_{\|a\|_2=1} \text{mineval}(\Sigma)\|D^{-1/2}a\|_2^2$$
$$= \text{mineval}(\Sigma) \min_{\|a\|_2=1} a^\intercal D^{-1} a = \text{mineval}(\Sigma)\text{mineval}(D^{-1}) = \frac{\text{mineval}(\Sigma)}{\text{maxeval}(D)} > 0,$$

where the second equality is due to Rayleigh-Ritz theorem, and the last inequality is due to Assumption 3.3 and Proposition A.3(vi). For part (ii), the maximum eigenvalue of $\hat{\Theta}$ is its spectral norm, i.e., $\|\hat{\Theta}\|_{\ell_2}$.

$$\|\hat{\Theta}_T\|_{\ell_2} \leq \|\hat{\Theta}_T - \Theta\|_{\ell_2} + \|\Theta\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right) + \|\Theta\|_{\ell_2} = O_p(1)$$

where the first equality is due to Proposition 3.1(i) and the last equality is due to part (i). The minimum eigenvalue of $\hat{\Theta}_T$ is $1/\text{maxeval}(\hat{\Theta}_T^{-1})$. Since $\|\Theta^{-1}\|_{\ell_2} = \text{maxeval}(\Theta^{-1}) = 1/\text{mineval}(\Theta) = O(1)$ by part (i) and $\|\hat{\Theta}_T - \Theta\|_{\ell_2} = O_p(\sqrt{n/T})$ by Proposition 3.1(i), we can invoke Lemma A.6 in Appendix A.5 to get

$$\|\hat{\Theta}_T^{-1} - \Theta^{-1}\|_{\ell_2} = O_p(\sqrt{n/T}),$$

whence we have

$$\|\hat{\Theta}_T^{-1}\|_{\ell_2} \leq \|\hat{\Theta}_T^{-1} - \Theta^{-1}\|_{\ell_2} + \|\Theta^{-1}\|_{\ell_2} = O_p(1).$$

Thus the minimum eigenvalue of $\hat{\Theta}_T$ is bounded away from zero by an absolute constant. $\square$

Define

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

The following proposition gives the rate of convergence for $\hat{H}_T$. The following proposition is also true when one replaces $\hat{H}_T$ with $\hat{H}_{T,D}$.

Proposition A.6. *Let Assumptions 3.1(i), 3.2(i) and 3.3 be satisfied. Then we have*

$$\|H\|_{\ell_2} = O(1), \qquad \|\hat{H}_T\|_{\ell_2} = O_p(1), \qquad \|\hat{H}_T - H\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right). \tag{A.12}$$

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

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

where the second equality is due to Proposition A.10 in Appendix A.5, and the last equality is due to Proposition A.5(ii). Now,

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

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 A.10 in Appendix A.5, and the last equality is because Proposition A.5 implies

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

Now

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

where the last equality is due to Proposition 3.1(i). The proposition then follows after invoking Lemma A.6 in Appendix A.5. □

Proposition A.7. *Given the $n^2 \times n(n+1)/2$ duplication matrix D_n and its Moore-Penrose generalised inverse $D_n^+ = (D_n^\intercal D_n)^{-1} D_n^\intercal$ (i.e., D_n is full-column rank), we have*

$$\|D_n^+\|_{\ell_2} = \|D_n^{+\intercal}\|_{\ell_2} = 1, \qquad \|D_n\|_{\ell_2} = \|D_n^\intercal\|_{\ell_2} = 2. \tag{A.13}$$

Proof. First note that $D_n^\intercal D_n$ is a diagonal matrix with diagonal entries either 1 or 2. Using the fact that for any matrix A, $AA^\intercal$ and $A^\intercal A$ have the same non-zero eigenvalues, we have

$$\begin{aligned}
\|D_n^{+\intercal}\|_{\ell_2}^2 &= \text{maxeval}(D_n^+ D_n^{+\intercal}) = \text{maxeval}((D_n^\intercal D_n)^{-1}) = 1 \\
\|D_n^+\|_{\ell_2}^2 &= \text{maxeval}(D_n^{+\intercal} D_n^+) = \text{maxeval}(D_n^+ D_n^{+\intercal}) = \text{maxeval}((D_n^\intercal D_n)^{-1}) = 1 \\
\|D_n\|_{\ell_2}^2 &= \text{maxeval}(D_n^\intercal D_n) = 2 \\
\|D_n^\intercal\|_{\ell_2}^2 &= \text{maxeval}(D_n D_n^\intercal) = \text{maxeval}(D_n^\intercal D_n) = 2
\end{aligned}$$

□

Proof of Theorem 3.1. We first show that (A.11) is satisfied with probability approaching 1 for $A = \Theta$ and $B = \hat{\Theta}_T - \Theta$. That is,

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

for some constant C.

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

where the first equality is due to Proposition 3.1(i), and the last equality is due to mineval$(t(\Theta - I) + I) > C > 0$ for some absolute constant C (implied by Proposition A.5(i)) and Assumption 3.2(i). Together with Proposition A.5(ii) and Lemma 2.12 in van der Vaart (1998), we can invoke Proposition A.4 stochastically with $A = \Theta$ and $B = \hat{\Theta}_T - \Theta$:

$$\log \hat{\Theta}_T - \log \Theta = \int_0^1 [t(\Theta - I) + I]^{-1} (\hat{\Theta}_T - \Theta)[t(\Theta - I) + I]^{-1} dt + O_p(\|\hat{\Theta}_T - \Theta\|_{\ell_2}^2). \quad \text{(A.14)}$$

(We can invoke Proposition A.4 stochastically because the remainder of the log linearization is zero when the perturbation is zero. Moreover, we have $\|\hat{\Theta}_T - \Theta\|_{\ell_2} \xrightarrow{p} 0$ under Assumption 3.2(i).) Note that (A.14) also holds with $\hat{\Theta}_T$ replaced by $\hat{\Theta}_{T,D}$ by repeating the same argument. That is,

$$\log \hat{\Theta}_{T,D} - \log \Theta = \int_0^1 [t(\Theta - I) + I]^{-1} (\hat{\Theta}_{T,D} - \Theta)[t(\Theta - I) + I]^{-1} dt + O_p(\|\hat{\Theta}_{T,D} - \Theta\|_{\ell_2}^2).$$

Now we can write

$$\begin{aligned}
\frac{\sqrt{T} c^\intercal (\hat{\theta}_{T,D} - \theta^0)}{\sqrt{\hat{G}_{T,D}}} &= \frac{\sqrt{T} c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H (D^{-1/2} \otimes D^{-1/2}) \operatorname{vec}(\hat{\Sigma}_T - \Sigma)}{\sqrt{\hat{G}_{T,D}}} \\
&\quad + \frac{\sqrt{T} c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ \operatorname{vec} O_p(\|\hat{\Theta}_{T,D} - \Theta\|_{\ell_2}^2)}{\sqrt{\hat{G}_{T,D}}} \\
&=: \hat{t}_{D,1} + \hat{t}_{D,2}.
\end{aligned}$$

Define

$$t_{D,1} := \frac{\sqrt{T} c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H (D^{-1/2} \otimes D^{-1/2}) \operatorname{vec}(\tilde{\Sigma}_T - \Sigma)}{\sqrt{G_D}}.$$

To prove Theorem 3.1, it suffices to show $t_{D,1} \xrightarrow{d} N(0,1)$, $t_{D,1} - \hat{t}_{D,1} = o_p(1)$, and $\hat{t}_{D,2} = o_p(1)$.

A.4.1 $t_{D,1} \xrightarrow{d} N(0,1)$

We now prove that $t_{D,1}$ is asymptotically distributed as a standard normal.

$$\begin{aligned}
t_{D,1} &= \frac{\sqrt{T} c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H (D^{-1/2} \otimes D^{-1/2}) \operatorname{vec}\left(\frac{1}{T}\sum_{t=1}^T \left[(x_t - \mu)(x_t - \mu)^\intercal - \mathbb{E}(x_t - \mu)(x_t - \mu)^\intercal\right]\right)}{\sqrt{G_D}} \\
&= \sum_{t=1}^T \frac{T^{-1/2} c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H (D^{-1/2} \otimes D^{-1/2}) \operatorname{vec}\left[(x_t - \mu)(x_t - \mu)^\intercal - \mathbb{E}(x_t - \mu)(x_t - \mu)^\intercal\right]}{\sqrt{G_D}} \\
&=: \sum_{t=1}^T U_{D,T,n,t}.
\end{aligned}$$

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

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

For any $\gamma > 2$,

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

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

$$\begin{aligned}
G_D &= c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H (D^{-1/2} \otimes D^{-1/2}) V (D^{-1/2} \otimes D^{-1/2}) H D_n^{+\intercal} W E (E^\intercal W E)^{-1} c \\
&\geq \text{mineval}(V)\text{mineval}(D^{-1} \otimes D^{-1})\text{mineval}(H^2)\text{mineval}(D_n^+ D_n^{+\intercal})\text{mineval}(W)\text{mineval}((E^\intercal W E)^{-1}) \\
&= \frac{\text{mineval}(V)\text{mineval}^2(H)}{\text{maxeval}(D \otimes D)\text{maxeval}(D_n^\intercal D_n)\text{maxeval}(W^{-1})\text{maxeval}(E^\intercal W E)} \\
&\geq \frac{\text{mineval}(V)\text{mineval}^2(H)}{\text{maxeval}(D \otimes D)\text{maxeval}(D_n^\intercal D_n)\text{maxeval}(W^{-1})\text{maxeval}(W)\text{maxeval}(E^\intercal E)}
\end{aligned}$$

where the first and third inequalities are true by repeatedly invoking the Rayleigh-Ritz theorem. Note that

$$\text{maxeval}(E^\intercal E) \leq \text{tr}(E^\intercal E) \leq s \cdot n, \tag{A.15}$$

where the last inequality is due to Proposition A.2. For future reference

$$\|E\|_{\ell_2} = \|E^\intercal\|_{\ell_2} = \sqrt{\text{maxeval}(E^\intercal E)} \leq \sqrt{sn}. \tag{A.16}$$

Since the minimum eigenvalue of H is bounded away from zero by an absolute constant by Proposition A.5(i), the maximum eigenvalue of D is bounded from above by an absolute constant (Proposition A.3(vi)), and $\text{maxeval}[D_n^\intercal D_n]$ is bounded from above since $D_n^\intercal D_n$ is a diagonal matrix with diagonal entries either 1 or 2, we have

$$\frac{1}{\sqrt{G_D}} = O(\sqrt{s \cdot n \cdot \kappa(W)}). \tag{A.17}$$

Then a sufficient condition for the Lindeberg condition is:

$$\begin{aligned}
&T^{1-\frac{\gamma}{2}} (sn\kappa(W))^{\gamma/2} \\
&\cdot \mathbb{E} \left| c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H (D^{-1/2} \otimes D^{-1/2}) \,\text{vec}\left[(x_t - \mu)(x_t - \mu)^\intercal - \mathbb{E}(x_t - \mu)(x_t - \mu)^\intercal\right] \right|^{\gamma} \\
&= o(1),
\end{aligned} \tag{A.18}$$

for some $\gamma > 2$. Note that

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

where the second equality is due to the fact that for any matrix A, $AA^\intercal$ and $A^\intercal A$ have the same non-zero eigenvalues, the third equality is due to $(A^\intercal)^{-1} = (A^{-1})^\intercal$, and the last equality is due to Assumption 3.3(ii). Thus

$$\|(E^\intercal WE)^{-1}E^\intercal W\|_{\ell_2} = O(\sqrt{\varpi\kappa(W)/n}). \tag{A.19}$$

We now verify (A.18).

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

where the first equality is because of (A.19), (A.12), and Proposition A.3(ii), the third inequality is due to Loeve's c_r inequality, the fourth inequality is due to Jensen's inequality, and the last equality is due to the definition of L_p norm. By Assumption 3.1(i), for any $i, j = 1, \dots, n$,

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

It follows from Lemma 2.2.1 in van der Vaart and Wellner (1996) that $\|x_{t,i}x_{t,j}\|_{\psi_1}\le(1+2K)/C$. Similarly we have $\mathbb{P}(|x_{t,i}|\ge\epsilon)\le Ke^{-C\epsilon^2}$, so $\|x_{t,i}\|_{\psi_1}\le\|x_{t,i}\|_{\psi_2}(\log 2)^{-1/2}\le\left[\frac{1+K}{C}\right]^{1/2}(\log 2)^{-1/2}$. Next,

$$\max_{1\le i\le n}|\mu_i| = \max_{1\le i\le n}|\mathbb{E}x_{t,i}|\le\max_{1\le i\le n}\mathbb{E}|x_{t,i}| = \max_{1\le i\le n}\|x_{t,i}\|_{L_1}\le\max_{1\le i\le n}\|x_{t,i}\|_{\psi_1} = O(1). \tag{A.20}$$

Then we have

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

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

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

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

$$\left\|\max_{1\le i,j\le n}\left|(x_t-\mu)_i(x_t-\mu)_j\right|\right\|_{L_\gamma}^\gamma\le(\gamma!)^\gamma\left\|\max_{1\le i,j\le n}\left|(x_t-\mu)_i(x_t-\mu)_j\right|\right\|_{\psi_1}^\gamma = O(\log^\gamma n). \tag{A.21}$$

Summing up the rates, we have

$$\begin{aligned}
&T^{1-\frac{\gamma}{2}}(sn\kappa(W))^{\gamma/2}\\
&\cdot\mathbb{E}\left|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal-\mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right|^\gamma\\
&= T^{1-\frac{\gamma}{2}}(sn\kappa(W))^{\gamma/2}(\varpi\kappa(W)n)^{\gamma/2}O(\log^\gamma n) = O\left(\frac{n^2\cdot\kappa^2(W)\cdot\log^3 n\cdot\varpi}{T^{1-\frac{2}{\gamma}}}\right)^{\gamma/2} = o(1)
\end{aligned}$$

by Assumption 3.2(ii). Thus, we have verified (A.18).

A.4.2 $t_{D,1} - \hat{t}_{D,1} = o_p(1)$

We now show that $t_{D,1} - \hat{t}_{D,1} = o_p(1)$. Let A_D and $\hat{A}_D$ denote the numerators of $t_{D,1}$ and $\hat{t}_{D,1}$, respectively.

$$t_{D,1} - \hat{t}_{D,1} = \frac{A_D}{\sqrt{G_D}} - \frac{\hat{A}_D}{\sqrt{\hat{G}_{T,D}}} = \frac{\sqrt{sn\kappa(W)}A_D}{\sqrt{sn\kappa(W)G_D}} - \frac{\sqrt{sn\kappa(W)}\hat{A}_D}{\sqrt{sn\kappa(W)\hat{G}_{T,D}}}.$$

Since we have already shown in (A.17) that $sn\kappa(W)G_D$ is bounded away from zero by an absolute constant, it suffices to show the denominators as well as numerators of $t_{D,1}$ and $\hat{t}_{D,1}$ are asymptotically equivalent.

A.4.3 Denominators of $t_{D,1}$ and $\hat{t}_{D,1}$

We first show that the denominators of $t_{D,1}$ and $\hat{t}_{D,1}$ are asymptotically equivalent, i.e.,

$$sn\kappa(W)|\hat{G}_{T,D} - G_D| = o_p(1).$$

Define

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

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

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

where $\|\cdot\|_\infty$ denotes the absolute elementwise maximum, the third equality is due to Proposition A.3(ii), Proposition A.10 in Appendix A.5, (A.12), (A.19), and (A.13), the second last equality is due to Proposition 8.2 in SM 8.2, and the last equality is due to Assumption 3.2(ii). We now prove $sn\kappa(W)|\tilde{G}_{T,D} - G_D| = o_p(1)$.

$$\begin{aligned}
&sn\kappa(W)|\tilde{G}_{T,D} - G_D| \\
&= sn\kappa(W)|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_{T,D}(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})\hat{H}_{T,D}D_n^{+\intercal}WE(E^\intercal WE)^{-1}c \\
&\qquad - c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})HD_n^{+\intercal}WE(E^\intercal WE)^{-1}c| \\
&\leq sn\kappa(W)\left|\text{maxeval}\left[(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})\right]\right|^2\|(\hat{H}_{T,D} - H)D_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2^2 \\
&\quad + 2sn\kappa(W)\|(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})HD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2 \\
&\qquad \cdot \|(\hat{H}_{T,D} - H)D_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2
\end{aligned} \tag{A.22}$$

where the inequality is due to Lemma A.7 in Appendix A.5. We consider the first term of (A.22) first.

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

where the second last equality is due to (A.12), (A.13), and (A.19), and the last equality is due to Assumption 3.2(ii). We now consider the second term of (A.22).

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

where the first equality is due to (A.12), (A.13), and (A.19), and the last equality is due to Assumption 3.2(ii). We have proved $|sn\kappa(W)\tilde{G}_{T,D}-sn\kappa(W)G_D|=o_p(1)$ and hence $|sn\kappa(W)\hat{G}_{T,D}-sn\kappa(W)G_D|=o_p(1)$.

A.4.4 Numerators of $t_{D,1}$ and $\hat{t}_{D,1}$

We now show that numerators of $t_{D,1}$ and $\hat{t}_{D,1}$ are asymptotically equivalent, i.e.,

$$
\sqrt{sn\kappa(W)}|A_D-\hat{A}_D|=o_p(1).
$$

This is relatively straight forward.

$$
\begin{aligned}
&\sqrt{Tsn\kappa(W)}\left|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}(\hat{\Sigma}_T-\Sigma-\tilde{\Sigma}_T+\Sigma)\right|\\
&=\sqrt{Tsn\kappa(W)}\left|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}(\hat{\Sigma}_T-\tilde{\Sigma}_T)\right|\\
&=\sqrt{Tsn\kappa(W)}\left|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}\left[(\bar{x}-\mu)(\bar{x}-\mu)^\intercal\right]\right|\\
&\leq\sqrt{Tsn\kappa(W)}\|(E^\intercal WE)^{-1}E^\intercal W\|_{\ell_2}\|D_n^+\|_{\ell_2}\|H\|_{\ell_2}\|D^{-1/2}\otimes D^{-1/2}\|_{\ell_2}\|\operatorname{vec}\left[(\bar{x}-\mu)(\bar{x}-\mu)^\intercal\right]\|_2\\
&=O(\sqrt{Tsn\kappa(W)})\sqrt{\varpi\kappa(W)/n}\|(\bar{x}-\mu)(\bar{x}-\mu)^\intercal\|_F\\
&\leq O(\sqrt{Tsn\kappa(W)})\sqrt{\varpi\kappa(W)/n}n\|(\bar{x}-\mu)(\bar{x}-\mu)^\intercal\|_\infty\\
&=O(\sqrt{Tsn^2\kappa^2(W)\varpi})\max_{1\leq i,j\leq n}\left|(\bar{x}-\mu)_i(\bar{x}-\mu)_j\right|=O_p(\sqrt{Tsn^2\kappa^2(W)\varpi})\log n/T\\
&=O_p\left(\sqrt{\frac{\log^3 n\cdot n^2\kappa^2(W)\varpi}{T}}\right)=o_p(1),
\end{aligned}
$$

where the third equality is due to (A.12), (A.13), and (A.19), the third last equality is due to (8.20), and the last equality is due to Assumption 3.2(ii).

A.4.5 $\hat{t}_{D,2}=o_p(1)$

Write

$$
\hat{t}_{D,2}=\frac{\sqrt{T}\sqrt{sn\kappa(W)}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\operatorname{vec}O_p(\|\hat{\Theta}_{T,D}-\Theta\|_{\ell_2}^2)}{\sqrt{sn\kappa(W)\hat{G}_{T,D}}}.
$$

Since the denominator of the preceding equation is bounded away from zero by an absolute constant with probability approaching one by (A.17) and that $|sn\kappa(W)\hat{G}_{T,D}-sn\kappa(W)G_D|=o_p(1)$, it suffices to show

$$
\sqrt{T}\sqrt{sn\kappa(W)}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\operatorname{vec}O_p(\|\hat{\Theta}_{T,D}-\Theta\|_{\ell_2}^2)=o_p(1).
$$

This is straightforward:

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

where the last equality is due to Assumption 3.2(ii). □

A.5

Definition A.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 A.3. *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: $\|A\|_{\ell_2}=\max_{\|x\|_2=1}|x^{\intercal}Ax|$.

Lemma A.4. *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*

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

Proof. See Vershynin (2011) p8. □

Theorem A.1 (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. $\square$

We can use Bernstein's inequality to establish a rate for the maximum.

Proposition A.8. *Suppose via Bernstein's inequality that we have for* $1 \le i \le n$

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Z_{t,i}\right| \ge \sigma_0^2\left[K\epsilon + \sqrt{2\epsilon}\right]\right) \le 2e^{-T\sigma_0^2\epsilon}.$$

Then

$$\max_{1\le i\le n}\left|\frac{1}{T}\sum_{t=1}^{T} Z_{t,i}\right| = O_p\left(\frac{\log n}{T} \vee \sqrt{\frac{\log n}{T}}\right).$$

Proof. We need to use joint asymptotics $n, T \to \infty$. We shall use the preceding inequality with $\epsilon = (2\log n)/(T\sigma_0^2)$. Fix $\varepsilon > 0$. These exist $N_\varepsilon := 2/\varepsilon$, T_ε and $M_\varepsilon := \max(4K, 4\sigma_0)$ such that for all $n > N_\varepsilon$ and $T > T_\varepsilon$ we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\le i\le n}\left|\frac{1}{T}\sum_{t=1}^{T} Z_{t,i}\right| \ge M_\varepsilon\left(\frac{\log n}{T} \vee \sqrt{\frac{\log n}{T}}\right)\right)\\
&\le \sum_{i=1}^{n}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Z_{t,i}\right| \ge \sigma_0^2\left[K\epsilon + \sqrt{2\epsilon}\right]\right) \le 2e^{\log n - 2\log n} = \frac{2}{n} < \varepsilon.
\end{aligned}$$

$\square$

Lemma A.5. *Let* A, B *be* $n \times n$ *positive semidefinite matrices and not both singular. Then*

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

Proof. See Horn and Johnson (1985) p410. $\square$

Lemma A.6. *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.

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

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

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

We now solve for $z_{n,T}$:

$$z_{n,T} = \frac{v_{n,T}^2 w_{n,T}}{1 - v_{n,T}w_{n,T}} = O_p(a_{n,T}).$$

$\square$

Theorem A.2 (Higham (2008) p269; Dieci, Morini, and Papini (1996))**.** *For $A \in \mathbb{C}^{n\times n}$ with no eigenvalues lying on the closed negative real axis $(-\infty, 0]$,*

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

Proposition A.9. *Let A, B be $n \times n$ real matrices. Suppose that A is symmetric, positive definite for all n and its minimum eigenvalue is 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*

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

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

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

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 real, symmetric, and 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. □

Proposition A.10. *Consider real matrices A $(m \times n)$ and B $(p \times q)$. Then*

$$\|A \otimes B\|_{\ell_2} = \|A\|_{\ell_2}\|B\|_{\ell_2}.$$

Proof.

$$\begin{aligned}\|A\otimes B\|_{\ell_2} &= \sqrt{\text{maxeval}[(A\otimes B)^\intercal(A\otimes B)]} = \sqrt{\text{maxeval}[(A^\intercal\otimes B^\intercal)(A\otimes B)]}\\ &= \sqrt{\text{maxeval}[A^\intercal A\otimes B^\intercal B]} = \sqrt{\text{maxeval}[A^\intercal A]\text{maxeval}[B^\intercal B]} = \|A\|_{\ell_2}\|B\|_{\ell_2},\end{aligned}$$

where the fourth equality is due to the fact that both $A^\intercal A$ and $B^\intercal B$ are symmetric, positive semidefinite. □

Proposition A.11. *Suppose we have subgaussian random variables $Z_{l,t,j}$ for $l = 1,\ldots,L$ ($L \geq 2$ fixed), $t = 1,\ldots,T$ and $j = 1,\ldots,p$. Z_{l_1,t_1,j_1} and Z_{l_2,t_2,j_2} are independent as long as $t_1 \neq t_2$ regardless of the values of other subscripts. Then,*

$$\max_{1\leq j\leq p}\max_{1\leq t\leq T}\mathbb{E}\left|\prod_{l=1}^{L} Z_{l,t,j}\right| \leq A = O(1),$$

for some positive constant A and

$$\max_{1\leq j\leq p}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\prod_{l=1}^{L}Z_{l,t,j} - \mathbb{E}\left[\prod_{l=1}^{L}Z_{l,t,j}\right]\right)\right| = O_p\left(\sqrt{\frac{(\log p)^{L+1}}{T}}\right).$$

Proof. See Proposition 3 of Kock and Tang (2018). □

Lemma A.7. *Let A be a $p \times p$ symmetric matrix and $\hat{v}, v \in \mathbb{R}^p$. Then*

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

Proof. See Lemma 3.1 in the supplementary material of van de Geer, Buhlmann, Ritov, and Dezeure (2014). □

Lemma A.8. *Let A and B be $m \times n$ and $p \times q$ matrices, respectively. There exists a unique permutation matrix $P_K := I_n \otimes K_{q,m} \otimes I_p$, where $K_{q,m}$ is the commutation matrix, such that*

$$\mathrm{vec}(A \otimes B) = P(\mathrm{vec}\, A \otimes \mathrm{vec}\, B).$$

Proof. Magnus and Neudecker (2007) Theorem 3.10 p55. □

Theorem A.3. *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. □

Lemma A.9 (van der Vaart (1998) p27)**.**

$$\frac{\chi_k^2 - k}{\sqrt{2k}} \xrightarrow{d} N(0, 1),$$

as $k \to \infty$.

Lemma A.10 (van der Vaart (2010) p41)**.** *For $T, n \in \mathbb{N}$ let $X_{T,n}$ be random vectors such that $X_{T,n} \rightsquigarrow X_n$ as $T \to \infty$ for every fixed n such that $X_n \rightsquigarrow X$ as $n \to \infty$. Then there exists a sequence $n_T \to \infty$ such that $X_{T,n_T} \rightsquigarrow X$ as $T \to \infty$.*

References

Akdemir, D. and A. K. Gupta (2011): "Array Variate Random Variables with Multiway Kronecker Delta Covariance Matrix Structure," *Journal of Algebraic Statistics*, 2, 98–113.

Allen, G. I. (2012): "Regularized Tensor Factorizations and Higher-Order Principal Components Analysis," *arXiv preprint arXiv:1202.2476v1.*

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, Z. and Y. Wu (1994): "Limiting Behaviour of M-Estimators of Regression Coefficients in High Dimensional Linear Models. I. Scale-Dependent Case," *Journal of Multivariate Analysis*, 51, 211–239.

Battey, H. and J. Fan (2017): "Simultaneous Estimation of a Logarithmically Sparse Covariance Matrix and its Inverse," *Working paper.*

Bauwens, L., S. Laurent, and J. V. K. Rombouts (2006): "Multivariate GARCH models: A survey," *Journal of Applied Econometrics*, 31, 79–109.

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.

Browne, M. W., R. C. MacCallum, C.-T. Kim, B. L. Andersen, and R. Glaser (2002): "When Fit Indices and Residuals are Incompatible," *Psychological Methods*, 7, 403–421.

Browne, M. W. and A. Shapiro (1991): "Invariance of Covariance Structures under Groups of Transformations," *Metrika*, 38, 345–355.

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

Cai, T. T., Z. Ren, and H. H. Zhou (2014): "Estimating Structured High-Dimensional Covariance and Precision Matrices: Optimal Rates and Adaptive Estimation," *Working Paper*.

Cai, T. T., C.-H. Zhang, and H. H. Zhou (2010): "Optimal Rates of Convergence for Covariance Matrix Estimation," *The Annals of Statistics*, 38, 2118–2144.

Cai, T. T. and H. H. Zhou (2012): "Optimal Rates of Convergence for Sparse Covariance Matrix Estimation," *The Annals of Statistics*, 40, 2389–2420.

Campbell, D. T. and E. J. O'Connell (1967): "Method Factors in Multitrait-Multimethod Matrices: Multiplicative Rather than Additive?" *Multivariate Behaviorial 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.

Chang, J., S. X. Chen, and X. Chen (2015): "High Dimensional Generalized Empirical Likelihood for Moment Restrictions with Dependent Data," *Journal of Econometrics*, 283–304.

Cohen, J. E., K. Usevich, and P. Comon (2016): "A Tour of Constrained Tensor Canonical Polyadic Decomposition," *Research Report*.

Constantinou, P., P. Kokoszka, and M. Reimherr (2015): "Testing Separability of Space-Time Functional Processes," *arXiv preprint arXiv: 1509.07017v1*.

Cudeck, R. (1988): "Multiplicative Models and MTMM Matrices," *Journal of Educational Statistics*, 13, 131–147.

Dieci, L., B. Morini, and A. Papini (1996): "Computational Techniques for Real Logarithms of Matrices," *SIAM Journal on Matrix Analysis and Applications*, 17, 570–593.

Dobra, A. (2014): "Graphical Modeling of Spatial Health Data," *arXiv preprint arXiv: 1411.6512v1*.

Engle, R. F. (2002): "Dynamical Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroscedastic Models," *Journal of Business and Economic Statistics*, 20, 339–350.

Engle, R. F. and K. F. Kroner (1995): "Multivariate simultaneous generalized ARCH," *Econometric Theory*, 11, 122–150.

Fama, E. F. and K. R. French (1993): "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 33, 3–56.

Fan, J., Y. Fan, and J. Lv (2008): "High Dimensional Covariance Matrix Estimation Using a Factor Model," *Journal of Econometrics*, 147, 186–197.

Fan, J., Y. Liao, and M. Mincheva (2013): "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 603680.

Fan, J., Y. Liao, and X. Shi (2015): "Risks of Large Portfolios," *Journal of Econometrics*, 186, 367–387.

Fosdick, B. K. and P. D. Hoff (2014): "Separable Factor Analysis with Applications to Mortality Data," *The Annals of Applied Statistics*, 8, 120147.

Gerard, D. and P. Hoff (2015): "Equivariant Minimax Dominators of the MLE in the Array Normal Model," *Journal of Multivariate Analysis*, 137, 42–49.

Gil', M. I. (2012): "Perturbations of the Matrix Logarithm," *Journal of Applied Analysis*, 18, 47–68.

He, X. and Q. Shao (2000): "On Parameters of Increasing Dimensions," *Journal of Multivariate Analysis*, 73, 120–135.

Higham, N. J. (2008): *Functions of Matrices: Theory and Computation*, SIAM.

Hoff, P. D. (2011): "Separable Covariance Arrays via the Tucker Product, with Applications to Multivariate Relational Data," *Bayesian Analysis*, 6, 179–196.

——— (2015): "Multilinear Tensor Regression for Longitudinal Relational Data," *The Annals of Applied Statistics*, 9, 11691193.

——— (2016): "Equivariant and Scale-Free Tucker Decomposition Models," *Bayesian Analysis*, 11, 627648.

Horn, R. and C. Johnson (1985): *Matrix Analysis*, Cambridge University Press.

——— (2013): *Matrix Analysis Second Edition*, Cambridge University Press.

Horn, R. A. and C. R. Johnson (1991): *Topics in Matrix Analysis*, Cambridge University Press.

Huber, P. J. (1973): "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799–821.

Kock, A. B. and H. Tang (2018): "Uniform Inference in High-Dimensional Dynamic Panel Data Models With Approximately Sparse Fixed Effects," *Econometric Theory forthcoming.*

Krijnen, W. P. (2004): "Convergence in Mean Square of Factor Predictors," *British Journal of Mathematical and Statistical Psychology*, 57, 311326.

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, 365411.

——— (2012): "Nonlinear Shrinkage Estimation of Large Dimensional Covariance Matrices," *Annals of Statistics*, 40, 1024–1060.

——— (2015): "Spectrum Estimation: a Unified Framework for Covariance Matrix Estimation and PCA in Large Dimensions," *Journal of Multivariate Analysis*, 139, 360–384.

Leiva, R. and A. Roy (2014): "Classification of Higher-Order Data with Separable Covariance and Structured Multiplicative or Additive Mean Models," *Communications in Statistics - Theory and Methods*, 43, 989–1012.

Leng, C. and C. Y. Tang (2012): "Sparse Matrix Graphical Models," *Journal of the American Statistical Association*, 107, 1187–1200.

Lewis, R. and G. C. Reinsel (1985): "Prediction of Multivariate Time Series by Autoregressive Model Fitting," *Journal of Multivariate Analysis*, 16, 393–411.

Li, L. and X. Zhang (2016): "Parsimonious Tensor Response Regression," *Journal of the American Statistical Association.*

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.

Mammen, E. (1989): "Asymptotics with Increasing Dimension for Robust Regression with Applications to the Bootstrap," *The Annals of Statistics*, 17, 382–400.

Manceura, A. M. and P. Dutilleul (2013): "Maximum Likelihood Estimation for the Tensor Normal Distribution: Algorithm, Minimum Sample Size, and Empirical Bias and Dispersion," *Journal of Computational and Applied Mathematics*, 239, 37–49.

Neudecker, H. and A. M. Wesselman (1990): "The Asymptotic Variance Matrix of the Sample Correlation Matrix," *Linear Algebra and Its Applications*, 127, 589–599.

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

Ning, Y. and H. Liu (2013): "High-Dimensional Semiparametric Bigraphical Models," *Biometrika*, 100, 655–670.

Ohlson, M., M. R. Ahmada, and D. von Rosen (2013): "The Multilinear Normal Distribution: Introduction and Some Basic Properties," *Journal of Multivariate Analysis*, 113, 37–47.

Onatski, A. (2009): "Testing Hypotheses about the Number of Factors in Large Factor Models," *Econometrica*, 77, 1447–1479.

Phillips, P. C. and H. R. Moon (1999): "Linear Regression Limit Theory for Nonstationary Panel Data," *Econometrica*, 67, 1057–1111.

Pitsianis, N. P. (1997): "The Kronecker Product in Approximation and Fast Transform Generation," *PhD Thesis.*

Portnoy, S. (1985): "Asymptotic Behaviour of M-estimators of p Regression Parameters when p^2/n is Large. II. Normal Approximation," *The Annals of Statistics*, 13, 1403–1417.

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

Singull, M., M. R. Ahmad, and D. von Rosen (2012): “More on the Kronecker Structured Covariance Matrix,” *Communications in Statistics-Theory and Methods*, 41, 25122523.

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, 11661202.

van der Vaart, A. (1998): *Asymptotic Statistics*, Cambridge University Press.

——— (2010): *Time Series.*

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

van Loan, C. F. (2000): “The Ubiquitous Kronecker Product,” *Journal of Computational and Applied Mathematics*, 123, 85–100.

Verhees, J. and T. J. Wansbeek (1990): “A Multimode Direct Product Model for Covariance Structure Analysis,” *British Journal of Mathematical and Statistical Psychology*, 43, 231240.

Vershynin, R. (2011): “Introduction to the non-asymptotic analysis of random matrices,” *Chapter 5 of: Compressed Sensing, Theory and Applications.*

——— (2012): “How Close is the Sample Covariance Matrix to the Actual Covariance Matrix?” *Journal of Theoretical Probability*, 25, 655–686.

Volfovsky, A. and P. D. Hoff (2014): “Hierarchical Array Priors for ANOVA Decompositions,” *The Annals of Applied Statistics*, 8, 19–41.

——— (2015): “Testing for Nodal Dependence in Relational Data Matrices,” *Journal of the American Statistical Association*, 110, 1037–1046.

Wang, W. and J. Fan (2016): “Asymptotics of Empirical Eigen-Structure for High Dimensional Spiked Covariance,” *The Annals of Statistics*, To appear.

Welsh, A. H. (1989): “On M-Processes and M-Estimation,” *The Annals of Statistics*, 17, 337–361.

Yao, J., S. Zheng, and Z. Bai (2015): *Large Sample Covariance Matrices and High-Dimensional Data Analysis*, Cambridge University Press.

Yin, J. and H. Li (2012): “Model Selection and Estimation in the Matrix Normal Graphical Model,” *Journal of Multivariate Analysis*, 107, 119140.

Yohai, V. J. and R. A. Maronna (1979): “Asymptotic Behaviour of M-Estimators for the Linear Model,” *The Annals of Statistics*, 7, 258–268.

Supplementary Material for "Estimation of a Multiplicative Correlation Structure in the Large Dimensional Case"

Christian M. Hafner* Université catholique de Louvain
Oliver B. Linton† University of Cambridge
Haihan Tang‡ Fudan University

February 28, 2018

8 Supplementary Material

8.1

Proposition 8.1. *Suppose that $A_1, A_2, \ldots, A_v$ are real symmetric and positive definite matrices of sizes $a_1 \times a_1, \ldots, a_v \times a_v$, respectively. Then*

$$\log(A_1 \otimes A_2 \otimes \cdots \otimes A_v)$$
$$= \log A_1 \otimes I_{a_2} \otimes \cdots \otimes I_{a_v} + I_{a_1} \otimes \log A_2 \otimes I_{a_3} \otimes \cdots \otimes I_{a_v} + \cdots + I_{a_1} \otimes I_{a_2} \otimes \cdots \otimes \log A_v.$$

Proof. Since $A_1, A_2, \ldots, A_v$ are real symmetric, they can be orthogonally diagonalized: for $i = 1, \ldots, v$,

$$A_i = U_i^\intercal \Lambda_i U_i,$$

where U_i is orthogonal, and $\Lambda_i = \mathrm{diag}(\lambda_{i,1}, \ldots, \lambda_{i,a_i})$ is a diagonal matrix containing the a_i eigenvalues of A_i. Positive definiteness of $A_1, A_2, \ldots, A_v$ ensures that their Kronecker product is positive definite. Then the logarithm of $A_1 \otimes A_2 \otimes \cdots \otimes A_v$ is:

$$\begin{aligned}\log(A_1 \otimes A_2 \otimes \cdots \otimes A_v) &= \log[(U_1 \otimes \cdots \otimes U_v)^\intercal (\Lambda_1 \otimes \cdots \otimes \Lambda_v)(U_1 \otimes \cdots \otimes U_v)] \\ &= (U_1 \otimes \cdots \otimes U_v)^\intercal \log(\Lambda_1 \otimes \cdots \otimes \Lambda_v)(U_1 \otimes \cdots \otimes U_v), \end{aligned} \tag{8.1}$$

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

Now let Λ_{2-v} denote $\Lambda_2 \otimes \cdots \otimes \Lambda_v$.

$$\begin{aligned}\log(\Lambda_1 \otimes \Lambda_{2-v}) &= \mathrm{diag}(\log(\lambda_{1,1}\Lambda_{2-v}), \ldots, \log(\lambda_{1,a_1}\Lambda_{2-v})) \\ &= \mathrm{diag}(\log(\lambda_{1,1} I_{a_2\cdots a_v}\Lambda_{2-v}), \ldots, \log(\lambda_{1,a_1} I_{a_2\cdots a_v}\Lambda_{2-v})) \\ &= \mathrm{diag}(\log(\lambda_{1,1} I_{a_2\cdots a_v}) + \log(\Lambda_{2-v}), \ldots, \log(\lambda_{1,a_1} I_{a_2\cdots a_v}) + \log(\Lambda_{2-v})) \\ &= \mathrm{diag}(\log(\lambda_{1,1} I_{a_2\cdots a_v}), \ldots, \log(\lambda_{1,a_1} I_{a_2\cdots a_v})) + \mathrm{diag}(\log(\Lambda_{2-v}), \ldots, \log(\Lambda_{2-v})) \\ &= \log(\Lambda_1) \otimes I_{a_2\cdots a_v} + I_{a_1} \otimes \log(\Lambda_{2-v}), \end{aligned} \tag{8.2}$$

*Institut de statistique, biostatistique et sciences actuarielles, and CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium. Email: `christian.hafner@uclouvain.be`

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

‡Corresponding author. Fanhai International School of Finance and School of Economics, Fudan University. Email: `hhtang@fudan.edu.cn`.

where the third equality holds only because $\lambda_{1,j} I_{a_2 \cdots a_v}$ and Λ_{2-v} have real positive eigenvalues only and commute for all $j = 1, \ldots, a_1$ (Higham (2008) p270 Theorem 11.3). Substitute (8.2) into (8.1):

$$\log(A_1 \otimes A_2 \otimes \cdots \otimes A_v) = (U_1 \otimes \cdots \otimes U_v)^\intercal \log(\Lambda_1 \otimes \cdots \otimes \Lambda_v)(U_1 \otimes \cdots \otimes U_v)$$
$$= (U_1 \otimes U_{2-v})^\intercal (\log \Lambda_1 \otimes I_{a_2 \cdots a_v} + I_{a_1} \otimes \log \Lambda_{2-v})(U_1 \otimes U_{2-v})$$
$$= (U_1 \otimes U_{2-v})^\intercal (\log \Lambda_1 \otimes I_{a_2 \cdots a_v})(U_1 \otimes U_{2-v}) + (U_1 \otimes U_{2-v})^\intercal (I_{a_1} \otimes \log \Lambda_{2-v})(U_1 \otimes U_{2-v})$$
$$= \log A_1 \otimes I_{a_2 \cdots a_v} + I_{a_1} \otimes \log A_{2-v},$$

where $U_{2-v} := U_2 \otimes \cdots \otimes U_v$ and $A_{2-v} := A_2 \otimes \cdots \otimes A_v$. This procedure can be repeated until we get the proposition. □

8.2

Proposition 8.2. *Let Assumptions 3.1(i), 3.2(i) be satisfied. Then*

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

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

$$\|\hat{V}_T - V\|_\infty := \max_{1 \le x,y \le n^2} |\hat{V}_{T,x,y} - V_{x,y}| = \max_{1 \le i,j,k,\ell \le n} |\hat{V}_{T,i,j,k,\ell} - V_{i,j,k,\ell}|$$
$$\le \max_{1 \le i,j,k,\ell \le n} \left| \frac{1}{T} \sum_{t=1}^T \tilde{x}_{t,i}\tilde{x}_{t,j}\tilde{x}_{t,k}\tilde{x}_{t,\ell} - \frac{1}{T} \sum_{t=1}^T \dot{x}_{t,i}\dot{x}_{t,j}\dot{x}_{t,k}\dot{x}_{t,\ell} \right| \tag{8.3}$$
$$+ \max_{1 \le i,j,k,\ell \le n} \left| \frac{1}{T} \sum_{t=1}^T \dot{x}_{t,i}\dot{x}_{t,j}\dot{x}_{t,k}\dot{x}_{t,\ell} - \mathbb{E}[\dot{x}_{t,i}\dot{x}_{t,j}\dot{x}_{t,k}\dot{x}_{t,\ell}] \right| \tag{8.4}$$
$$+ \max_{1 \le i,j,k,\ell \le n} \left| \left(\frac{1}{T} \sum_{t=1}^T \tilde{x}_{t,i}\tilde{x}_{t,j}\right)\left(\frac{1}{T} \sum_{t=1}^T \tilde{x}_{t,k}\tilde{x}_{t,\ell}\right) - \left(\frac{1}{T} \sum_{t=1}^T \dot{x}_{t,i}\dot{x}_{t,j}\right)\left(\frac{1}{T} \sum_{t=1}^T \dot{x}_{t,k}\dot{x}_{t,\ell}\right) \right| \tag{8.5}$$
$$+ \max_{1 \le i,j,k,\ell \le n} \left| \left(\frac{1}{T} \sum_{t=1}^T \dot{x}_{t,i}\dot{x}_{t,j}\right)\left(\frac{1}{T} \sum_{t=1}^T \dot{x}_{t,k}\dot{x}_{t,\ell}\right) - \mathbb{E}[\dot{x}_{t,i}\dot{x}_{t,j}]\mathbb{E}[\dot{x}_{t,k}\dot{x}_{t,\ell}] \right| \tag{8.6}$$

8.2.1 (8.4)

By Assumption 3.1(i), $x_{t,i}, x_{t,j}, x_{t,k}, x_{t,\ell}$ are subgaussian random variables. We now show that $\dot{x}_{t,i}, \dot{x}_{t,j}, \dot{x}_{t,k}, \dot{x}_{t,\ell}$ are also uniformly subgaussian. Without loss of generality consider $\dot{x}_{t,i}$.

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

where the fifth inequality is due to the decoupling inequality $2xy \le x^2/2 + 2y^2$, and the last equality is due to (A.20). We now consider (8.4). Invoke Proposition A.11 in Appendix A.5:

$$\max_{1 \le i,j,k,\ell \le n} \left| \frac{1}{T} \sum_{t=1}^T \dot{x}_{t,i}\dot{x}_{t,j}\dot{x}_{t,k}\dot{x}_{t,\ell} - \mathbb{E}\dot{x}_{t,i}\dot{x}_{t,j}\dot{x}_{t,k}\dot{x}_{t,\ell} \right| = O_p\left(\sqrt{\frac{\log^5 n^4}{T}}\right). \tag{8.7}$$

8.2.2 (8.6)

We now consider (8.6).

$$\begin{aligned}
&\max_{1\le i,j,k,\ell\le n}\left|\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,i}\dot{x}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,k}\dot{x}_{t,\ell}\right)-\mathbb{E}[\dot{x}_{t,i}\dot{x}_{t,j}]\mathbb{E}[\dot{x}_{t,k}\dot{x}_{t,\ell}]\right| \\
&\le \max_{1\le i,j,k,\ell\le n}\left|\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,i}\dot{x}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,k}\dot{x}_{t,\ell}-\mathbb{E}[\dot{x}_{t,k}\dot{x}_{t,\ell}]\right)\right| \qquad (8.8)\\
&\quad + \max_{1\le i,j,k,\ell\le n}\left|\mathbb{E}[\dot{x}_{t,k}\dot{x}_{t,\ell}]\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,i}\dot{x}_{t,j}-\mathbb{E}[\dot{x}_{t,i}\dot{x}_{t,j}]\right)\right|. \qquad (8.9)
\end{aligned}$$

Consider (8.8).

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

where the first equality is due to Proposition A.11 in Appendix A.5 and the last equality is due to Assumption 3.2(i). Now consider (8.9).

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

where the equality is due to Proposition A.11 in Appendix A.5. Thus

$$\max_{1\le i,j,k,\ell\le n}\left|\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,i}\dot{x}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,k}\dot{x}_{t,\ell}\right)-\mathbb{E}[\dot{x}_{t,i}\dot{x}_{t,j}]\mathbb{E}[\dot{x}_{t,k}\dot{x}_{t,\ell}]\right|=O_p\left(\sqrt{\frac{\log^3 n^2}{T}}\right). \qquad (8.10)$$

8.2.3 (8.3)

We first give a rate for $\max_{1\le i\le n}|\bar{x}_i-\mu_i|$. The index i is arbitrary and could be replaced with j,k,ℓ. By Assumption 3.1(i), $\{x_{t,i}\}_{t=1}^{T}$ are independent subgaussian random variables. For $\epsilon>0$, $\mathbb{P}(|x_{t,i}|\ge\epsilon)\le Ke^{-C\epsilon^2}$. It follows from Lemma 2.2.1 in van der Vaart and Wellner (1996) that $\|x_{t,i}\|_{\psi_2}\le(1+K)^{1/2}/C^{1/2}$. Then $\|x_{t,i}-\mathbb{E}x_{t,i}\|_{\psi_2}\le\|x_{t,i}\|_{\psi_2}+\mathbb{E}\|x_{t,i}\|_{\psi_2}\le\frac{2(1+K)^{1/2}}{C^{1/2}}$. Next, using the second last inequality in van der Vaart and Wellner (1996) p95, we have

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

Then, by the definition of the Orlicz norm, $\mathbb{E}\left[e^{W|x_{t,i}-\mathbb{E}x_{t,i}|}\right]\le 2$. Use Fubini's theorem to expand out the exponential moment. It is easy to see that $x_{t,i}-\mathbb{E}x_{t,i}$ satisfies the moment conditions of Bernstein's inequality in Appendix A.5 with $A = \frac{1}{W}$ and $\sigma_0^2 = \frac{2}{W^2}$. Now invoke Bernstein's inequality for all $\epsilon > 0$

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

Invoking Proposition A.8 in Appendix A.5, we have

$$\max_{1\leq i\leq n}|\bar{x}_i - \mu_i| = \max_{1\leq i\leq n}\left|\frac{1}{T}\sum_{t=1}^{T}(x_{t,i} - \mathbb{E}x_{t,i})\right| = O_p\left(\frac{\log n}{T} \vee \sqrt{\frac{\log n}{T}}\right) = O_p\left(\sqrt{\frac{\log n}{T}}\right), \quad (8.11)$$

where the last equality is due to Assumption 3.2(i). Then we also have

$$\max_{1\leq i\leq n}|\bar{x}_i| = \max_{1\leq i\leq n}|\bar{x}_i - \mu_i + \mu_i| \leq \max_{1\leq i\leq n}|\bar{x}_i - \mu_i| + \max_{1\leq i\leq n}|\mu_i| = O_p\left(\sqrt{\frac{\log n}{T}}\right) + O(1) = O_p(1). \quad (8.12)$$

We now consider (8.3):

$$\max_{1\leq i,j,k,\ell\leq n}\left|\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{t,i}\tilde{x}_{t,j}\tilde{x}_{t,k}\tilde{x}_{t,\ell} - \frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,i}\dot{x}_{t,j}\dot{x}_{t,k}\dot{x}_{t,\ell}\right|.$$

With tedious expansion, simplification and recognition the indices i, j, k, ℓ are completely symmetric, we can bound (8.3) by

$$\max_{1\leq i,j,k,\ell\leq n}\left|\bar{x}_i\bar{x}_j\bar{x}_k\bar{x}_\ell - \mu_i\mu_j\mu_k\mu_\ell\right| \quad (8.13)$$

$$+ 4\max_{1\leq i,j,k,\ell\leq n}\left|\bar{x}_i\left(\bar{x}_j\bar{x}_k\bar{x}_\ell - \mu_j\mu_k\mu_\ell\right)\right| \quad (8.14)$$

$$+ 6\max_{1\leq i,j,k,\ell\leq n}\left|\left(\frac{1}{T}\sum_{t=1}^{T}x_{t,i}x_{t,j}\right)\left(\bar{x}_k\bar{x}_\ell - \mu_k\mu_\ell\right)\right| \quad (8.15)$$

$$+ 4\max_{1\leq i,j,k,\ell\leq n}\left|\left(\frac{1}{T}\sum_{t=1}^{T}x_{t,i}x_{t,j}x_{t,k}\right)\left(\bar{x}_\ell - \mu_\ell\right)\right|. \quad (8.16)$$

We consider (8.13) first. (8.13) can be bounded by repeatedly invoking triangular inequalities (e.g., inserting terms like $\mu_i\bar{x}_j\bar{x}_k\bar{x}_\ell$) using (A.20), (8.12) and (8.11). (8.13) is of order $O_p(\sqrt{\log n/T})$. (8.14) is of order $O_p(\sqrt{\log n/T})$ by a similar argument. (8.15) and (8.16) are of the same order $O_p(\sqrt{\log n/T})$ using a similar argument provided that both $\max_{1\leq i,j\leq n}|\sum_{t=1}^{T}x_{t,i}x_{t,j}|/T$ and $\max_{1\leq i,j,k\leq n}|\sum_{t=1}^{T}x_{t,i}x_{t,j}x_{t,k}|/T$ are $O_p(1)$; these follow from Proposition A.11 in Appendix A.5. Thus

$$\max_{1\leq i,j,k,\ell\leq n}\left|\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{t,i}\tilde{x}_{t,j}\tilde{x}_{t,k}\tilde{x}_{t,\ell} - \frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,i}\dot{x}_{t,j}\dot{x}_{t,k}\dot{x}_{t,\ell}\right| = O_p(\sqrt{\log n/T}). \quad (8.17)$$

8.2.4 (8.5)

We now consider (8.5).

$$\max_{1\leq i,j,k,\ell\leq n}\left|\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{t,i}\tilde{x}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{t,k}\tilde{x}_{t,\ell}\right) - \left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,i}\dot{x}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,k}\dot{x}_{t,\ell}\right)\right|$$

$$\leq \max_{1\leq i,j,k,\ell\leq n}\left|\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{t,i}\tilde{x}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{x}_{t,k}\tilde{x}_{t,\ell} - \dot{x}_{t,k}\dot{x}_{t,\ell}\right)\right)\right| \quad (8.18)$$

$$+ \max_{1\leq i,j,k,\ell\leq n}\left|\left(\frac{1}{T}\sum_{t=1}^{T}\dot{x}_{t,k}\dot{x}_{t,\ell}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{x}_{t,i}\tilde{x}_{t,j} - \dot{x}_{t,i}\dot{x}_{t,j}\right)\right)\right| \quad (8.19)$$

It suffices to give a bound for (8.18) as the bound for (8.19) is of the same order and follows through similarly. First, it is easy to show that $\max_{1\leq i,j\leq n}|\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{t,i}\tilde{x}_{t,j}|=\max_{1\leq i,j\leq n}|\frac{1}{T}\sum_{t=1}^{T}x_{t,i}x_{t,j}-\bar{x}_i\bar{x}_j|=O_p(1)$ (using Proposition A.11 in Appendix A.5). Next

$$\max_{1\leq k,\ell\leq n}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{x}_{t,k}\tilde{x}_{t,\ell}-\dot{x}_{t,k}\dot{x}_{t,\ell}\right)\right|=\max_{1\leq k,\ell\leq n}\left|-(\bar{x}_k-\mu_k)(\bar{x}_\ell-\mu_\ell)\right|=O_p\left(\frac{\log n}{T}\right). \tag{8.20}$$

The proposition follows after summing up the rates for (8.7), (8.10), (8.17) and (8.20). □

8.3

Proposition 8.3. *Let Assumptions 3.1(i), 3.2(i) and 3.3 be satisfied. Then we have*

$$\|P\|_{\ell_2}=O(1),\qquad \|\hat{P}_T\|_{\ell_2}=O_p(1),\qquad \|\hat{P}_T-P\|_{\ell_2}=O_p\left(\sqrt{\frac{n}{T}}\right). \tag{8.21}$$

Proof. The proofs for $\|P\|_{\ell_2}=O(1)$ and $\|\hat{P}_T\|_{\ell_2}=O_p(1)$ are exactly the same, so we only give the proof for the latter.

$$\begin{aligned}\|\hat{P}_T\|_{\ell_2}&=\|I_{n^2}-D_nD_n^+(I_n\otimes\hat{\Theta}_T)M_d\|_{\ell_2}\leq 1+\|D_nD_n^+(I_n\otimes\hat{\Theta}_T)M_d\|_{\ell_2}\\&\leq 1+\|D_n\|_{\ell_2}\|D_n^+\|_{\ell_2}\|I_n\otimes\hat{\Theta}_T\|_{\ell_2}\|M_d\|_{\ell_2}=1+2\|I_n\|_{\ell_2}\|\hat{\Theta}_T\|_{\ell_2}=O_p(1)\end{aligned}$$

where the second equality is due to (A.13) and Proposition A.10 in Appendix A.5, and last equality is due to Proposition A.5(ii). Now,

$$\begin{aligned}&\|\hat{P}_T-P\|_{\ell_2}=\|I_{n^2}-D_nD_n^+(I_n\otimes\hat{\Theta}_T)M_d-(I_{n^2}-D_nD_n^+(I_n\otimes\Theta)M_d)\|_{\ell_2}\\&=\|D_nD_n^+(I_n\otimes\hat{\Theta}_T)M_d-D_nD_n^+(I_n\otimes\Theta)M_d)\|_{\ell_2}=\|D_nD_n^+(I_n\otimes(\hat{\Theta}_T-\Theta))M_d\|_{\ell_2}\\&=O_p(\sqrt{n/T}),\end{aligned}$$

where the last equality is due to Proposition 3.1(i). □

Proof of Theorem 3.2. We write

$$\begin{aligned}&\frac{\sqrt{T}c^\intercal(\hat{\theta}_T-\theta^0)}{\sqrt{\hat{G}_T}}\\&=\frac{\sqrt{T}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H\operatorname{vec}(\hat{\Theta}_T-\Theta)}{\sqrt{\hat{G}_T}}+\frac{\sqrt{T}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\operatorname{vec}O_p(\|\hat{\Theta}_T-\Theta\|_{\ell_2}^2)}{\sqrt{\hat{G}_T}}\\&=\frac{\sqrt{T}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H\left.\frac{\partial\operatorname{vec}\Theta}{\partial\operatorname{vec}\Sigma}\right|_{\Sigma=\tilde{\Sigma}_T^{(i)}}\operatorname{vec}(\hat{\Sigma}_T-\Sigma)}{\sqrt{\hat{G}_T}}\\&\quad+\frac{\sqrt{T}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\operatorname{vec}O_p(\|\hat{\Theta}_T-\Theta\|_{\ell_2}^2)}{\sqrt{\hat{G}_T}}\\&=:\hat{t}_1+\hat{t}_2,\end{aligned}$$

where $\left.\frac{\partial\operatorname{vec}\Theta}{\partial\operatorname{vec}\Sigma}\right|_{\Sigma=\tilde{\Sigma}_T^{(i)}}$ denotes the matrix whose jth row is the jth row of the Jacobian matrix $\frac{\partial\operatorname{vec}\Theta}{\partial\operatorname{vec}\Sigma}$ evaluated at $\operatorname{vec}\tilde{\Sigma}_T^{(j)}$, which is a point between $\operatorname{vec}\Sigma$ and $\operatorname{vec}\hat{\Sigma}_T$, for $j=1,\dots,n^2$.

Define

$$t_1:=\frac{\sqrt{T}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+HP(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}(\tilde{\Sigma}_T-\Sigma)}{\sqrt{G}}.$$

To prove Theorem 3.2, it suffices to show $t_1\xrightarrow{d}N(0,1)$, $t_1-\hat{t}_1=o_p(1)$, and $\hat{t}_2=o_p(1)$. The proof is similar to that of Theorem 3.1, so we will be concise for the parts which are almost identical to that of Theorem 3.1.

8.3.1 $t_1 \xrightarrow{d} N(0,1)$

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

$$t_1 = \frac{\sqrt{T}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+HP(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}\left(\frac{1}{T}\sum_{t=1}^T\left[(x_t-\mu)(x_t-\mu)^\intercal - \mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right)}{\sqrt{G}}$$
$$= \sum_{t=1}^T \frac{T^{-1/2}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+HP(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal - \mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]}{\sqrt{G}}$$
$$=: \sum_{t=1}^T U_{T,n,t}.$$

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 (e.g., Phillips and Moon (1999) Theorem 2 p1070): for all $\varepsilon > 0$,

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

For any $\gamma > 2$,

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

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

$$\begin{aligned}
G &= c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+HP(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal HD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\\
&\geq \operatorname{mineval}\left(E^\intercal WD_n^+HP(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal HD_n^{+\intercal}WE\right)\|(E^\intercal WE)^{-1}c\|_2^2\\
&\geq \frac{n}{\varpi}\operatorname{mineval}^2(W)c(E^\intercal WE)^{-2}c \geq \frac{n}{\varpi}\operatorname{mineval}^2(W)\operatorname{mineval}\left((E^\intercal WE)^{-2}\right)\\
&= \frac{n\cdot\operatorname{mineval}^2(W)}{\varpi\operatorname{maxeval}^2(E^\intercal WE)} \geq \frac{n}{\varpi\operatorname{maxeval}^2(W^{-1})\operatorname{maxeval}^2(W)\operatorname{maxeval}^2(E^\intercal E)}\\
&= \frac{n}{\varpi\kappa^2(W)\operatorname{maxeval}^2(E^\intercal E)}
\end{aligned}$$

where the second inequality is due to Assumption 3.5(ii). Using (A.15), we have

$$\frac{1}{\sqrt{G}} = O(\sqrt{s^2\cdot n\cdot\kappa^2(W)\cdot\varpi}). \tag{8.22}$$

Then a sufficient condition for the Lindeberg condition is:

$$\begin{aligned}
&T^{1-\frac{\gamma}{2}}(s^2n\kappa^2(W)\varpi)^{\gamma/2}\\
&\cdot\mathbb{E}\left|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+HP(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal - \mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right|^\gamma\\
&= o(1),
\end{aligned} \tag{8.23}$$

for some $\gamma > 2$. The verification will be exactly the same as that of (A.18). In the end, we have

$$\begin{aligned}
&T^{1-\frac{\gamma}{2}}(s^2n\kappa^2(W)\varpi)^{\gamma/2}\\
&\cdot\mathbb{E}\left|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+HP(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal - \mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right|^\gamma\\
&= T^{1-\frac{\gamma}{2}}(s^2n\kappa^2(W)\varpi)^{\gamma/2}(\varpi\kappa(W)n)^{\gamma/2}O(\log^\gamma n) = O\left(\frac{n^2\cdot\kappa^3(W)\cdot\log^4 n\cdot\varpi^2}{T^{1-\frac{2}{\gamma}}}\right)^{\gamma/2} = o(1)
\end{aligned}$$

by Assumption 3.2(ii).

8.3.2 $t_1 - \hat{t}_1 = o_p(1)$

We now show that $t_1 - \hat{t}_1 = o_p(1)$. Let A and $\hat{A}$ denote the numerators of t_1 and $\hat{t}_1$, respectively.

$$t_1 - \hat{t}_1 = \frac{A}{\sqrt{G}} - \frac{\hat{A}}{\sqrt{\hat{G}_T}} = \frac{\sqrt{s^2 n\kappa^2(W)\varpi}A}{\sqrt{s^2 n\kappa^2(W)\varpi G}} - \frac{\sqrt{s^2 n\kappa^2(W)\varpi}\hat{A}}{\sqrt{s^2 n\kappa^2(W)\varpi \hat{G}_T}}.$$

Since we have already shown in (8.22) that $s^2 n\kappa^2(W)\varpi G$ is bounded away from zero by an absolute constant, it suffices to show the denominators as well as numerators of t_1 and $\hat{t}_1$ are asymptotically equivalent.

8.3.3 Denominators of t_1 and $\hat{t}_1$

We first show that the denominators of t_1 and $\hat{t}_1$ are asymptotically equivalent, i.e.,

$$s^2 n\kappa^2(W)\varpi|\hat{G}_T - G| = o_p(1).$$

Define

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

By the triangular inequality: $s^2 n\kappa^2(W)\varpi|\hat{G}_T - G| \leq s^2 n\kappa^2(W)\varpi|\hat{G}_T - \tilde{G}_T| + s^2 n\kappa^2(W)\varpi|\tilde{G}_T - G|$. First, we prove $s^2 n\kappa^2(W)\varpi|\hat{G}_T - \tilde{G}_T| = o_p(1)$.

$$\begin{aligned}
&s^2 n\kappa^2(W)\varpi|\hat{G}_T - \tilde{G}_T| \\
&= s^2 n\kappa^2(W)\varpi|c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T\hat{P}_T(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2})\hat{V}_T(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2})\hat{P}_T^\intercal\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c \\
&\qquad - c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T\hat{P}_T(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2})V(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2})\hat{P}_T^\intercal\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c| \\
&= s^2 n\kappa^2(W)\varpi \\
&\quad \cdot |c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T\hat{P}_T(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2})(\hat{V}_T - V)(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2})\hat{P}_T^\intercal\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c| \\
&\leq s^2 n\kappa^2(W)\varpi\|\hat{V}_T - V\|_\infty\|(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2})\hat{P}_T^\intercal\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_1^2 \\
&\leq s^2 n^3\kappa^2(W)\varpi\|\hat{V}_T - V\|_\infty\|(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2})\hat{P}_T^\intercal\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2^2 \\
&\leq s^2 n^3\kappa^2(W)\varpi\|\hat{V}_T - V\|_\infty\|(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2})\|_{\ell_2}^2\|\hat{P}_T^\intercal\|_{\ell_2}^2\|\hat{H}_T\|_{\ell_2}^2\|D_n^{+\intercal}\|_{\ell_2}^2\|WE(E^\intercal WE)^{-1}\|_{\ell_2}^2 \\
&= O_p(s^2 n^2\kappa^3(W)\varpi^2)\|\hat{V}_T - V\|_\infty = O_p\left(\sqrt{\frac{n^4\kappa^6(W)s^4\varpi^4\log^5 n^4}{T}}\right) = o_p(1),
\end{aligned}$$

where $\|\cdot\|_\infty$ denotes the absolute elementwise maximum, the third equality is due to Proposition A.3(v), Proposition A.10 in Appendix A.5, (A.12), (A.19), (A.13) and (8.21), the second last equality is due to Proposition 8.2 in SM 8.2, and the last equality is due to Assumption 3.2(ii).

We now prove $s^2 n\kappa^2(W)\varpi|\tilde{G}_T - G| = o_p(1)$. Define

$$\begin{aligned}
\tilde{G}_{T,a} &:= c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T\hat{P}_T(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})\hat{P}_T^\intercal\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c \\
\tilde{G}_{T,b} &:= c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T P(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal\hat{H}_T D_n^{+\intercal}WE(E^\intercal WE)^{-1}c.
\end{aligned}$$

We use triangular inequality again

$$s^2 n\kappa^2(W)\varpi|\tilde{G}_T - G| \leq s^2 n\kappa^2(W)\varpi|\tilde{G}_T - \tilde{G}_{T,a}| + s^2 n\kappa^2(W)\varpi|\tilde{G}_{T,a} - \tilde{G}_{T,b}| + s^2 n\kappa^2(W)\varpi|\tilde{G}_{T,b} - G|. \tag{8.24}$$

We consider the first term on the right hand side of (8.24).

$$
\begin{aligned}
&s^2n\kappa^2(W)\varpi|\tilde{G}_T-\tilde{G}_{T,a}|=\\
&s^2n\kappa^2(W)\varpi|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T\hat{P}_T(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2})V(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2})\hat{P}_T^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\\
&\qquad-c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T\hat{P}_T(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})\hat{P}_T^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c|\\
&\le s^2n\kappa^2(W)\varpi\big|\text{maxeval}(V)\big|^2\|(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2}-D^{-1/2}\otimes D^{-1/2})\hat{P}_T^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2^2\\
&\quad+s^2n\kappa^2(W)\varpi\|V(D^{-1/2}\otimes D^{-1/2})\hat{P}_T^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2\\
&\qquad\cdot\|(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2}-D^{-1/2}\otimes D^{-1/2})\hat{P}_T^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2 \qquad (8.25)
\end{aligned}
$$

where the inequality is due to Lemma A.7 in Appendix A.5. We consider the first term of (8.25) first.

$$
\begin{aligned}
&s^2n\kappa^2(W)\varpi\big|\text{maxeval}(V)\big|^2\|(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2}-D^{-1/2}\otimes D^{-1/2})\hat{P}_T^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2^2\\
&=O(s^2n\kappa^2(W)\varpi)\|\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2}-D^{-1/2}\otimes D^{-1/2}\|_{\ell_2}^2\|\hat{P}_T^\intercal\|_{\ell_2}^2\|\hat{H}_T\|_{\ell_2}^2\|D_n^{+\intercal}\|_{\ell_2}^2\|WE(E^\intercal WE)^{-1}\|_{\ell_2}^2\\
&=O_p(s^2n\kappa^3(W)\varpi^2/T)=o_p(1),
\end{aligned}
$$

where the second last equality is due to (A.12), (A.13), (A.19), (8.21) and Proposition A.3(vii), and the last equality is due to Assumption 3.2(ii).

We now consider the second term of (8.25).

$$
\begin{aligned}
&2s^2n\kappa^2(W)\varpi\|V(D^{-1/2}\otimes D^{-1/2})\hat{P}_T^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2\\
&\qquad\cdot\|(\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2}-D^{-1/2}\otimes D^{-1/2})\hat{P}_T^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2\\
&\le O(s^2n\kappa^2(W)\varpi)\|\hat{D}_T^{-1/2}\otimes\hat{D}_T^{-1/2}-D^{-1/2}\otimes D^{-1/2}\|_{\ell_2}\|\hat{P}_T^\intercal\|_{\ell_2}^2\|\hat{H}_T\|_{\ell_2}^2\|D_n^{+\intercal}\|_{\ell_2}^2\|WE(E^\intercal WE)^{-1}\|_{\ell_2}^2\\
&=O(\sqrt{s^4n\kappa^6(W)\varpi^4/T})=o_p(1),
\end{aligned}
$$

where the first equality is due to (A.12), (A.13), (A.19), (8.21) and Proposition A.3(vii), and the last equality is due to Assumption 3.2(ii). We have proved $s^2n\kappa^2(W)\varpi|\tilde{G}_T-\tilde{G}_{T,a}|=o_p(1)$.

We consider the second term on the right hand side of (8.24).

$$
\begin{aligned}
&s^2n\kappa^2(W)\varpi|\tilde{G}_{T,a}-\tilde{G}_{T,b}|=\\
&s^2n\kappa^2(W)\varpi|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T\hat{P}_T(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})\hat{P}_T^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\\
&\qquad-c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_TP(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c|\\
&\le s^2n\kappa^2(W)\varpi\big|\text{maxeval}[(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})]\big|^2\|(\hat{P}_T-P)^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2^2\\
&\quad+2s^2n\kappa^2(W)\varpi\|(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2\\
&\qquad\cdot\|(\hat{P}_T-P)^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2 \qquad (8.26)
\end{aligned}
$$

where the inequality is due to Lemma A.7 in Appendix A.5. We consider the first term of (8.26) first.

$$
\begin{aligned}
&s^2n\kappa^2(W)\varpi\big|\text{maxeval}[(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})]\big|^2\|(\hat{P}_T-P)^\intercal\hat{H}_TD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2^2\\
&=O(s^2n\kappa^2(W)\varpi)\|\hat{P}_T^\intercal-P^\intercal\|_{\ell_2}^2\|\hat{H}_T\|_{\ell_2}^2\|D_n^{+\intercal}\|_{\ell_2}^2\|WE(E^\intercal WE)^{-1}\|_{\ell_2}^2\\
&=O_p(s^2n\kappa^3(W)\varpi^2/T)=o_p(1),
\end{aligned}
$$

where the second last equality is due to (A.12), (A.13), (A.19), and (8.21), and the last equality is due to Assumption 3.2(ii).

We now consider the second term of (8.26).

$$
\begin{aligned}
&2s^2 n\kappa^2(W)\varpi \|(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal \hat{H}_T D_n^{+\intercal} WE(E^\intercal WE)^{-1}c\|_2 \\
&\qquad \cdot \|(\hat{P}_T - P)^\intercal \hat{H}_T D_n^{+\intercal} WE(E^\intercal WE)^{-1}c\|_2 \\
&\le O(s^2 n\kappa^2(W)\varpi)\|\hat{P}_T^\intercal - P^\intercal\|_{\ell_2}^2 \|\hat{H}_T\|_{\ell_2}^2 \|D_n^{+\intercal}\|_{\ell_2}^2 \|WE(E^\intercal WE)^{-1}\|_{\ell_2}^2 \\
&= O(\sqrt{s^4 n\kappa^6(W)\varpi^4/T}) = o_p(1),
\end{aligned}
$$

where the first equality is due to (A.12), (A.13), (A.19), and (8.21), and the last equality is due to Assumption 3.2(ii). We have proved $s^2 n\kappa^2(W)\varpi|\tilde{G}_{T,a} - \tilde{G}_{T,b}| = o_p(1)$.

We consider the third term on the right hand side of (8.24).

$$
\begin{aligned}
&s^2 n\kappa^2(W)\varpi|\tilde{G}_{T,b} - G| = \\
&s^2 n\kappa^2(W)\varpi|c^\intercal (E^\intercal WE)^{-1}E^\intercal W D_n^{+}\hat{H}_T P(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal \hat{H}_T D_n^{+\intercal} WE(E^\intercal WE)^{-1}c \\
&\qquad - c^\intercal (E^\intercal WE)^{-1}E^\intercal W D_n^{+} HTP(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal H D_n^{+\intercal} WE(E^\intercal WE)^{-1}c| \\
&\le s^2 n\kappa^2(W)\varpi\left|\text{maxeval}[P(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal]\right|^2 \|(\hat{H}_T - H)D_n^{+\intercal} WE(E^\intercal WE)^{-1}c\|_2^2 \\
&\quad + 2s^2 n\kappa^2(W)\varpi\|P(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal H D_n^{+\intercal} WE(E^\intercal WE)^{-1}c\|_2 \\
&\qquad \cdot \|(\hat{H}_T - H)D_n^{+\intercal} WE(E^\intercal WE)^{-1}c\|_2 \qquad (8.27)
\end{aligned}
$$

where the inequality is due to Lemma A.7 in Appendix A.5. We consider the first term of (8.27) first.

$$
\begin{aligned}
&s^2 n\kappa^2(W)\varpi\left|\text{maxeval}[P(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal]\right|^2 \|(\hat{H}_T - H)D_n^{+\intercal} WE(E^\intercal WE)^{-1}c\|_2^2 \\
&= O(s^2 n\kappa^2(W)\varpi)\|\hat{H}_T - H\|_{\ell_2}^2 \|D_n^{+\intercal}\|_{\ell_2}^2 \|WE(E^\intercal WE)^{-1}\|_{\ell_2}^2 \\
&= O_p(s^2 n\kappa^3(W)\varpi^2/T) = o_p(1),
\end{aligned}
$$

where the second last equality is due to (A.12), (A.13), and (A.19), and the last equality is due to Assumption 3.2(ii).

We now consider the second term of (8.27).

$$
\begin{aligned}
&2s^2 n\kappa^2(W)\varpi\|P(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal H D_n^{+\intercal} WE(E^\intercal WE)^{-1}c\|_2 \\
&\qquad \cdot \|(\hat{H}_T - H)D_n^{+\intercal} WE(E^\intercal WE)^{-1}c\|_2 \\
&\le O(s^2 n\kappa^2(W)\varpi)\|\hat{H}_T - H\|_{\ell_2}^2 \|D_n^{+\intercal}\|_{\ell_2}^2 \|WE(E^\intercal WE)^{-1}\|_{\ell_2}^2 \\
&= O(\sqrt{s^4 n\kappa^6(W)\varpi^4/T}) = o_p(1),
\end{aligned}
$$

where the first equality is due to (A.12), (A.13), and (A.19), and the last equality is due to Assumption 3.2(ii). We have proved $s^2 n\kappa^2(W)\varpi|\tilde{G}_{T,b} - G| = o_p(1)$. Hence we have proved $s^2 n\kappa^2(W)\varpi|\tilde{G}_T - G| = o_p(1)$.

8.3.4 Numerators of t_1 and $\hat{t}_1$

We now show that numerators of t_1 and $\hat{t}_1$ are asymptotically equivalent, i.e.,

$$
\sqrt{s^2 n\kappa^2(W)\varpi}|A - \hat{A}| = o_p(1).
$$

Note that

$$\begin{aligned}
\hat{A} &= \sqrt{T}c^{\intercal}(E^{\intercal}WE)^{-1}E^{\intercal}WD_n^{+}H\left.\frac{\partial \operatorname{vec}\Theta}{\partial \operatorname{vec}\Sigma}\right|_{\Sigma=\tilde{\Sigma}_T^{(i)}}\operatorname{vec}(\hat{\Sigma}_T-\Sigma)\\
&= \sqrt{T}c^{\intercal}(E^{\intercal}WE)^{-1}E^{\intercal}WD_n^{+}H\left.\frac{\partial \operatorname{vec}\Theta}{\partial \operatorname{vec}\Sigma}\right|_{\Sigma=\tilde{\Sigma}_T^{(i)}}\operatorname{vec}(\hat{\Sigma}_T-\tilde{\Sigma}_T)\\
&\quad+\sqrt{T}c^{\intercal}(E^{\intercal}WE)^{-1}E^{\intercal}WD_n^{+}H\left.\frac{\partial \operatorname{vec}\Theta}{\partial \operatorname{vec}\Sigma}\right|_{\Sigma=\tilde{\Sigma}_T^{(i)}}\operatorname{vec}(\tilde{\Sigma}_T-\Sigma)\\
&=: \hat{A}_a+\hat{A}_b.
\end{aligned}$$

To show $\sqrt{s^2n\kappa^2(W)\varpi}|A-\hat{A}|=o_p(1)$, it suffices to show $\sqrt{s^2n\kappa^2(W)\varpi}|\hat{A}_b-A|=o_p(1)$ and $\sqrt{s^2n\kappa^2(W)\varpi}|\hat{A}_a|=o_p(1)$. We first show that $\sqrt{s^2n\kappa^2(W)\varpi}|\hat{A}_b-A|=o_p(1)$.

$$\begin{aligned}
&\sqrt{s^2n\kappa^2(W)\varpi}|\hat{A}_b-A|\\
&=\sqrt{s^2n\kappa^2(W)\varpi}\left|\sqrt{T}c^{\intercal}(E^{\intercal}WE)^{-1}E^{\intercal}WD_n^{+}H\left[\left.\frac{\partial \operatorname{vec}\Theta}{\partial \operatorname{vec}\Sigma}\right|_{\Sigma=\tilde{\Sigma}_T^{(i)}}-P(D^{-1/2}\otimes D^{-1/2})\right]\operatorname{vec}(\tilde{\Sigma}_T-\Sigma)\right|\\
&\leq \sqrt{Ts^2n\kappa^2(W)\varpi}\|(E^{\intercal}WE)^{-1}E^{\intercal}W\|_{\ell_2}\|D_n^{+}\|_{\ell_2}\|H\|_{\ell_2}\\
&\quad\cdot\left\|\left.\frac{\partial \operatorname{vec}\Theta}{\partial \operatorname{vec}\Sigma}\right|_{\Sigma=\tilde{\Sigma}_T^{(i)}}-P(D^{-1/2}\otimes D^{-1/2})\right\|_{\ell_2}\|\operatorname{vec}(\tilde{\Sigma}_T-\Sigma)\|_2\\
&=O(\sqrt{Ts^2n\kappa^2(W)\varpi})\sqrt{\varpi\kappa(W)/n}O_p\left(\sqrt{\frac{n}{T}}\right)\|\tilde{\Sigma}_T-\Sigma\|_F\leq O(\sqrt{ns^2\kappa^3(W)\varpi^2})\sqrt{n}\|\tilde{\Sigma}_T-\Sigma\|_{\ell_2}\\
&=O(\sqrt{ns^2\kappa^3(W)\varpi^2})\sqrt{n}O_p\left(\sqrt{\frac{n}{T}}\right)=O_p\left(\sqrt{\frac{n^3s^2\kappa^3(W)\varpi^2}{T}}\right)=o_p(1),
\end{aligned}$$

where the second equality is due to Assumption 3.5(i), the third equality is due to Lemma A.1, and final equality is due to Assumption 3.2(ii).

We now show that $\sqrt{s^2n\kappa^2(W)\varpi}|\hat{A}_a|=o_p(1)$.

$$\begin{aligned}
&\sqrt{s^2n\kappa^2(W)\varpi T}\left|c^{\intercal}(E^{\intercal}WE)^{-1}E^{\intercal}WD_n^{+}H\left.\frac{\partial \operatorname{vec}\Theta}{\partial \operatorname{vec}\Sigma}\right|_{\Sigma=\tilde{\Sigma}_T^{(i)}}\operatorname{vec}(\hat{\Sigma}_T-\tilde{\Sigma}_T)\right|\\
&=\sqrt{s^2n\kappa^2(W)\varpi T}\left|c^{\intercal}(E^{\intercal}WE)^{-1}E^{\intercal}WD_n^{+}H\left.\frac{\partial \operatorname{vec}\Theta}{\partial \operatorname{vec}\Sigma}\right|_{\Sigma=\tilde{\Sigma}_T^{(i)}}\operatorname{vec}\left[(\bar{x}-\mu)(\bar{x}-\mu)^{\intercal}\right]\right|\\
&\leq \sqrt{s^2n\kappa^2(W)\varpi T}\|(E^{\intercal}WE)^{-1}E^{\intercal}W\|_{\ell_2}\|D_n^{+}\|_{\ell_2}\|H\|_{\ell_2}\left\|\left.\frac{\partial \operatorname{vec}\Theta}{\partial \operatorname{vec}\Sigma}\right|_{\Sigma=\tilde{\Sigma}_T^{(i)}}\right\|_{\ell_2}\|\operatorname{vec}\left[(\bar{x}-\mu)(\bar{x}-\mu)^{\intercal}\right]\|_2\\
&=O(\sqrt{Ts^2n\kappa^2(W)\varpi})\sqrt{\varpi\kappa(W)/n}\|(\bar{x}-\mu)(\bar{x}-\mu)^{\intercal}\|_F\\
&\leq O(\sqrt{Ts^2n\kappa^2(W)\varpi})\sqrt{\varpi\kappa(W)/n}n\|(\bar{x}-\mu)(\bar{x}-\mu)^{\intercal}\|_\infty\\
&=O(\sqrt{Ts^2n^2\kappa^3(W)\varpi^2})\max_{1\leq i,j\leq n}\left|(\bar{x}-\mu)_i(\bar{x}-\mu)_j\right|=O_p(\sqrt{Ts^2n^2\kappa^3(W)\varpi^2})\log n/T\\
&=O_p\left(\sqrt{\frac{\log^4 n\cdot n^2\kappa^3(W)\varpi^2}{T}}\right)=o_p(1),
\end{aligned}$$

where the third last equality is due to (8.20), the last equality is due to Assumption 3.2(ii), and the second equality is due to (A.12), (A.13), (A.19), and the fact that

$$\begin{aligned}
\left\|\left.\frac{\partial \operatorname{vec}\Theta}{\partial \operatorname{vec}\Sigma}\right|_{\Sigma=\tilde{\Sigma}_T^{(i)}}\right\|_{\ell_2}&=\left\|\left.\frac{\partial \operatorname{vec}\Theta}{\partial \operatorname{vec}\Sigma}\right|_{\Sigma=\tilde{\Sigma}_T^{(i)}}-P(D^{-1/2}\otimes D^{-1/2})\right\|_{\ell_2}+\left\|P(D^{-1/2}\otimes D^{-1/2})\right\|_{\ell_2}\\
&=O_p\left(\sqrt{\frac{n}{T}}\right)+O(1)=O_p(1).
\end{aligned}$$

8.3.5 $\hat{t}_2 = o_p(1)$

Write
$$\hat{t}_2 = \frac{\sqrt{T}\sqrt{s^2 n\kappa^2(W)\varpi} c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ \operatorname{vec} O_p(\|\hat{\Theta}_T - \Theta\|_{\ell_2}^2)}{\sqrt{s^2 n\kappa^2(W)\varpi \hat{G}_T}}.$$
Since the denominator of the preceding equation is bounded away from zero by an absolute constant with probability approaching one by (8.22) and that $s^2 n\kappa^2(W)\varpi|\hat{G}_T - G| = o_p(1)$, it suffices to show
$$\sqrt{T}\sqrt{s^2 n\kappa^2(W)\varpi} c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ \operatorname{vec} O_p(\|\hat{\Theta}_T - \Theta\|_{\ell_2}^2) = o_p(1).$$
This is straightforward:
$$\begin{aligned}
&|\sqrt{Ts^2 n\kappa^2(W)\varpi} c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ \operatorname{vec} O_p(\|\hat{\Theta}_T - \Theta\|_{\ell_2}^2)| \\
&\leq \sqrt{Ts^2 n\kappa^2(W)\varpi} \|c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+\|_2 \|\operatorname{vec} O_p(\|\hat{\Theta}_T - \Theta\|_{\ell_2}^2)\|_2 \\
&= O(\sqrt{Ts^2\kappa^3(W)\varpi^2}) \|O_p(\|\hat{\Theta}_T - \Theta\|_{\ell_2}^2)\|_F = O(\sqrt{Tns^2\kappa^3(W)\varpi^2}) \|O_p(\|\hat{\Theta}_T - \Theta\|_{\ell_2}^2)\|_{\ell_2} \\
&= O(\sqrt{Tns^2\kappa^3(W)\varpi^2}) O_p(\|\hat{\Theta}_T - \Theta\|_{\ell_2}^2) = O_p\left(\sqrt{\frac{n^3 s^2\kappa^3(W)\varpi^2}{T}}\right) = o_p(1),
\end{aligned}$$
where the last equality is due to Assumption 3.2(ii). □

8.4

Proof of Proposition 4.1. At each step, we take the symmetry of $\Omega(\theta)$ into account.
$$\begin{aligned}
&d\ell_{T,D}(\theta, \mu) \\
&= -\frac{T}{2} d\log\left|D^{1/2}\exp(\Omega)D^{1/2}\right| - \frac{T}{2} d\mathrm{tr}\left(\frac{1}{T}\sum_{t=1}^{T}(x_t - \mu)^\intercal D^{-1/2}[\exp(\Omega)]^{-1}D^{-1/2}(x_t - \mu)\right) \\
&= -\frac{T}{2} d\log\left|D^{1/2}\exp(\Omega)D^{1/2}\right| - \frac{T}{2} d\mathrm{tr}\left(D^{-1/2}\tilde{\Sigma}_T D^{-1/2}[\exp(\Omega)]^{-1}\right) \\
&= -\frac{T}{2}\mathrm{tr}\left(\left[D^{1/2}\exp(\Omega)D^{1/2}\right]^{-1} D^{1/2} d\exp(\Omega) D^{1/2}\right) - \frac{T}{2} d\mathrm{tr}\left(D^{-1/2}\tilde{\Sigma}_T D^{-1/2}[\exp(\Omega)]^{-1}\right) \\
&= -\frac{T}{2}\mathrm{tr}\left(\left[\exp(\Omega)\right]^{-1} d\exp(\Omega)\right) - \frac{T}{2}\mathrm{tr}\left(D^{-1/2}\tilde{\Sigma}_T D^{-1/2} d[\exp(\Omega)]^{-1}\right) \\
&= -\frac{T}{2}\mathrm{tr}\left(\left[\exp(\Omega)\right]^{-1} d\exp(\Omega)\right) + \frac{T}{2}\mathrm{tr}\left(D^{-1/2}\tilde{\Sigma}_T D^{-1/2}[\exp(\Omega)]^{-1} d\exp(\Omega)[\exp(\Omega)]^{-1}\right) \\
&= \frac{T}{2}\mathrm{tr}\left(\left\{[\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}[\exp(\Omega)]^{-1} - \left[\exp(\Omega)\right]^{-1}\right\} d\exp(\Omega)\right) \\
&= \frac{T}{2}\left[\operatorname{vec}\left(\left\{[\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}[\exp(\Omega)]^{-1} - \left[\exp(\Omega)\right]^{-1}\right\}^\intercal\right)\right]^\intercal \operatorname{vec} d\exp(\Omega) \\
&= \frac{T}{2}\left[\operatorname{vec}\left([\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}[\exp(\Omega)]^{-1} - \left[\exp(\Omega)\right]^{-1}\right)\right]^\intercal \operatorname{vec} d\exp(\Omega),
\end{aligned}$$
where in the second equality we used the definition of $\tilde{\Sigma}_T$, 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 $\mathrm{tr}(AB) = (\operatorname{vec}[A^\intercal])^\intercal \operatorname{vec} B$, and the eighth equality is due to that matrix function preserves symmetry and we can interchange inverse and transpose operators.

The following differential of matrix exponential can be inferred from (10.15) in Higham (2008) p238:
$$d\exp(\Omega) = \int_0^1 e^{(1-t)\Omega}(d\Omega)e^{t\Omega}dt.$$

Therefore,

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

Hence,

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

and

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

Now we derive the Hessian matrix.

$$dy = \frac{T}{2}E^{\intercal}D_n^{\intercal}(d\Psi_1)\Psi_2+\frac{T}{2}E^{\intercal}D_n^{\intercal}\Psi_1d\Psi_2 = \frac{T}{2}(\Psi_2^{\intercal}\otimes E^{\intercal}D_n^{\intercal})\operatorname{vec}d\Psi_1+\frac{T}{2}E^{\intercal}D_n^{\intercal}\Psi_1d\Psi_2. \quad (8.28)$$

Consider $d\Psi_1$ first.

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

where

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

Therefore,

$$\begin{aligned}\operatorname{vec}d\Psi_1 &= \int_0^1 \operatorname{vec}\left(A\otimes e^{(1-t)\Omega}\right)dt+\int_0^1 \operatorname{vec}\left(e^{t\Omega}\otimes B\right)dt\\ &= \int_0^1 P_K\left(\operatorname{vec}A\otimes\operatorname{vec}e^{(1-t)\Omega}\right)dt+\int_0^1 P_K\left(\operatorname{vec}e^{t\Omega}\otimes\operatorname{vec}B\right)dt\\ &= \int_0^1 P_K\left(I_{n^2}\otimes\operatorname{vec}e^{(1-t)\Omega}\right)\operatorname{vec}Adt+\int_0^1 P_K\left(\operatorname{vec}e^{t\Omega}\otimes I_{n^2}\right)\operatorname{vec}Bdt\\ &= \int_0^1 P_K\left(I_{n^2}\otimes\operatorname{vec}e^{(1-t)\Omega}\right)\int_0^1 e^{st\Omega}\otimes e^{(1-s)t\Omega}ds\cdot\operatorname{vec}d(t\Omega)dt\\ &\quad+\int_0^1 P_K\left(\operatorname{vec}e^{t\Omega}\otimes I_{n^2}\right)\int_0^1 e^{s(1-t)\Omega}\otimes e^{(1-s)(1-t)\Omega}ds\cdot\operatorname{vec}d((1-t)\Omega)dt\\ &= \int_0^1 P_K\left(I_{n^2}\otimes\operatorname{vec}e^{(1-t)\Omega}\right)\int_0^1 e^{st\Omega}\otimes e^{(1-s)t\Omega}ds\cdot tdtD_nEd\theta\\ &\quad+\int_0^1 P_K\left(\operatorname{vec}e^{t\Omega}\otimes I_{n^2}\right)\int_0^1 e^{s(1-t)\Omega}\otimes e^{(1-s)(1-t)\Omega}ds\cdot(1-t)dtD_nEd\theta \quad (8.29)\end{aligned}$$

where $P_K := I_n \otimes K_{n,n} \otimes I_n$, the second equality is due to Lemma A.8 in Appendix A.5. We now consider $d\Psi_2$.

$$\begin{aligned}
d\Psi_2 &= d\,\text{vec}\left([\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}[\exp(\Omega)]^{-1} - \left[\exp(\Omega)\right]^{-1}\right) \\
&= \text{vec}\left(d[\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}[\exp(\Omega)]^{-1}\right) \\
&\quad + \left([\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}d[\exp(\Omega)]^{-1}\right) - \text{vec}\left(d\left[\exp(\Omega)\right]^{-1}\right) \\
&= \text{vec}\left(-[\exp(\Omega)]^{-1}d\exp(\Omega)[\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}[\exp(\Omega)]^{-1}\right) \\
&\quad + \text{vec}\left(-[\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}[\exp(\Omega)]^{-1}d\exp(\Omega)[\exp(\Omega)]^{-1}\right) \\
&\quad + \text{vec}\left([\exp(\Omega)]^{-1}d\exp(\Omega)[\exp(\Omega)]^{-1}\right) \\
&= \left([\exp(\Omega)]^{-1} \otimes [\exp(\Omega)]^{-1}\right)\text{vec}\,d\exp(\Omega) \\
&\quad - \left([\exp(\Omega)]^{-1} \otimes [\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}[\exp(\Omega)]^{-1}\right)\text{vec}\,d\exp(\Omega) \\
&\quad - \left([\exp(\Omega)]^{-1}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}[\exp(\Omega)]^{-1} \otimes [\exp(\Omega)]^{-1}\right)\text{vec}\,d\exp(\Omega) \qquad (8.30)
\end{aligned}$$

Substituting (8.29) and (8.30) into (8.28) yields the result:

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

□

8.5

Proposition 8.4. *Suppose Assumptions 3.1(i), 3.2(i) and 3.3 hold. Then*

(i)

$$\Xi = \int_0^1\int_0^1 \Theta^{t+s-1} \otimes \Theta^{1-t-s}dtds$$

has minimum eigenvalue bounded away from zero by an absolute constant and maximum eigenvalue bounded from above by an absolute constant.

(ii)

$$\hat{\Xi}_T = \int_0^1\int_0^1 \hat{\Theta}_T^{t+s-1} \otimes \hat{\Theta}_T^{1-t-s}dtds$$

has minimum eigenvalue bounded away from zero by an absolute constant and maximum eigenvalue bounded from above by an absolute constant with probability approaching 1.

(iii)

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

(iv)

$$\|\Psi_1\|_{\ell_2} = \left\| \int_0^1 e^{t\Omega(\theta)} \otimes e^{(1-t)\Omega(\theta)} dt \right\|_{\ell_2} = O(1).$$

Proof. The proofs for the first two parts are the same, so we only give one for part (i). Under assumptions of this proposition, we can invoke Proposition A.5(i) to have eigenvalues of Θ to be bounded away from zero and from above by absolute positive constants. Let $\lambda_1, \ldots, \lambda_n$ denote these eigenvalues. Suppose $\Theta = Q^\intercal \text{diag}(\lambda_1, \ldots, \lambda_n) Q$ (orthogonal diagonalization). By definition of matrix function, we have

$$\begin{aligned}
\Theta^{(t+s-1)} &= Q^\intercal \text{diag}(\lambda_1^{(t+s-1)}, \ldots, \lambda_n^{(t+s-1)}) Q \\
\Theta^{(1-s-t)} &= Q^\intercal \text{diag}(\lambda_1^{(1-s-t)}, \ldots, \lambda_n^{(1-s-t)}) Q \\
\Theta^{(t+s-1)} \otimes \Theta^{(1-s-t)} &= (Q \otimes Q)^\intercal \left[\text{diag}(\lambda_1^{(t+s-1)}, \ldots, \lambda_n^{(t+s-1)}) \otimes \text{diag}(\lambda_1^{(1-s-t)}, \ldots, \lambda_n^{(1-s-t)}) \right] (Q \otimes Q) \\
&=: (Q \otimes Q)^\intercal M_2 (Q \otimes Q),
\end{aligned}$$

where M_2 is an $n^2 \times n^2$ diagonal matrix whose $[(i-1)n+j]$th diagonal entry is

$$\begin{cases} 1 & \text{if } i = j \\ 1 & \text{if } i \neq j, \lambda_i = \lambda_j \\ \left(\frac{\lambda_i}{\lambda_j}\right)^{s+t-1} & \text{if } i \neq j, \lambda_i \neq \lambda_j \end{cases}$$

for $i, j = 1, \ldots, n$.

Thus

$$\int_0^1 \int_0^1 \Theta^{t+s-1} \otimes \Theta^{1-t-s} dt ds = (Q \otimes Q)^\intercal \int_0^1 \int_0^1 M_2 dt ds (Q \otimes Q)$$

where $\int_0^1 \int_0^1 M_2 ds dt$ is an $n^2 \times n^2$ diagonal matrix whose $[(i-1)n+j]$th diagonal entry is

$$\begin{cases} 1 & \text{if } i = j \\ 1 & \text{if } i \neq j, \lambda_i = \lambda_j \\ \frac{1}{\left[\log\left(\frac{\lambda_i}{\lambda_j}\right)\right]^2} \frac{\lambda_j}{\lambda_i} \left[\frac{\lambda_i}{\lambda_j} - 1\right]^2 & \text{if } i \neq j, \lambda_i \neq \lambda_j \end{cases}$$

To see this,

$$\begin{aligned}
&\int_0^1 \int_0^1 \left(\frac{\lambda_i}{\lambda_j}\right)^{s+t-1} ds dt = \frac{\lambda_j}{\lambda_i} \int_0^1 \left(\frac{\lambda_i}{\lambda_j}\right)^s ds \int_0^1 \left(\frac{\lambda_i}{\lambda_j}\right)^t dt \\
&= \frac{\lambda_j}{\lambda_i} \left[\int_0^1 \left(\frac{\lambda_i}{\lambda_j}\right)^s ds\right]^2 = \frac{\lambda_j}{\lambda_i} \left[\left[\frac{\left(\frac{\lambda_i}{\lambda_j}\right)^s}{\log\left(\frac{\lambda_i}{\lambda_j}\right)}\right]_0^1\right]^2 = \frac{1}{\left[\log\left(\frac{\lambda_i}{\lambda_j}\right)\right]^2} \frac{\lambda_j}{\lambda_i} \left[\frac{\lambda_i}{\lambda_j} - 1\right]^2.
\end{aligned}$$

For part (iii), we have

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

First, note that for any $t, s \in [0, 1]$, $\|\hat{\Theta}_T^{t+s-1}\|_{\ell_2}$ and $\|\Theta^{1-t-s}\|_{\ell_2}$ are $O_p(1)$ and $O(1)$, respectively. For example, diagonalize Θ, apply the function $f(x) = x^{1-t-s}$, and take the spectral norm.

The proposition would then follow if we show that

$$
\max_{t,s\in[0,1]} \|\hat{\Theta}_T^{1-t-s} - \Theta^{1-t-s}\|_{\ell_2} = O_p(\sqrt{n/T}), \qquad \max_{t,s\in[0,1]} \|\hat{\Theta}_T^{t+s-1} - \Theta^{t+s-1}\|_{\ell_2} = O_p(\sqrt{n/T}).
$$

It suffices to give a proof for the first equation, as the proof for the second is similar.

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

where the first inequality is due to Theorem A.3 in Appendix A.5, and the second equality is due to the fact that all the eigenvalues of Θ are bounded away from zero and infinity by absolute positive constants. Now use Proposition 3.1 to get

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

The result follows after recognising $\exp(o_p(1)) = O_p(1)$.

For part (iv), since $\Theta = Q^\intercal \mathrm{diag}(\lambda_1, \ldots, \lambda_n) Q$, we have

$$
\Theta^t = Q^\intercal \mathrm{diag}(\lambda_1^t, \ldots, \lambda_n^t) Q, \qquad \Theta^{1-t} = Q^\intercal \mathrm{diag}(\lambda_1^{1-t}, \ldots, \lambda_n^{1-t}) Q.
$$

Then

$$
\begin{aligned}
\Theta^t \otimes \Theta^{1-t} &= (Q \otimes Q)^\intercal \left[\mathrm{diag}(\lambda_1^t, \ldots, \lambda_n^t) \otimes \mathrm{diag}(\lambda_1^{1-t}, \ldots, \lambda_n^{1-t}) \right] (Q \otimes Q) \\
&=: (Q \otimes Q)^\intercal M_3 (Q \otimes Q),
\end{aligned}
$$

where M_3 is an $n^2 \times n^2$ diagonal matrix whose $[(i-1)n + j]$th diagonal entry is

$$
\begin{cases}
1 & \text{if } i = j \\
1 & \text{if } i \neq j, \lambda_i = \lambda_j \\
\lambda_j \left(\frac{\lambda_i}{\lambda_j} \right)^t & \text{if } i \neq j, \lambda_i \neq \lambda_j
\end{cases}
$$

for $i, j = 1, \ldots, n$.

Thus

$$\Psi_1 = \int_0^1 \Theta^t \otimes \Theta^{1-t} dt = (Q \otimes Q)^\intercal \int_0^1 M_3 dt (Q \otimes Q)$$

where $\int_0^1 M_3 dt$ is an $n^2 \times n^2$ diagonal matrix whose $[(i-1)n + j]$th diagonal entry is

$$\begin{cases} 1 & \text{if } i = j \\ 1 & \text{if } i \neq j, \lambda_i = \lambda_j \\ \frac{\lambda_i - \lambda_j}{\log \lambda_i - \log \lambda_j} & \text{if } i \neq j, \lambda_i \neq \lambda_j \end{cases}$$

To see this,

$$\lambda_j \int_0^1 \left(\frac{\lambda_i}{\lambda_j}\right)^t dt = \lambda_j \int_0^1 \left(\frac{\lambda_i}{\lambda_j}\right)^t dt = \lambda_j \left[\frac{\left(\frac{\lambda_i}{\lambda_j}\right)^t}{\log\left(\frac{\lambda_i}{\lambda_j}\right)}\right]_0^1 = \frac{1}{\log\left(\frac{\lambda_i}{\lambda_j}\right)} \lambda_j \left[\frac{\lambda_i}{\lambda_j} - 1\right].$$

□

Proposition 8.5. *Suppose Assumptions 3.1(i), 3.2(i) and 3.3 hold. Then*

(i)

$$\|\hat{\Upsilon}_{T,D} - \Upsilon_D\|_{\ell_2} = O_p\left(sn\sqrt{\frac{n}{T}}\right).$$

(ii)

$$\|\hat{\Upsilon}_{T,D}^{-1} - \Upsilon_D^{-1}\|_{\ell_2} = O_p\left(\varpi^2 s\sqrt{\frac{1}{nT}}\right).$$

Proof. For part (i),

$$\|\hat{\Upsilon}_{T,D} - \Upsilon_D\|_{\ell_2} = \frac{1}{2}\|E^\intercal D_n^\intercal(\hat{\Xi}_T - \Xi)D_n E\|_{\ell_2} \leq \frac{1}{2}\|E^\intercal\|_{\ell_2}\|D_n^\intercal\|_{\ell_2}\|\hat{\Xi}_T - \Xi\|_{\ell_2}\|D_n\|_{\ell_2}\|E\|_{\ell_2}$$
$$= O(1)\|\hat{\Xi}_T - \Xi\|_{\ell_2}\|E\|_{\ell_2}^2 = O_p\left(sn\sqrt{\frac{n}{T}}\right),$$

where the second equality is due to (A.13), and the last equality is due to (A.16) and Proposition 8.4(iii).

For part (ii),

$$\|\hat{\Upsilon}_{T,D}^{-1} - \Upsilon_D^{-1}\|_{\ell_2} = \|\hat{\Upsilon}_{T,D}^{-1}(\Upsilon_D - \hat{\Upsilon}_{T,D})\Upsilon_D^{-1}\|_{\ell_2} \leq \|\hat{\Upsilon}_{T,D}^{-1}\|_{\ell_2}\|\Upsilon_D - \hat{\Upsilon}_{T,D}\|_{\ell_2}\|\Upsilon_D^{-1}\|_{\ell_2}$$
$$= O_p(\varpi^2/n^2)O_p\left(sn\sqrt{\frac{n}{T}}\right) = O_p\left(s\varpi^2\sqrt{\frac{1}{nT}}\right).$$

□

Proof of Theorem 4.1. We first show that $\hat{\Upsilon}_{T,D}$ is invertible with probability approaching 1, so that our estimator $\tilde{\theta}_{T,D} := \hat{\theta}_{T,D} - \hat{\Upsilon}_{T,D}^{-1}\frac{\partial \ell_{T,D}(\hat{\theta}_{T,D},\bar{x})}{\partial \theta^\intercal}/T$ is well defined. It suffices to show that

$\hat{\Upsilon}_{T,D}$ has minimum eigenvalue bounded away from zero by an absolute constant with probability approaching one.

$$\begin{aligned}\text{mineval}(\hat{\Upsilon}_{T,D}) &= \frac{1}{2}\text{mineval}(E^\intercal D_n^\intercal \hat{\Xi}_T D_n E) \geq \text{mineval}(\hat{\Xi}_T)\text{mineval}(D_n^\intercal D_n)\text{mineval}(E^\intercal E)/2 \\ &\geq C\frac{n}{\varpi},\end{aligned}$$

for some absolute positive constant C with probability approaching one, where the second inequality is due to Proposition 8.4(ii) and Assumption 3.3(ii). Hence $\hat{\Upsilon}_{T,D}$ has minimum eigenvalue bounded away from zero by an absolute constant with probability approaching one. Also as a by-product

$$\|\hat{\Upsilon}_{T,D}^{-1}\|_{\ell_2} = \frac{1}{\text{mineval}(\hat{\Upsilon}_{T,D})} = O_p\left(\frac{\varpi}{n}\right) \qquad \|\Upsilon_D^{-1}\|_{\ell_2} = \frac{1}{\text{mineval}(\Upsilon_D)} = O\left(\frac{\varpi}{n}\right). \tag{8.31}$$

From the definition of $\tilde{\theta}_{T,D}$, for any $b \in \mathbb{R}^s$ with $\|b\|_2 = 1$ we can write

$$\begin{aligned}&\sqrt{T}b^\intercal(\hat{\Upsilon}_{T,D})(\tilde{\theta}_{T,D} - \theta) = \sqrt{T}b^\intercal(\hat{\Upsilon}_{T,D})(\hat{\theta}_{T,D} - \theta) - \sqrt{T}b^\intercal\frac{1}{T}\frac{\partial \ell_{T,D}(\hat{\theta}_{T,D}, \bar{x})}{\partial \theta^\intercal} \\ &= \sqrt{T}b^\intercal(\hat{\Upsilon}_{T,D})(\hat{\theta}_{T,D} - \theta) - \sqrt{T}b^\intercal\frac{1}{T}\frac{\partial \ell_{T,D}(\theta, \bar{x})}{\partial \theta^\intercal} - \sqrt{T}b^\intercal\Upsilon_D(\hat{\theta}_{T,D} - \theta) + o_p(1) \\ &= \sqrt{T}b^\intercal(\hat{\Upsilon}_{T,D} - \Upsilon_D)(\hat{\theta}_{T,D} - \theta) - b^\intercal\sqrt{T}\frac{1}{T}\frac{\partial \ell_{T,D}(\theta, \bar{x})}{\partial \theta^\intercal} + o_p(1)\end{aligned}$$

where the second equality is due to Assumption 4.1 and the fact that $\hat{\theta}_{T,D}$ is $C_E\sqrt{n\log n/T}$-consistent. Defining $a^\intercal = b^\intercal(-\hat{\Upsilon}_{T,D})$, we write

$$\begin{aligned}\sqrt{T}\frac{a^\intercal}{\|a\|_2}(\tilde{\theta}_{T,D} - \theta) &= \sqrt{T}\frac{a^\intercal}{\|a\|_2}\hat{\Upsilon}_{T,D}^{-1}(\hat{\Upsilon}_{T,D} - \Upsilon)(\hat{\theta}_{T,D} - \theta) \\ &\quad - \frac{a^\intercal}{\|a\|_2}\hat{\Upsilon}_{T,D}^{-1}\sqrt{T}\frac{1}{T}\frac{\partial \ell_{T,D}(\theta, \bar{x})}{\partial \theta^\intercal} + \frac{o_p(1)}{\|a\|_2}.\end{aligned}$$

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

$$\frac{1}{\|a\|_2} = O_p\left(\frac{\varpi}{n}\right).$$

Thus without loss of generality, we have

$$\sqrt{T}b^\intercal(\tilde{\theta}_{T,D} - \theta) = \sqrt{T}b^\intercal\hat{\Upsilon}_{T,D}^{-1}(\hat{\Upsilon}_{T,D} - \Upsilon_D)(\hat{\theta}_{T,D} - \theta) - b^\intercal\hat{\Upsilon}_{T,D}^{-1}\sqrt{T}\frac{1}{T}\frac{\partial \ell_{T,D}(\theta, \bar{x})}{\partial \theta^\intercal} + o_p(\varpi/n).$$

We now determine a rate for the first term on the right side. This is straightforward

$$\begin{aligned}&\sqrt{T}|b^\intercal\hat{\Upsilon}_{T,D}^{-1}(\hat{\Upsilon}_{T,D} - \Upsilon_D)(\hat{\theta}_{T,D} - \theta)| \leq \sqrt{T}\|b\|_2\|\hat{\Upsilon}_{T,D}^{-1}\|_{\ell_2}\|\hat{\Upsilon}_{T,D} - \Upsilon_D\|_{\ell_2}\|\hat{\theta}_{T,D} - \theta\|_2 \\ &= \sqrt{T}O_p(\varpi/n)snO_p(\sqrt{n/T})O_p(\sqrt{n\varpi\kappa(W)/T}) = O_p\left(\sqrt{\frac{n^2\log^2 n\varpi^3\kappa(W)}{T}}\right),\end{aligned}$$

where the first equality is due to (8.31), Proposition 8.5(i) and the rate of convergence for the minimum distance estimator $\hat{\theta}_T$ ($\hat{\theta}_{T,D}$). Thus

$$\begin{aligned}\sqrt{T}b^\intercal(\tilde{\theta}_{T,D} - \theta) &= -b^\intercal\hat{\Upsilon}_{T,D}^{-1}\sqrt{T}\frac{1}{T}\frac{\partial \ell_{T,D}(\theta, \bar{x})}{\partial \theta^\intercal} + \text{rem} \\ \text{rem} &= O_p\left(\sqrt{\frac{n^2\log^2 n\varpi^3\kappa(W)}{T}}\right) + o_p(\varpi/n)\end{aligned}$$

whence, if we divide by $\sqrt{b^\intercal \hat{\Upsilon}_{T,D}^{-1} b}$, we have

$$\frac{\sqrt{T} b^\intercal (\tilde{\theta}_{T,D} - \theta)}{\sqrt{b^\intercal \hat{\Upsilon}_{T,D}^{-1} b}} = \frac{-b^\intercal \hat{\Upsilon}_{T,D}^{-1} \sqrt{T} \frac{\partial \ell_{T,D}(\theta, \bar{x})}{\partial \theta^\intercal} / T}{\sqrt{b^\intercal \hat{\Upsilon}_{T,D}^{-1} b}} + \frac{\text{rem}}{\sqrt{b^\intercal \hat{\Upsilon}_{T,D}^{-1} b}} =: \hat{t}_{os,D,1} + t_{os,D,2}.$$

Define

$$t_{os,D,1} := \frac{-b^\intercal \Upsilon_D^{-1} \sqrt{T} \frac{\partial \ell_{T,D}(\theta, \mu)}{\partial \theta^\intercal} / T}{\sqrt{b^\intercal \Upsilon_D^{-1} b}}.$$

To prove Theorem 4.1, it suffices to show $-t_{os,D,1} \xrightarrow{d} N(0,1)$, $\hat{t}_{os,D,1} - t_{os,D,1} = o_p(1)$, and $t_{os,D,2} = o_p(1)$.

8.5.1 $-t_{os,D,1} \xrightarrow{d} N(0,1)$

We now prove that $-t_{os,D,1}$ is asymptotically distributed as a standard normal. It is not difficult to show $\mathbb{E}[-t_{os,D,1}] = 0$ and $\text{var}(-t_{os,D,1}) = 1$ under assumption of normality (Assumption 3.1(ii)). Write

$$t_{os,D,1} := \frac{-b^\intercal \Upsilon_D^{-1} \sqrt{T} \frac{\partial \ell_{T,D}(\theta, \mu)}{\partial \theta^\intercal} / T}{\sqrt{b^\intercal \Upsilon_D^{-1} b}} =$$

$$\sum_{t=1}^{T} \frac{\frac{1}{2} b^\intercal \Upsilon_D^{-1} E^\intercal D_n^\intercal \Psi_1 (\Theta^{-1} \otimes \Theta^{-1})(D^{-1/2} \otimes D^{-1/2}) T^{-1/2} \operatorname{vec}\left[(x_t - \mu)(x_t - \mu)^\intercal - \mathbb{E}(x_t - \mu)(x_t - \mu)^\intercal\right]}{\sqrt{b^\intercal \Upsilon_D^{-1} b}}$$

$$=: \sum_{t=1}^{T} U_{os,D,T,n,t}.$$

The proof is very similar to that of $t_{D,1} \xrightarrow{d} N(0,1)$ in Section A.4.1. We now just need to verify the following Lindeberg condition for a double indexed process: for all $\varepsilon > 0$,

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

For any $\gamma > 2$,

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

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

$$b^\intercal \Upsilon_D^{-1} b = 2 b^\intercal \left(E^\intercal D_n^\intercal \Xi D_n E\right)^{-1} b \geq 2\text{mineval}\left(\left(E^\intercal D_n^\intercal \Xi D_n E\right)^{-1}\right)$$
$$= \frac{2}{\text{maxeval}\left(E^\intercal D_n^\intercal \Xi D_n E\right)}.$$

Since,

$$\text{maxeval}\left(E^\intercal D_n^\intercal \Xi D_n E\right) \leq \text{maxeval}(\Xi)\text{maxeval}(D_n^\intercal D_n)\text{maxeval}(E^\intercal E) \leq Csn,$$

for some positive constant C. Thus we have

$$\frac{1}{\sqrt{b^\intercal \Upsilon_D^{-1} b}} = O(\sqrt{sn}). \tag{8.32}$$

Then a sufficient condition for the Lindeberg condition is:

$$\begin{aligned}
&T^{1-\frac{\gamma}{2}}(sn)^{\gamma/2}\cdot \\
&\mathbb{E}\left| b^\intercal \Upsilon_D^{-1} E^\intercal D_n^\intercal \Psi_1 (\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2}) \operatorname{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal - \mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right|^\gamma \\
&= o(1),
\end{aligned} \tag{8.33}$$

for some $\gamma > 2$. We now verify (8.33). We shall be concise as the proof is very similar to that in Section A.4.1.

$$\begin{aligned}
&\mathbb{E}\left| b^\intercal \Upsilon_D^{-1} E^\intercal D_n^\intercal \Psi_1 (\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2}) \operatorname{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal - \mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right|^\gamma \\
&\le \| b^\intercal \Upsilon_D^{-1} E^\intercal D_n^\intercal \Psi_1 (\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2})\|^\gamma \mathbb{E}\left\| \operatorname{vec}\left[(x_t-\mu)(x_t-\mu)^\intercal - \mathbb{E}(x_t-\mu)(x_t-\mu)^\intercal\right]\right\|^\gamma \\
&\lesssim \left(\|\Upsilon_D^{-1}\|_{\ell_2}\|E^\intercal\|_{\ell_2}\|D_n^\intercal\|_{\ell_2}\|\Psi_1\|_{\ell_2}\|\Theta^{-1}\otimes\Theta^{-1}\|_{\ell_2}\|D^{-1/2}\otimes D^{-1/2})\|_{\ell_2}\right)^\gamma \mathbb{E}\|(x_t-\mu)(x_t-\mu)^\intercal\|_F^\gamma \\
&= O\left(s\varpi^2/n\right)^{\gamma/2} n^\gamma \left\| \max_{1\le i,j\le n} \left|(x_t-\mu)_i(x_t-\mu)_j\right|\right\|_{L_\gamma}^\gamma \\
&= O\left(s\varpi^2/n\right)^{\gamma/2} n^\gamma O(\log^\gamma n),
\end{aligned}$$

where the second last equality is due to Proposition 8.4(iv), (A.16) and (8.31), and the last equality is due to (A.21). Summing up the rates, we have

$$T^{1-\frac{\gamma}{2}}(sn)^{\gamma/2} O\left(s\varpi^2/n\right)^{\gamma/2} n^\gamma O(\log^\gamma n) = O\left(\frac{n^2\varpi^2\log^4 n}{T^{1-\frac{2}{\gamma}}}\right)^{\gamma/2} = o(1),$$

by Assumption 3.2(iii). Thus, we have verified (8.33).

8.5.2 $\hat{t}_{os,D,1} - t_{os,D,1} = o_p(1)$

We now show that $\hat{t}_{os,D,1} - t_{os,D,1} = o_p(1)$. Let $A_{os,D}$ and $\hat{A}_{os,D}$ denote the numerators of $t_{os,D,1}$ and $\hat{t}_{os,D,1}$, respectively.

$$\hat{t}_{os,D,1} - t_{os,D,1} = \frac{\hat{A}_{os,D}}{\sqrt{b^\intercal \hat{\Upsilon}_{T,D}^{-1} b}} - \frac{A_{os,D}}{\sqrt{b^\intercal \Upsilon_D^{-1} b}} = \frac{\sqrt{sn}\hat{A}_{os,D}}{\sqrt{snb^\intercal \hat{\Upsilon}_{T,D}^{-1} b}} - \frac{\sqrt{sn}A_{os,D}}{\sqrt{snb^\intercal \Upsilon_D^{-1} b}}$$

Since we have already shown in (8.32) that $snb^\intercal \Upsilon_D^{-1} b$ is bounded away from zero by an absolute constant, it suffices to show the denominators as well as numerators of $\hat{t}_{os,D,1}$ and $t_{os,D,1}$ are asymptotically equivalent.

8.5.3 Denominators of $\hat{t}_{os,D,1}$ and $t_{os,D,1}$

We need to show

$$sn|b^\intercal(\hat{\Upsilon}_{T,D}^{-1} - \Upsilon_D^{-1})b| = o_p(1).$$

This is straightforward.

$$\begin{aligned}
&sn|b^\intercal(\hat{\Upsilon}_{T,D}^{-1} - \Upsilon_D^{-1})b| \le sn\|\hat{\Upsilon}_{T,D}^{-1} - \Upsilon_D^{-1})\|_{\ell_2} = snO_p\left(s\varpi^2\sqrt{\frac{1}{nT}}\right) = O_p\left(s^2\varpi^2\sqrt{\frac{n}{T}}\right) \\
&= o_p(1),
\end{aligned}$$

where the last equality is due to Assumption 3.2(iii).

8.5.4 Numerators of $\hat{t}_{os,D,1}$ and $t_{os,D,1}$

We now show

$$\sqrt{sn}\left|b^\intercal\hat{\Upsilon}_{T,D}^{-1}\sqrt{T}\frac{\partial \ell_{T,D}(\theta,\bar{x})}{\partial\theta^\intercal}/T-b^\intercal\Upsilon_D^{-1}\sqrt{T}\frac{\partial \ell_{T,D}(\theta,\mu)}{\partial\theta^\intercal}/T\right|=o_p(1).$$

Using triangular inequality, we have

$$\begin{aligned}
&\sqrt{sn}\left|b^\intercal\hat{\Upsilon}_{T,D}^{-1}\sqrt{T}\frac{\partial \ell_{T,D}(\theta,\bar{x})}{\partial\theta^\intercal}/T-b^\intercal\Upsilon_D^{-1}\sqrt{T}\frac{\partial \ell_{T,D}(\theta,\mu)}{\partial\theta^\intercal}/T\right|\\
&\leq\sqrt{sn}\left|b^\intercal\hat{\Upsilon}_{T,D}^{-1}\sqrt{T}\frac{\partial \ell_{T,D}(\theta,\bar{x})}{\partial\theta^\intercal}/T-b^\intercal\Upsilon_D^{-1}\sqrt{T}\frac{\partial \ell_{T,D}(\theta,\bar{x})}{\partial\theta^\intercal}/T\right|\\
&\quad+\sqrt{sn}\left|b^\intercal\Upsilon_D^{-1}\sqrt{T}\frac{\partial \ell_{T,D}(\theta,\bar{x})}{\partial\theta^\intercal}/T-b^\intercal\Upsilon_D^{-1}\sqrt{T}\frac{\partial \ell_{T,D}(\theta,\mu)}{\partial\theta^\intercal}/T\right| \qquad (8.34)
\end{aligned}$$

We first show that the first term of (8.34) is $o_p(1)$.

$$\begin{aligned}
&\sqrt{sn}\left|b^\intercal(\hat{\Upsilon}_{T,D}^{-1}-\Upsilon_D^{-1})\sqrt{T}\frac{\partial \ell_{T,D}(\theta,\bar{x})}{\partial\theta^\intercal}/T\right|\\
&=\sqrt{sn}\left|b^\intercal(\hat{\Upsilon}_{T,D}^{-1}-\Upsilon_D^{-1})\sqrt{T}\frac{1}{2}E^\intercal D_n^\intercal\Psi_1(\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}(\hat{\Sigma}_T-\Sigma)\right|\\
&\lesssim\sqrt{sn}\|\hat{\Upsilon}_{T,D}^{-1}-\Upsilon_D^{-1}\|_{\ell_2}\sqrt{T}\|E^\intercal\|_{\ell_2}\|\hat{\Sigma}_T-\Sigma\|_F\\
&\lesssim\sqrt{sn}\varpi^2 s\sqrt{1/(nT)}\sqrt{T}\sqrt{sn}\sqrt{n}\|\hat{\Sigma}_T-\Sigma\|_{\ell_2}\lesssim\sqrt{sn}\varpi^2 s\sqrt{1/(nT)}\sqrt{T}\sqrt{sn}\sqrt{n}\sqrt{n/T}\\
&=O_p\left(\sqrt{\frac{n^3s^4\varpi^4}{T}}\right)=o_p(1),
\end{aligned}$$

where the last equality is due to Assumption 3.2(iii).

We now show that the second term of (8.34) is $o_p(1)$.

$$\begin{aligned}
&\sqrt{sn}\left|b^\intercal\Upsilon_D^{-1}\sqrt{T}\left(\frac{\partial \ell_{T,D}(\theta,\bar{x})}{\partial\theta^\intercal}/T-\frac{\partial \ell_{T,D}(\theta,\mu)}{\partial\theta^\intercal}/T\right)\right|\\
&=\sqrt{sn}\left|b^\intercal\Upsilon_D^{-1}\sqrt{T}\frac{1}{2}E^\intercal D_n^\intercal\Psi_1(\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}(\hat{\Sigma}_T-\tilde{\Sigma}_T)\right|\\
&=O(\sqrt{sn})\|\Upsilon_D^{-1}\|_{\ell_2}\sqrt{T}\|E\|_{\ell_2}\|\hat{\Sigma}_T-\tilde{\Sigma}_T\|_F=O_p(\sqrt{sn})\frac{\varpi}{n}\sqrt{T}\sqrt{sn}n\frac{\log n}{T}\\
&=O_p\left(\sqrt{\frac{\log^4 n\cdot n^2\varpi^2}{T}}\right)=o_p(1),
\end{aligned}$$

where the third last equality is due to (8.20), and the last equality is due to Assumption 3.2(iii).

8.5.5 $t_{os,D,2}=o_p(1)$

To prove $t_{os,D,2}=o_p(1)$, it suffices to show that $\sqrt{sn}|\text{rem}|=o_p(1)$. This is delivered by Assumption 3.2(iii). $\square$

8.6

Proof of Proposition 3.2. We only give a proof for part (i), as that for part (ii) is similar. Note that under H_0,

$$\begin{aligned}
\sqrt{T}g_{T,D}(\theta)&=\sqrt{T}[\operatorname{vech}(\log\hat{\Theta}_{T,D})-E\theta]=\sqrt{T}[\operatorname{vech}(\log\hat{\Theta}_{T,D})-\operatorname{vech}(\log\Theta)]\\
&=\sqrt{T}D_n^+\operatorname{vec}(\log\hat{\Theta}_{T,D}-\log\Theta).
\end{aligned}$$

Thus we can adopt the same method as in Theorem 3.1 to establish the asymptotic distribution of $\sqrt{T}g_{T,D}(\theta)$. In fact, it will be much simpler here because we fixed n. We should have

$$\sqrt{T}g_{T,D}(\theta) \xrightarrow{d} N(0,S), \qquad S := D_n^+ H(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})HD_n^{+\intercal}, \tag{8.35}$$

where S is positive definite given the assumptions of this proposition. The closed-form solution for $\hat{\theta}_T = \hat{\theta}_{T,D}$ has been given in (3.3), but this is not important. We only need that $\hat{\theta}_{T,D}$ sets the first derivative of the objective function to zero:

$$E^\intercal W g_{T,D}(\hat{\theta}_{T,D}) = 0. \tag{8.36}$$

Notice that

$$g_{T,D}(\hat{\theta}_{T,D}) - g_{T,D}(\theta) = -E(\hat{\theta}_{T,D} - \theta). \tag{8.37}$$

Pre-multiply (8.37) by $\frac{\partial g_{T,D}(\hat{\theta}_{T,D})}{\partial \theta^\intercal}W = -E^\intercal W$ to give

$$-E^\intercal W[g_{T,D}(\hat{\theta}_{T,D}) - g_{T,D}(\theta)] = E^\intercal W E(\hat{\theta}_{T,D} - \theta),$$

whence we obtain

$$\hat{\theta}_{T,D} - \theta = -(E^\intercal W E)^{-1}E^\intercal W[g_{T,D}(\hat{\theta}_{T,D}) - g_{T,D}(\theta)]. \tag{8.38}$$

Substitute (8.38) into (8.37)

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

where the second equality is due to (8.36). Using (8.35), we have

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

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

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

Thus

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

because $\hat{S}_{T,D}$ is a consistent estimate of S given (A.12) and Proposition 8.2, which hold under the assumptions of this proposition. The asymptotic covariance matrix in the preceding display is idempotent and has rank $n(n+1)/2 - s$. Thus, under H_0,

$$Tg_{T,D}(\hat{\theta}_{T,D})^\intercal \hat{S}_{T,D}^{-1}g_{T,D}(\hat{\theta}_{T,D}) \xrightarrow{d} \chi^2_{n(n+1)/2-s}.$$

□

Proof of Corollary 3.3. We only give a proof for part (i), as that for part (ii) is similar. From (3.5) and the Slutsky lemma, we have for every fixed n (and hence v and s)

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

as $T \to \infty$. Then invoke Lemma A.9 in Appendix A.5

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

as $n \to \infty$ under H_0. Next invoke Lemma A.10 in Appendix A.5, there exists a sequence $n = n_T$ such that

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

as $T \to \infty$. □

Proof of Corollary 3.1. Theorem 3.1 and a result we proved before, namely,

$$|\hat{G}_{T,D} - G_D| = |c^\intercal \hat{J}_{T,D} c - c^\intercal J_D c| = o_p\left(\frac{1}{sn\kappa(W)}\right), \tag{8.39}$$

imply

$$\sqrt{T} c^\intercal (\hat{\theta}_{T,D} - \theta^0) \xrightarrow{d} N(0, c^\intercal J_D c). \tag{8.40}$$

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

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

so we can invoke (8.40) with $c = Ab/\|Ab\|_2$:

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

which is equivalent to

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

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

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

Since we have shown in the mathematical display above (A.15) that J_D is positive definite and A has full-column rank, $A^\intercal J_D A$ is positive definite and its negative square root exists. Hence,

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

Next from (8.39),

$$\left|b^\intercal B b\right| := \left|b^\intercal A^\intercal \hat{J}_{T,D} A b - b^\intercal A^\intercal J_D A b\right| = o_p\left(\frac{1}{sn\kappa(W)}\right) \|Ab\|_2^2 \leq o_p\left(\frac{1}{sn\kappa(W)}\right) \|A\|_{\ell_2}^2 \|b\|_2^2.$$

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

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

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

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

Then

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

Thus we proved

$$B = A^{\intercal}\hat{J}_{T,D}A - A^{\intercal}J_D A = o_p(1),$$

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

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

□

Proposition 8.6. *For any positive definite matrix Θ,*

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

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

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

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

Substitute the preceding equation into the second last

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

Since E is arbitrary, the result follows. □

Example 8.1. *In the special case of normality, $V = 2D_n D_n^+(\Sigma \otimes \Sigma)$ (Magnus and Neudecker (1986) Lemma 9). Then G_D could be simplified into*

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

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

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

References

HIGHAM, N. J. (2008): *Functions of Matrices: Theory and Computation*, SIAM.

MAGNUS, J. R. AND H. NEUDECKER (1986): "Symmetry, 0-1 Matrices and Jacobians a Review," *Econometric Theory*, 157–190.

PHILLIPS, P. C. AND H. R. MOON (1999): "Linear Regression Limit Theory for Nonstationary Panel Data," *Econometrica*, 67, 1057–1111.

VAN DER VAART, A. AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*, Springer.