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

April 16, 2019

Abstract

We propose a Kronecker product model for correlation or covariance matrices in the large dimensional 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 rates of convergence and central limit theorems (CLT) for our estimators in the large dimensional case. A specification test and tools for Kronecker product model selection and inference are provided. In a Monte Carlo study where a Kronecker product model is correctly specified, our estimators exhibit superior performance. In an empirical application to portfolio choice for S&P500 daily returns, we demonstrate that certain Kronecker product models are good approximations to the general covariance matrix.

Keywords: Correlation matrix, Kronecker product, matrix logarithm; multiway array data; portfolio choice.

JEL classification C55; C58; G11.

*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.

‡Corresponding author. Fanhai International School of Finance and School of Economics, Fudan University, 220 Handan Road, Yangpu District, Shanghai, 200433, China. 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 assumed that some observed variables are related to the key unobserved traits through a factor model, and then use the covariance matrix of the observed variables to deduce properties of the latent traits. Anderson (1984) is a classic statistical reference that studies the estimation of covariance matrices and hypotheses testing about them in the low dimensional case (i.e., the dimension of the covariance matrix, n, is small compared with the sample size T).

More recent work has considered the case where n is large along with T. This is because 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 when compared to zero but still less than one, we call this the *large dimensional* case in this article. (We reserve the phrase "the high dimensional case" for $n > T$.) 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 alone). Standard statistical methods such as principal component analysis (PCA) and canonical-correlation analysis (CCA), do not directly generalize to the large dimensional 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 to be estimated, or to use some device, such as shrinkage or a factor model, to reduce dimension. Most of this literature assumes i.i.d. data.

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

$$\Delta^* = \Delta_1^* \otimes \Delta_2^* \otimes \cdots \otimes \Delta_v^*, \tag{1.1}$$

where Δ_j^* is an $n_j \times n_j$ dimensional real symmetric, positive definite *sub-matrix* such that $n = n_1 \times \cdots \times n_v$. We require that $n_j \in \mathbb{Z}$ and $n_j \geq 2$ for all j; the $\{n_j\}_{j=1}^v$ need not be distinct. We suppose that Δ is the covariance or correlation matrix of an observable series with sample size T and $\{\Delta^*\}$ is a model for Δ.

We study the Kronecker product model in the large dimensional case. Since n tends to infinity in the joint asymptotics, there are two main cases: (1) $n_j \to \infty$ for $j = 1, \ldots, v$ and v is fixed; (2) $\{n_j\}_{j=1}^v$ are all fixed and $v \to \infty$. We shall study case (2) in detail because of its dimensionality reduction property. In this case, the number of parameters of a Kronecker product model grows *logarithmically* with n. In particular, we 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*.

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, which in the simplest case are obtainable in closed form. In the large dimensional case the eigenvalue distribution can be quite general, and there is no spikedness property as in strict factor models (Johnstone and Onatskiy (2018)). The inverse covariance matrix, its determinant, and other key quantities are easily obtained from the corresponding quantities of the sub-matrices, which facilitates computation and analysis.

We primarily focus on correlation matrices rather than covariance matrices. This is partly because the asymptotic theory for the correlation matrix model nests that for the covariance matrix model, and partly because this will allow us to adopt a more flexible approach to approximating a general covariance matrix: We can allow the diagonal elements of the covariance matrix to be unrestricted (and they can be estimated 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. To avoid confusion, we would like to remark that if a Kronecker product model is correctly specified for a correlation matrix, its corresponding covariance matrix need not have a Kronecker product structure, and vice versa. In other words, log sparsity on a correlation matrix does not necessarily imply that its corresponding covariance matrix has log sparsity, and vice versa.

We show that the logarithm of a Kronecker product model is linear in its unknown parameters. We use this as a basis to propose a minimum distance (MD) estimator that is in closed form. We establish a crude upper bound rate of convergence for the MD estimator under the joint asymptotics, but we anticipate that this bound could be improved with better technology and we leave this for future research. There is a large literature on the optimal rate of convergence for estimation of high-dimensional covariance matrices 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. In addition, we allow for weakly dependent data, whereas the above cited papers all assume i.i.d. structures.

Next, we 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 discuss in Section 2 what kind of data gives rise to a Kronecker product model. However, a given covariance or correlation matrix might not exactly correspond to a Kronecker product; in which case a Kronecker product model is misspecified, so $\Delta \notin \{\Delta^*\}$. The previous literature on Kronecker product models did not touch this, 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. We show that for a given Kronecker product model there always exists a member in it that is closest to the covariance or correlation matrix in some sense to be made precise shortly.

We provide some simulation evidence that the Kronecker product model works very well when it is correctly specified. In the empirical study, we apply the Kronecker product model to S&P500 daily stock returns and compare it with Ledoit and Wolf (2004)'s linear shrinkage estimator as well as with Ledoit and Wolf (2017)'s direct nonlinear shrinkage estimator. We find that the minimum variance portfolio implied by a Kronecker product model is almost as good as that constructed from Ledoit and Wolf (2004)'s linear shrinkage estimator. In future work we aim to improve the practical performance of our method by combining it with other approaches such as factor models and by improving the estimation methodology.

1.1 Literature Review

The Kronecker product model has been previously considered in the psychometric literature (see Campbell and O'Connell (1967), Swain (1975), Cudeck (1988), Verhees and Wansbeek (1990) etc). In a multitrait-multimethod (MTMM) context, "multi-mode" data give rise to a Kronecker product model naturally (we will further discuss this in Section 2). Verhees and Wansbeek (1990) outlined several estimation methods of the model based on the least squares and maximum likelihood principles, and provided large sample variances under assumptions of Gaussianity and fixed n. There is a growing Bayesian 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), and Yin and Li (2012). In this literature, they also work with fixed n.

In the spatial literature, there are a number of studies that consider a Kronecker product structure for the correlation matrix of a random field, see for example Loh and Lam (2000).

This article is the first one studying Kronecker product models in the large dimensional case. Our work is also among the first exploiting 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. Even if we look at covariance matrices for the purpose of comparison, the Kronecker product model imposes different sparsity restrictions - as 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 look at different estimators.

1.2 Roadmap

The rest of the article 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 model as well as an empirical application. Section 7 concludes. Major proofs are to be found in Appendix; the remaining proofs are put in Supplementary Material (SM in what follows).

1.3 Notation

Let A be an $m \times n$ matrix. Let $\operatorname{vec} A$ denote the 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 $n(n-1)/2$ supradiagonal 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 generalized inverse of D_n. In particular, $D_n^+ = (D_n^\intercal D_n)^{-1} D_n^\intercal$ because

D_n is full-column-rank. We use $\text{vecl}(A)$ to denote the vectorization operator of the lower off-diagonal elements of A (so this operator excludes the diagonal elements unlike the related $\text{vech}(\cdot)$ operator).

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. Let $\text{maxeval}(\cdot)$ and $\text{mineval}(\cdot)$ 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 $\|A\|_F := [\text{tr}(A^\intercal A)]^{1/2} \equiv [\text{tr}(AA^\intercal)]^{1/2} \equiv \|\text{vec}\, A\|_2$ and $\|A\|_{\ell_2} := \max_{\|x\|_2=1} \|Ax\|_2 \equiv \sqrt{\text{maxeval}(A^\intercal A)}$ denote the Frobenius norm and spectral norm (ℓ_2 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 $n\times n$ real random matrices X_T and Y_T. Notation $X_T = O_p(\|Y_T\|)$, where $\|\cdot\|$ is some matrix norm, means that for every real $\varepsilon > 0$, there exist $M_\varepsilon > 0$, $N_\varepsilon > 0$ and $T_\varepsilon > 0$ such that for all $n > N_\varepsilon$ and $T > T_\varepsilon$, $\mathbb{P}(\|X_T\|/\|Y_T\| > M_\varepsilon) < \varepsilon$. Notation $X_T = o_p(\|Y_T\|)$, where $\|\cdot\|$ is some matrix norm, means that $\|X_T\|/\|Y_T\| \xrightarrow{p} 0$ as $n, T \to \infty$ simultaneously. Landau notation in this article, unless otherwise stated, should be interpreted in the sense that $n, T \to \infty$ simultaneously.

Let $a \vee b$ and $a \wedge b$ denote $\max(a, b)$ and $\min(a, b)$, respectively. 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. Notation $\sigma(\cdot)$ defines sigma algebra.

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 The Kronecker Product Model

In this section we provide more details on the model. Consider an n-dimensional weakly stationary time series vector y_t, where $\mu := \mathbb{E}y_t$ and covariance matrix $\Sigma := \mathbb{E}[(y_t - \mu)(y_t - \mu)^\intercal]$. Let D be the diagonal matrix containing the diagonal entries of Σ.[1] The correlation matrix of y_t, denoted Θ, is $\Theta := D^{-1/2}\Sigma D^{-1/2}$. A Kronecker product model for the covariance or correlation matrix is given by (1.1). The factorization $n = n_1 \times \cdots \times n_v$ could be the prime factorization, which exists for any integer n, or it could be an aggregation of that. For example, if $n = 2^8$, one factorization is $2 \times 2 \times \cdots \times 2$, called the minimal factorization, at the other extreme $2^8 \times 1$ is the maximal factorization (we do not consider the maximal factorization in this article). One also has $4 \times 4 \times 4 \times 4$ and $2 \times 16 \times 2 \times 4$ etc. Highly composite numbers such as 2^8 offer many possible factorizations, but if n is not composite or not composite enough, one can add a vector of pseudo variables to the system until the final dimension is composite enough.[2]

Let Δ denote Σ or Θ according to the modelling purpose. If $\Delta \in \{\Delta^*\}$, we say that the Kronecker product model $\{\Delta^*\}$ is *correctly specified*. Otherwise the Kronecker product model $\{\Delta^*\}$ is *misspecified*. We first make clearer when a Kronecker product model is correctly specified (see Verhees and Wansbeek (1990) and Cudeck (1988) for more discussion). A Kronecker product arises when data have some *multiplicative array* structure. For example, suppose that $u_{j,k}$ are error terms in a panel regression model with $j = 1, \ldots, J$ and $k = 1, \ldots, K$. 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 $JK \times 1$ vector containing all the elements of $u_{j,k}$, $\gamma = (\gamma_1, \ldots, \gamma_J)^\intercal$, and $f = (f_1, \ldots, f_K)^\intercal$.

[1] Matrix D should not be confused with the duplication matrix D_n defined in Notation.

[2] It is recommended to add a vector of independent variables $z_t \sim N(0, I_k)$ such that $(y_t^\intercal, z_t^\intercal)^\intercal$ is an $n' \times 1$ random vector with $n' \times n'$ correlation matrix

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

Suppose that γ, f are random, where γ is independent of f, and both vectors have mean zero. Then,

$$\mathbb{E}[uu^\intercal] = \mathbb{E}[\gamma\gamma^\intercal] \otimes \mathbb{E}[ff^\intercal].$$

In this case the covariance matrix of u is a Kronecker product of two sub-matrices. If one dimension were time and the other were firm, then this implies that the covariance matrix of u is the product of a covariance matrix representing cross-sectional dependence and a covariance matrix representing the time series dependence.

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 $u_{i_1,i_2,\ldots,i_v} = \varepsilon_{1,i_1}\varepsilon_{2,i_2}\cdots\varepsilon_{v,i_v}$, $i_j = 1,\ldots,n_j$ for $j = 1,\ldots,v$, 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. Such multiplicative effects may be a valid description of a data generating process.[3]

In earlier versions of this article we emphasized the Kronecker product model for the covariance matrix. We now focus primarily on the correlation matrix for the reasons mentioned in the introduction and leave the diagonal variance matrix D unrestricted. For the present discussion we assume that D (as well as μ) is known. A Kronecker product model for Θ is given by (1.1) with Δ^* and $\{\Delta_j^*\}_{j=1}^v$ replaced by Θ^* and $\{\Theta_j^*\}_{j=1}^v$, respectively. Since Θ is a correlation matrix, this implies that the diagonal entries of Θ_j^* must be the same, although this diagonal entry could differ as j varies (so long as the diagonal entries of the implied Θ^* are one). Without loss of generality, we may impose a normalization constraint that all the diagonal entries of sub-matrices $\{\Theta_j^*\}_{j=1}^v$ are equal to one.

The Kronecker product model substantially reduces the number of parameters to estimate. 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! Since we do not restrict the diagonal matrix we have an additional n variance parameters,[4] so overall the correlation matrix version of the model has more parameters and more flexibility than the covariance matrix version of the model. The Kronecker product model induces sparsity. Specifically, although Θ^* is not sparse, the matrix $\log\Theta^*$ is sparse, where log denotes the (principal) matrix logarithm defined through the eigendecomposition of a real symmetric, positive definite matrix (see Higham (2008) p20 for a definition). This is due to a property of Kronecker products (see Lemma 8.1 in SM 8.1 for derivation), that

$$\log\Theta^* = \log\Theta_1^* \otimes I_{n_2} \otimes \cdots \otimes I_{n_v} + I_{n_1} \otimes \log\Theta_2^* \otimes I_{n_3} \otimes \cdots \otimes I_{n_v} + \cdots + I_{n_1} \otimes I_{n_2} \otimes \cdots \otimes \log\Theta_v^*, \quad (2.1)$$

[3]For example, 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.

[4]These parameters can be estimated in a first step by standard methods.

whence we see that $\log \Theta^*$ has many zero elements, generated by identity sub-matrices.[5] That is, we can write $\text{vech}(\log \Theta^*) = E\theta^*$ for some matrix E of zeros and ones and vector θ^* containing the unrestricted elements of $\log \Theta_1^*, \ldots, \log \Theta_v^*$.

We next discuss some further identification/parameterization issues. First of all, sub-matrices $\{\Theta_j^*\}_{j=1}^v$ are uniquely identified by the following argument based on the architecture of Θ^*. Suppose that $\Theta^* = \tilde{\Theta}_1^* \otimes \cdots \otimes \tilde{\Theta}_v^*$ for other sub-matrices $\{\tilde{\Theta}_j^*\}_{j=1}^v$, with diagonal elements being one, whose dimensions agree with those of $\{\Theta_j^*\}_{j=1}^v$. Let $\rho_{j,k\ell}^*$ and $\tilde{\rho}_{j,k\ell}^*$ denote a typical off-diagonal element of Θ_j^* and $\tilde{\Theta}_j^*$, respectively ($k, \ell = 1, \ldots, n_j$, $k \neq \ell$). Note that $\rho_{j,k\ell}^*$ appears, on its own, in some elements of Θ^*, so does $\tilde{\rho}_{j,k\ell}^*$ in the same positions. We must have $\rho_{j,k\ell}^* = \tilde{\rho}_{j;k\ell}^*$ for all $k, \ell = 1, \ldots, n_j$, $k \neq \ell$ and $j = 1, \ldots, v$. Therefore, sub-matrices $\{\Theta_j^*\}_{j=1}^v$ are identified from Θ^* once $\{n_j\}_{j=1}^v$ are specified, or equivalently $\{\rho_{j,k\ell}^* : k, \ell = 1, \ldots, n_j, k \neq \ell\}_{j=1}^v$ are identified from Θ^* once $\{n_j\}_{j=1}^v$ are specified. We call $\{\rho_{j,k\ell}^* : k, \ell = 1, \ldots, n_j, k \neq \ell\}_{j=1}^v$ the *original parameters* of some member Θ^* in the Kronecker product model. If $\{\Theta_j^*\}_{j=1}^v$ are positive definite correlation matrices, then so is Θ^*.

Second, the matrix logarithm of a correlation matrix has a complicated structure, with its diagonal elements taking any non-positive values and its off-diagonal elements taking any values (Archakov and Hansen (2018) Lemma 2). As an illustration, suppose that

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

then

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

There are $\sum_{j=1}^v n_j(n_j+1)/2$ parameters in $\{\log \Theta_j^*\}_{j=1}^v$; we call these *log parameters* of some member Θ^* in the Kronecker product model. On the other hand, Θ^* has only $\sum_{j=1}^v n_j(n_j-1)/2$ original parameters. For each Θ_j^*, its $n_j(n_j-1)/2$ original parameters completely pin down its $n_j(n_j+1)/2$ 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 the original parameters to the 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).

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*

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

[5] A final property of the Kronecker product model is that it 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

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

Third, there are several ways to achieve identification of $\{\log\Theta_j^*\}_{j=1}^v$ given $\log\Theta^*$ (i.e., identification of log parameters of Θ^*). We outline two methods. The *fill and shrink* method estimates the log parameters without imposing the restrictions implied by that $\{\Theta_j^*\}_{j=1}^v$ being correlation matrices, and then imposes those restrictions afterwards. In this case at the estimation stage, we must impose $v-1$ identification restrictions on the log parameters because in (2.1) the diagonal elements of $\log\Theta^*$ are sums of diagonal elements from $\{\log\Theta_j^*\}_{j=1}^v$ (see Example 8.2 in SM 8.1). There are several ways to impose these $v-1$ identification restrictions. For example, one can set $\mathrm{tr}(\log\Theta_j^*)$ to be some fixed value for $j=1,\ldots,v-1$, or one can set $v-1$ diagonal elements of $\{\log\Theta_j^*\}_{j=1}^v$ to be zero (see Examples 8.1 and 8.2 in SM 8.1 for illustrations). Then in Theorem A.1 in Appendix A.1 we show that there exists an $n(n+1)/2\times s$ full column rank, deterministic matrix E such that

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

where $\theta^*\in\mathbb{R}^s$ are the *unrestricted* log parameters of Θ^*, where $s := \sum_{j=1}^v n_j(n_j+1)/2-(v-1) = O(\log n)$. So far we have not imposed those restrictions that $\{\Theta_j^*\}_{j=1}^v$ are correlation matrices. Nevertheless, $D^{1/2}\exp(\log\Theta^*)D^{1/2}$ will always be a covariance matrix, and one could re-compute the correlation matrix from $D^{1/2}\exp(\log\Theta^*)D^{1/2}$ by re-normalization. Alternatively, one could use minimum distance estimation to shrink $\exp(\log\Theta_j^*)$ to a correlation matrix for $j=1,\ldots,v$ (see SM 8.2 for a discussion).

On the other hand, the *shrink and fill* method identifies a *subset* of unrestricted log parameters and then fills in the remainder afterwards. A recent paper of Archakov and Hansen (2018) proposed a neat way to achieve this. Let $\tilde{\theta}_j^* := \mathrm{vecl}(\log\Theta_j^*)$ and we can identify $\{\tilde{\theta}_j^*\}_{j=1}^v$ from $\mathrm{vecl}(\log\Theta^*)$. Then we can use $\tilde{\theta}_j^*$ to uniquely determine the diagonal elements of $\log\Theta_j^*$ by some function $\phi:\mathbb{R}^{n_j(n_j-1)/2}\to\mathbb{R}^{n_j}$, which can be obtained numerically (in the case $n_j=2$ there exists a closed form, see Example 2.1). Archakov and Hansen (2018) gave a concrete algorithm to do this and established its validity.

We shall use the fill and shrink method in what follows; in particular we set the first diagonal entry of $\log\Theta_j^*$ to zero for $j=1,\ldots,v-1$. To summarise, in order to estimate a correlation matrix Θ using a Kronecker product model Θ^*, there are two approaches. First, one can estimate the original parameters using the principle of maximum likelihood (see Section 4.1) or nonlinear minimum distance. Then form an estimate of Θ^*. Second, one can estimate the unrestricted log parameters θ^* using the principle of minimum distance (see Section 3) or maximum likelihood (see Section 4.1). Form an estimate of $\exp(\log\Theta^*)$ and then recover the estimated correlation matrix using either re-normalization or shrinkage. To study the theoretical properties of a Kronecker product model in large dimension, the second approach is more appealing as log parameters are additive from (2.1) while original parameters are multiplicative in nature; additive objects are easier to analyse theoretically than multiplicative objects.

3 Linear Minimum Distance Estimator

In this section, we define a class of estimators of the (unrestricted) log parameters θ^* of some member in the Kronecker product model (1.1), which are linear in the log sample correlation matrix, and give its asymptotic properties.

3.1 Estimation

We observe a sample $\{y_t\}_{t=1}^T$. Define the sample covariance matrix and sample correlation matrix

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

where $\bar{y} := (1/T)\sum_{t=1}^T y_t$ 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 by Θ^0, which is closest to the correlation matrix Θ in the following sense:

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

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

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

whence we can solve

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

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, we have $\theta^0 = (E^\intercal WE)^{-1}E^\intercal W\operatorname{vech}(\log\Theta) = (E^\intercal WE)^{-1}E^\intercal WE\theta = \theta$. In this case, $\hat{\theta}_T$ is indeed estimating the elements of the correlation matrix Θ.

In practice the MD estimator is easy to compute. The matrix $E^\intercal WE$ is of dimensions $s\times s$ and is highly structured (at least in the diagonal W case). One only needs a user-defined function in some software to generate the matrix E before one can use formula (3.3) to compute the MD estimator.[6]

3.2 Rate of Convergence

We now introduce some assumptions for our theoretical analysis. These conditions are sufficient but far from necessary.

Assumption 3.1.

(i) For all t, for every $a\in\mathbb{R}^n$ with $\|a\|_2 = 1$, there exist absolute constants $K_1 > 1, K_2 > 0, r_1 > 0$ such that[7]

$$\mathbb{E}\left[\exp\left(K_2|a^\intercal y_t|^{r_1}\right)\right] \leq K_1.$$

(ii) The time series $\{y_t\}_{t=1}^T$ are normally distributed.

Assumption 3.2. *There exist absolute constants $K_3 > 0$ and $r_2 > 0$ such that for all $h\in\mathbb{N}$*

$$\alpha(h) \leq \exp\left(-K_3 h^{r_2}\right),$$

where $\alpha(h)$ is the α-mixing (i.e., strong mixing) coefficients of y_t which are defined by $\alpha(0) = 1/2$ and for $h\in\mathbb{N}$

$$\alpha(h) := 2\sup_t \sup_{\substack{A\in\sigma(\cdots,y_{t-1},y_t)\\ B\in\sigma(y_{t+h},y_{t+h+1},\cdots)}} \left|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)\right|,$$

where $\sigma(\cdot)$ defines sigma algebra.

[6]We have written a user-defined function in Matlab which can return E within a few seconds for fairly large n, say, $n = 625$. It is available upon request.

[7]"Absolute constants" mean constants that are independent of both n and T.

Assumption 3.3.

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

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

$$\frac{n^4 \varpi^4 \kappa^6(W)(\log^5 n)\log^2(1+T)}{T} = o(1)$$

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 3.4(ii).

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

(a)

$$\frac{n^4 \varpi^4 \kappa(W)(\log^5 n)\log^2(1+T)}{T} = o(1),$$

(b)

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

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 3.4(ii).

Assumption 3.4.

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

(ii) Suppose

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

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

Assumption 3.1(i) is standard in high-dimensional theoretical work (e.g., Fan, Liao, and Mincheva (2011), Chang, Qiu, Yao, and Zou (2018) etc). In essence it assumes that a random vector has some exponential-type tail probability (c.f. Lemma A.2 in Appendix A.3), which allows us to invoke some concentration inequality such as a version of the Bernstein's inequality (e.g., Theorem A.2 in Appendix A.5). The parameter r_1 restricts the size of the tail of y_t - the smaller r_1, the heavier the tail. When $r_1 = 2$, y_t is said to be *subgaussian*, when $r_1 = 1$, y_t is said to be *subexponential*, and when $0 < r_1 < 1$, y_t is said to be *semiexponential*.

Needless to say, Assumption 3.1(i) is stronger than a finite polynomial moment assumption as it assumes the existence of some exponential moment. In a setting of independent observations, Vershynin (2012) replaced Assumption 3.1(i) with a finite polynomial moment condition and established a rate of convergence for covariance matrices, which is slightly worse than what we have in Theorem 3.1(i) for correlation matrices. For dependent data, relaxation of the subgaussian assumption is currently an active research area in probability theory and statistics. One of the recent work is Wu and Wu (2016) in which they relaxed subgaussianity to a finite polynomial moment condition in high-dimensional linear models with help of Nagaev-type inequalities. Thus Assumption 3.1(i) is likely to be relaxed when new probabilistic tools become available.

Assumption 3.1(ii), which will only be used in Section 4 for one-step estimation, implies Assumption 3.1(i) with $0 < r_1 \leq 2$. Assumption 3.1(ii) is not needed for the minimum distance estimation (Theorem 3.2 or 3.3) though.

Assumption 3.2 assumes that $\{y_t\}_{t=1}^{T}$ is alpha mixing (i.e., strong mixing) because $\alpha(h) \to 0$ as $h \to \infty$. In fact, we require it to decrease at an exponential rate. The bigger r_2 gets, the faster the decay rate and the less dependence $\{y_t\}_{t=1}^T$ exhibits. This assumption covers a wide range of time series. It is well known that both classical ARMA and GARCH processes are strong mixing with mixing coefficients which decrease to zero at an exponential rate (see Section 2.6.1 of Fan and Yao (2003) and the references therein).

Assumption 3.3(i) is for the derivation of a rate of convergence of $\hat{\Theta}_T - \Theta$ in terms of spectral norm. To establish the *same* rate of convergence of $\hat{\Sigma}_T - \Sigma$ in terms of spectral norm, 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 (see Lemma A.14 in Appendix A.5).

Assumptions 3.3(ii) and (iii) are *sufficient* conditions for the asymptotic normality of the minimum distance estimators (Theorems 3.2 and 3.3) and one-step estimator (Theorem 4.2), respectively. Assumption 3.3(ii) or (iii) necessarily requires $n^4/T \to 0$. At first glance, it looks restrictive, but we would like to emphasize that this is only a sufficient condition. We will have more to say on this assumption in the discussions following Theorem 3.2.

Assumption 3.4(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 Lemma A.7(i) in Appendix A.4) and its logarithm is well-defined. Assumption 3.4(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 Lemma A.1 in Appendix A.1). In words, Assumption 3.4(ii) says that the minimum eigenvalue of $E^\intercal E/n$ is allowed to slowly drift to zero.

The following theorem establishes an upper bound on the rate of convergence for the minimum distance estimator $\hat{\theta}_T$. To arrive at this, we restrict r_1 and r_2 such that $1/r_1 + 1/r_2 > 1$. However, this is not a necessary condition.

Theorem 3.1.

(i) Suppose Assumptions 3.1(i), 3.2, 3.3(i) and 3.4(i) hold with $1/r_1 + 1/r_2 > 1$. Then

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

where $\|\cdot\|_{\ell_2}$ is the spectral norm.

(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, 3.3(i) and 3.4 hold with $1/r_1 + 1/r_2 > 1$. 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 3.4(ii).

Proof. See Appendix A.3. □

Theorem 3.1(i) provides a rate of convergence of the spectral norm of $\hat{\Theta}_T - \Theta$, which is a stepping stone for the rest of theoretical results. This rate is the same as that of $\|\hat{\Sigma}_T - \Sigma\|_{\ell_2}$. The rate $\sqrt{n/T}$ is optimal in the sense that it cannot be improved without a further structural assumption on Θ or Σ.

Theorem 3.1(ii) is of independent interest as it relates $\|\log \hat{\Theta}_T - \log \Theta\|_{\ell_2}$ to $\|\hat{\Theta}_T - \Theta\|_{\ell_2}$. It is due to Gil' (2012).

Theorem 3.1(iii) gives a rate of convergence of the minimum distance estimator $\hat{\theta}_T$. Note that 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 a sum of n^2 terms each of which has a convergence 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 the 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 for $\hat{\theta}_T$ should be $\sqrt{\log n/T}$. In this sense, Theorem 3.1(iii) does not demonstrate the full advantages of a Kronecker product model. Because of the severe non-linearity introduced by the matrix logarithm it is a challenging problem to prove a faster rate of convergence for $\|\hat{\theta}_T - \theta^0\|_2$.

3.3 Asymptotic Normality

We define y_t's natural filtration $\mathcal{F}_t := \sigma(y_t, y_{t-1}, \ldots, y_1)$ and $\mathcal{F}_0 = \{\emptyset, \emptyset^c\}$.

Assumption 3.5.

(i) Suppose that $\{y_t - \mu, \mathcal{F}_t\}$ is a martingale difference sequence; that is $\mathbb{E}[y_t - \mu | \mathcal{F}_{t-1}] = 0$ for all $t = 1, \ldots, T$.

(ii) Suppose that $\{y_t y_t^\intercal - \mathbb{E}[y_t y_t^\intercal], \mathcal{F}_t\}$ is a martingale difference sequence; that is

$$\mathbb{E}\left[y_{t,i} y_{t,j} - \mathbb{E}[y_{t,i} y_{t,j}] | \mathcal{F}_{t-1}\right] = 0$$

for all $i, j = 1, \ldots, n$, $t = 1, \ldots, T$.

Assumption 3.5 allows us to establish inference results within a martingale framework. Outside this martingale framework, one encounters the issue of long-run variance whenever one tries to get some inference result. This is particularly challenging in the large dimensional case and we hence shall not consider it in this article.

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 these cases. We first comment on the plausibility or relevance of case (i). We present five situations/arguments to show that case (i) is relevant and these are by no means exhaustive. First, one could use higher frequency data to estimate the individual variances and thereby utilise a very large sample size. But that is not an option for estimating correlations because of the non-synchronicity problem, which is acute in the large dimensional case (Park, Hong, and Linton (2016)). Second, one could have unbalanced low frequency data meaning that each firm has a long time series but they start and finish at different times such that the overlap, which is relevant for estimation of correlations, can be quite a bit smaller. In that situation one might standardise marginally using all the individual time series data and then estimate pairwise correlations using the smaller overlapping data. Third, we could have a global parametric model for D and μ, but a local (in time) Kronecker product model for correlations, i.e., $\Theta(u)$ varies with rescaled time $u = t/T$. In this situation, the initial estimation of D and μ can be done at a faster rate than estimation of the time varying correlation $\Theta(u)$, so effectively D and μ could be treated as known quantities. Fourth, case (i) reflects our two-step estimation procedure where variances are estimated first without imposing any model structure. This is a common approach in dynamic volatility model estimation such as the DCC model of Engle and Sheppard (2001) and the GO-GARCH model (van der Weide (2002)). Indeed, in many of the early articles in that literature standard errors for dynamic parameters of the correlation process were constructed without regard to the effect of the initial procedure. Finally, we note that theoretically estimation of μ and D is well understood even in the high dimensional case, so in keeping with much practice in the literature we do not emphasize estimation of μ and D again.

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

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

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

$$\tilde{\Sigma}_T := \frac{1}{T} \sum_{t=1}^{T} (y_t - \mu)(y_t - \mu)^\intercal. \tag{3.4}$$

$$\begin{aligned} V &:= \operatorname{var}\left(\operatorname{vec}\left[(y_t - \mu)(y_t - \mu)^\intercal\right]\right) \\ &= \mathbb{E}\left[(y_t - \mu)(y_t - \mu)^\intercal \otimes (y_t - \mu)(y_t - \mu)^\intercal\right] - \mathbb{E}\left[(y_t - \mu) \otimes (y_t - \mu)\right] \mathbb{E}\left[(y_t - \mu)^\intercal \otimes (y_t - \mu)^\intercal\right]. \end{aligned}$$

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 (a, b)th entry of V is

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

where $\mu_i = \mathbb{E} y_{t,i}$ (similarly for μ_j, μ_k, μ_ℓ), $a, b \in \{1, \ldots, n^2\}$ 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 3.6. *Suppose that 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 3.6 is also a standard regularity condition. It is automatically satisfied under normality given Assumptions 3.3(i) and 3.4(i) (via Lemma A.4(vi) in Appendix A.3). Assumption 3.6 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 3.2 and Theorem 3.3 remain unchanged, but this rate will need to be incorporated in Assumption 3.3(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

$$\hat{H}_{T,D} := \int_0^1 [t(\hat{\Theta}_{T,D} - I) + I]^{-1} \otimes [t(\hat{\Theta}_{T,D} - I) + I]^{-1} dt. \tag{3.5}$$

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

$$\hat{V}_{T,a,b} \equiv \hat{V}_{T,i,j,k,\ell} := \frac{1}{T}\sum_{t=1}^{T}(y_{t,i}-\bar{y}_i)(y_{t,j}-\bar{y}_j)(y_{t,k}-\bar{y}_k)(y_{t,\ell}-\bar{y}_\ell) - \left(\frac{1}{T}\sum_{t=1}^{T}(y_{t,i}-\bar{y}_i)(y_{t,j}-\bar{y}_j)\right)\left(\frac{1}{T}\sum_{t=1}^{T}(y_{t,k}-\bar{y}_k)(y_{t,\ell}-\bar{y}_\ell)\right),$$

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

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

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

In the special case of normality, $c^\intercal J_D c$ could be simplified into (see Example 8.3 in SM 8.8 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 $c^\intercal \hat{J}_{T,D} c$:

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

Theorem 3.2. *Let Assumptions 3.1(i), 3.2, 3.3(ii), 3.4, 3.5 and 3.6 be satisfied with $1/r_1 + 1/r_2 > 1$. In particular we set $r_1 = 2$. Then*

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

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

Proof. See Appendix A.4. □

Theorem 3.2 is a version of the large-dimensional CLT, whose proof is mathematically non-trivial. To simplify the technicality, we assume subgaussianity ($r_1 = 2$). 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 3.3(ii); in particular, the dimension of parameter cannot exceed the sample size. Assumption 3.3(ii) necessarily requires $n^4/T \to 0$. In Lewis and Reinsel (1985), Saikkonen and Lutkepohl (1996), Chang, Chen, and Chen (2015), they require $n^3/T \to 0$ for establishment of a CLT for an n-dimensional parameter. Hence there is much room of improvement for Assumption 3.3(ii) because the dimension of θ^0 is $s = O(\log n)$. The difficulty for this relaxation is again, as we had mentioned when we discussed the rate of convergence of $\hat{\theta}_T$ (Theorem 3.1), due to the severe non-linearity introduced by matrix logarithm. In this sense Assumption 3.3(ii) is only a sufficient condition; the same reasoning applies to Assumption 3.3(iii).

Our 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, 3.3(ii), 3.4, 3.5 and 3.6 be satisfied with $1/r_1 + 1/r_2 > 1$. In particular we set $r_1 = 2$. Given a full-column-rank $s \times k$ matrix A where k is finite with $\|A\|_{\ell_2} = O(\sqrt{\log n \cdot n\kappa(W)})$, 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 8.8. □

Note that the condition $\|A\|_{\ell_2} = O(\sqrt{\log n \cdot n\kappa(W)})$ 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,

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

the scalar $c^\intercal J_D c$ 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., $V = 2D_n D_n^+(\Sigma \otimes \Sigma)$), the preceding display further simplifies to

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

by Lemmas 11 and 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

$$\hat{H}_T := \int_0^1 [t(\hat{\Theta}_T - I) + I]^{-1} \otimes [t(\hat{\Theta}_T - I) + I]^{-1} dt. \tag{3.6}$$

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

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

where F_{ii} is an $n \times n$ matrix with one in its (i,i)th position and zeros elsewhere. Matrix M_d is an $n^2 \times n^2$ 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 $n \times n$ matrix A in $\operatorname{vec} A$. Matrix P first appeared in (4.6) of Neudecker and Wesselman (1990). Note that for any correlation matrix Θ, matrix P is an idempotent matrix of rank $n^2 - n$ and has n rows of zeros. Neudecker and Wesselman (1990) proved that

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

that is, the derivative $\frac{\partial \operatorname{vec} \Theta}{\partial \operatorname{vec} \Sigma}$ is a function of Σ.

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

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

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

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

Assumption 3.7.

(i) For every positive constant C

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

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

(ii) The $s \times s$ matrix

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

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

$$\text{mineval}\left(E^\intercal W D_n^+ HP(D^{-1/2} \otimes D^{-1/2}) V(D^{-1/2} \otimes D^{-1/2}) P^\intercal H D_n^{+\intercal} WE \right) \geq \frac{n}{\varpi} \text{mineval}^2(W).$$

Assumption 3.7(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 made 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 3.7(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 3.7(i) is chosen to be $\sqrt{n/T}$ because it is the rate of convergence of

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

which could be easily deduced from the proof of Theorem 3.3. This rate $\sqrt{n/T}$ could even be relaxed to $\sqrt{n^2/T}$ as the part of the proof of Theorem 3.3 which requires Assumption 3.7(i) is not the "binding" part of the whole proof.

We now examine Assumption 3.7(ii). The $s \times s$ matrix

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

is symmetric and positive semidefinite. By Observation 7.1.8 of Horn and Johnson (2013), its rank is equal to $\operatorname{rank}(E^\intercal W D_n^+ HP)$, if $(D^{-1/2} \otimes D^{-1/2}) V(D^{-1/2} \otimes D^{-1/2})$ is positive definite. In other words, Assumption 3.7(ii) is assuming $\operatorname{rank}(E^\intercal W D_n^+ HP) = s$, 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 $\Theta = I_n$ and $W = I_{n(n+1)/2}$. In this special case

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

The second part of Assumption 3.7(ii) postulates a lower bound for its minimum eigenvalue. The rate $\text{mineval}^2(W) n/\varpi$ is specified as such because of Assumption 3.4(ii). Other magnitudes of the rate are also possible as long as the proof of Theorem 3.3 goes through.

Theorem 3.3. *Let Assumptions 3.1(i), 3.2, 3.3(ii), 3.4, 3.5, 3.6 and 3.7 be satisfied with $1/r_1 + 1/r_2 > 1$. In particular we set $r_1 = 2$. Then*

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

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

Proof. See SM 8.4. □

Again Theorem 3.3 is a version of the large-dimensional CLT, whose proof is mathematically non-trivial. It has the same structure as that of Theorem 3.2. However $c^\intercal \hat{J}_T c$ differs from $c^\intercal \hat{J}_{T,D} c$ reflecting the difference between $c^\intercal J c$ and $c^\intercal J_D c$. 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, 3.3(ii), 3.4, 3.5, 3.6 and 3.7 be satisfied with $1/r_1 + 1/r_2 > 1$. In particular we set $r_1 = 2$. Given a full-column-rank $s \times k$ matrix A where k is finite with $\|A\|_{\ell_2} = O(\sqrt{\log^2 n \cdot n\kappa^2(W)\varpi})$, 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. □

The condition $\|A\|_{\ell_2} = O(\sqrt{\log^2 n \cdot n\kappa^2(W)\varpi})$ 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. In the case of both μ and D unknown, the infeasible optimal weighting matrix will be

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

3.4 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 $n = n_1 \times \cdots \times n_v$. That is,

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

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

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

Theorem 3.4. *Fix n (and hence v and s).*

(i) Suppose μ is unknown but D is known. Let Assumptions 3.1(i), 3.2, 3.4, 3.5 and 3.6 be satisfied with $1/r_1 + 1/r_2 > 1$. In particular we set $r_1 = 2$. Thus, under H_0,

$$T g_{T,D}(\hat{\theta}_{T,D})^\intercal \hat{S}_{T,D}^{-1} g_{T,D}(\hat{\theta}_{T,D}) \xrightarrow{d} \chi^2_{n(n+1)/2-s}, \tag{3.7}$$

where

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

(ii) Suppose both μ and D are unknown. Let Assumptions 3.1(i), 3.2, 3.4, 3.5, 3.6, and 3.7 be satisfied with $1/r_1 + 1/r_2 > 1$. In particular we set $r_1 = 2$. 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 8.7. □

Note that $\hat{S}_{T,D}^{-1}$ and $\hat{S}_T^{-1}$ are the feasible versions of optimal weighting matrices $W_{D,op}$ and W_{op}, respectively. From Theorem 3.4, 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), 3.2, 3.4, 3.5 and 3.6 be satisfied with $1/r_1 + 1/r_2 > 1$. In particular we set $r_1 = 2$. 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), 3.2, 3.4, 3.5, 3.6, and 3.7 be satisfied with $1/r_1 + 1/r_2 > 1$. In particular we set $r_1 = 2$. Under H_0,

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

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

Proof. See SM 8.7. □

4 QMLE and One-Step Estimator

4.1 QMLE

In the context of Gaussian quasi-maximum likelihood estimation (QMLE), given a factorization $n = n_1 \times \cdots \times n_v$, we shall additionally assume that the Kronecker product model $\{\Theta^*\}$ is correctly specified (i.e. $\text{vech}(\log \Theta) = E\theta$). Let $\rho \in [-1,1]^{s_\rho}$ be the original parameters of some member of the Kronecker product model; we have mentioned that $s_\rho = \sum_{j=1}^{v} n_j(n_j - 1)/2$. Given Assumption 3.5, the log likelihood function in terms of original parameters ρ for a sample $\{y_1, y_2, \ldots, y_T\}$ is given by

$$\ell_T(\mu, D, \rho) = -\frac{Tn}{2}\log(2\pi) - \frac{T}{2}\log\left|D^{1/2}\Theta(\rho)D^{1/2}\right| - \frac{1}{2}\sum_{t=1}^{T}(y_t - \mu)^\intercal D^{-1/2}\Theta(\rho)^{-1}D^{-1/2}(y_t - \mu). \tag{4.1}$$

Write $\Omega = \Omega(\theta) := \log \Theta$. Given Assumption 3.5, the log likelihood function in terms of log parameters θ for a sample $\{y_1, y_2, \ldots, y_T\}$ is given by

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

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. Theorem 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 case 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 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 (4.2) 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 (4.2) reduces to

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

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

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

Theorem 4.1.

(i) The $s \times 1$ score function of (4.3) with respect to θ takes the following form[8]

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

where $\tilde{\Sigma}_T$ is defined in (3.4).

[8]The likelihood function (4.3) implicitly assumes Assumption 3.5 and positive definiteness of Θ.

(ii) The $s \times s$ block of the Hessian matrix of (4.3) corresponding to θ takes the following form

$$\frac{\partial^2 \ell_{T,D}(\theta,\mu)}{\partial\theta\partial\theta^\intercal} =$$

$$-\frac{T}{4}E^\intercal D_n^\intercal \int_0^1\int_0^1 \left(e^{-st\Omega}\otimes e^{-(1-s)t\Omega}Ae^{-(1-t)\Omega}+e^{-(1-t)\Omega}Ae^{-(1-s)t\Omega}\otimes e^{-st\Omega}\right)ds\cdot tdtD_nE$$
$$-\frac{T}{4}E^\intercal D_n^\intercal \int_0^1\int_0^1 \left(e^{-(1-s)t\Omega}Ae^{-(1-t)\Omega}\otimes e^{-st\Omega}+e^{-st\Omega}\otimes e^{-(1-t)\Omega}Ae^{-(1-s)t\Omega}\right)ds\cdot tdtD_nE$$

where $A := D^{-1/2}\tilde{\Sigma}_T D^{-1/2}$. Symmetry of $\frac{\partial^2 \ell_{T,D}(\theta,\mu)}{\partial\theta\partial\theta^\intercal}$ is in an obvious way.

(iii) The negative normalized expected Hessian matrix evaluated at the true parameter θ takes the following form

$$\Upsilon_D := \mathbb{E}\left[-\frac{1}{T}\frac{\partial^2 \ell_{T,D}(\theta,\mu)}{\partial\theta\partial\theta^\intercal}\right]$$
$$= \frac{1}{2}E^\intercal D_n^\intercal\int_0^1\int_0^1\left(e^{-st\Omega}\otimes e^{st\Omega}+e^{st\Omega}\otimes e^{-st\Omega}\right)ds\cdot tdtD_nE \qquad (4.4)$$
$$= \frac{1}{2}E^\intercal D_n^\intercal\Psi\left(e^{-\Omega}\otimes e^{-\Omega}\right)\Psi D_nE, \qquad (4.5)$$

where $\Psi := \int_0^1 e^{t\Omega}\otimes e^{(1-t)\Omega}dt$.

(iv) Under normality (i.e., $V = 2D_nD_n^+(\Sigma\otimes\Sigma)$), we have the well-known relation

$$\Upsilon_D = \mathbb{E}\left[\frac{1}{T}\frac{\partial \ell_{T,D}(\theta,\mu)}{\partial\theta^\intercal}\frac{\partial \ell_{T,D}(\theta,\mu)}{\partial\theta}\right].$$

Proof. See SM 8.5. □

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

$$\tilde{\theta}_{T,D} := \hat{\theta}_{T,D} - \frac{1}{T}\hat{\Upsilon}_{T,D}^{-1}\frac{\partial \ell_{T,D}(\hat{\theta}_{T,D},\bar{y})}{\partial\theta^\intercal}, \qquad (4.6)$$

where $\hat{\Upsilon}_{T,D}$ is a plug-in estimator of Υ_D and is defined as $\frac{1}{2}E^\intercal D_n^\intercal\left[\int_0^1\int_0^1 \hat{\Theta}_{T,D}^{t+s-1}\otimes\hat{\Theta}_{T,D}^{1-t-s}dtds\right]D_nE$ (We show in SM 8.6 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

To provide the large sample theory for the one-step estimator $\tilde{\theta}_{T,D}$, we make the following assumption.

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

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

Assumption 4.1 is one of the sufficient conditions needed for the asymptotic normality of $\tilde{\theta}_{T,D}$ (Theorem 4.2). 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 $\frac{1}{T}\frac{\partial \ell_{T,D}(\theta,\bar{y})}{\partial \theta^\intercal}$ 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 the display in Assumption 4.1 to $o_p(\sqrt{n/(\varpi^2\log n)})$ by examining the proof of Theorem 4.2.

Theorem 4.2. *Suppose that the Kronecker product model $\{\Theta^*\}$ is correctly specified. Let Assumptions 3.1(ii), 3.2, 3.3(iii), 3.4, 3.5, and 4.1 be satisfied with $1/r_1 + 1/r_2 > 1$ and $r_1 = 2$. Then*

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

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

Proof. See SM 8.6. □

Theorem 4.2 is a version of the large-dimensional CLT, whose proof is mathematically non-trivial. It has the same structure as that of Theorem 3.2 or Theorem 3.3. Note that under Assumption 3.1(ii), the QMLE is actually the maximum likelihood estimator (MLE). If we replace normality (Assumption 3.1(ii)) with the subgaussian assumption (Assumption 3.1(i) with $r_1 = 2$) - that is the Gaussian likelihood is not correctly specified - although the norm consistency of $\tilde{\theta}_{T,D}$ should still hold, the asymptotic variance in Theorem 4.2 needs to be changed to have a sandwich formula. Theorem 4.2 says that $\sqrt{T}c^\intercal(\tilde{\theta}_{T,D}-\theta) \xrightarrow{d} N\left(0, c^\intercal\left(\mathbb{E}\left[-\frac{1}{T}\frac{\partial^2\ell_{T,D}(\theta,\mu)}{\partial\theta\partial\theta^\intercal}\right]\right)^{-1}c\right)$. In the fixed n case, this estimator achieves the parametric efficiency bound by recognising 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 $H^{-1} = \int_0^1 e^{t\log\Theta}\otimes e^{(1-t)\log\Theta}dt = \Psi$ (see Lemma 8.10 in SM 8.8), 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 $\hat{\theta}_{T,D}(W_{D,op})$ 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, 3.3(iii), 3.4, 3.5, and 4.1 be satisfied with $1/r_1 + 1/r_2 > 1$ and $r_1 = 2$. Given a full-column-rank $s\times k$ matrix A where k is finite with $\|A\|_{\ell_2} = O(\sqrt{\log n\cdot n})$, 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 $\|A\|_{\ell_2} = O(\sqrt{\log n\cdot n})$ 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 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 a Kronecker product model is pinned down by the structure of multiway arrays - the Kronecker product model is correctly specified. This issue will pop up when one uses Kronecker product models to approximate a general covariance or correlation matrix - all Kronecker product models are then misspecified. The rest of discussions in this section will be based on this approximation framework.

First, if one permutes the data, the performance of a given Kronecker product model is likely to change. However, based on our experience, the performance of a Kronecker product model is not that sensitive to the ordering of the data. We will illustrate this in the empirical study. Moreover, usually one fixes the ordering of the data before considering the issue of covariance matrix estimation. Thus, Kronecker product models have a second-mover advantage: the choice of a Kronecker product model depends on the ordering of the data.

Second, if one fixes the ordering of the data as well as a factorization $n = n_1 \times \cdots \times n_v$, but 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 also obtains a different Kronecker product model. Suppose that n has the prime factorization $n = p_1 \times p_2 \times \cdots \times p_v$ 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 baseline model has dimensions $(p_1, p_2, \ldots, p_v)$, but there are many possible aggregations of this, for example, $\big((p_1 \times p_2), \ldots, (p_{v-1} \times p_v)\big)$.

To address the second and third issues, we might choose among Kronecker product models using some model selection criterion which penalizes models with more parameters. For example, we may define the Bayesian Information Criterion (BIC) in terms of the original parameters ρ:

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

where ℓ_T is the log likelihood function defined in (4.1), and s_ρ is the dimension of ρ. We seek the Kronecker product model with the minimum preceding display. Typically there are not so many factorizations to consider, so this is not too computationally burdensome.

6 Monte Carlo Simulations and an Application

In this section, we first provide a set of Monte Carlo simulations that evaluate the performance of the QMLE and MD estimator, and then give a small application of our Kronecker product model to daily stock returns.

6.1 Monte Carlo Simulations

We simulate T random vectors y_t of dimension n according to

$$y_t = \Sigma^{1/2} z_t, \qquad z_t \sim N(0, I_n) \qquad \Sigma = \Sigma_1 \otimes \Sigma_2 \otimes \cdots \otimes \Sigma_v, \tag{6.1}$$

where $n = 2^v$ and $v \in \mathbb{N}$. That is, the sub-matrices Σ_i are 2×2 for $i = 1, \ldots, v$. These sub-matrices Σ_j are generated with unit variances and off-diagonal elements drawn *randomly* from a uniform distribution on $(-1, 1)$. This ensures positive definiteness of Σ. Note that we have two sources of randomness in this data generating process: random innovations (z_t) and random off-diagonal elements of the Σ_i for $i = 1, \ldots, v$. Due to the unit variances, Σ is also the correlation matrix Θ of y_t, but the econometrician is unaware of this: He applies a Kronecker product model to the correlation matrix Θ. We consider the correctly specified case, i.e., the Kronecker product model has a factorization $n = 2^v$. The sample size is set to $T = 300$ while we vary v (hence n). We set the Monte Carlo simulations to 1000.

We shall consider the QMLE and MD estimator. For the QMLE, we estimate the original parameters ρ and obtain an estimator for Θ (and hence Σ) directly. Recalling (4.1), we could use $\ell_T(\bar{y}, \hat{D}_T, \rho)$ to optimise ρ. For the MD estimator, we estimate the log parameters θ^0 via formula (3.3), obtain an estimator for $\log\Theta$, and finally obtain an estimator for Θ (and hence Σ) via matrix exponential. In the MD case, we need to specify a choice of the weighting matrix W. Given its sheer dimension $(n(n+1)/2 \times n(n+1)/2)$, any non-sparse W will be a huge computational burden in terms of memory for the MD estimator. Hence we consider two diagonal weighting matrices

$$W_1 = I_{n(n+1)/2}, \qquad W_2 = \left[D_n^+ \left(\hat{D}_T \otimes \hat{D}_T\right) D_n^{+\intercal}\right]^{-1}.$$

In the latter case, the MD estimator is inversely weighted by the sample variances. Weighting matrix W_2 resembles, but is not the same as, a feasible version of the optimal weighting matrix W_{op}. The choice of W_2 is based on heuristics. In an unreported simulation, we also consider the optimally weighted MD estimator. The optimally weighted MD estimator is extremely computationally intensive and its finite sample performance is not as good as those weighted by W_1 or W_2. This is probably because a data-driven, large-dimensional weighting matrix introduces additional sizeable estimation errors in small samples - such a phenomenon has been well documented in the GMM framework by Andersen and Sørensen (1996).

We compare our estimators with Ledoit and Wolf (2017)'s direct nonlinear shrinkage estimator (the LW2017 estimator hereafter).[9]

Given a generic estimator $\tilde{\Sigma}$ of the covariance matrix Σ and in each simulation, we can compute

$$1 - \frac{\|\tilde{\Sigma} - \Sigma\|_F^2}{\|\hat{\Sigma}_T - \Sigma\|_F^2}.$$

The median of the preceding display is calculated among all the simulations and denoted RI in terms of Σ. Criterion RI is closely related to the *percentage relative improvement in average loss* (PRIAL) criterion in Ledoit and Wolf (2004).[10] As PRIAL, RI measures the performance of the estimator $\tilde{\Sigma}$ with respect to the sample covariance estimator $\hat{\Sigma}_T$. Note that RI $\in (-\infty, 1]$: A negative value means $\tilde{\Sigma}$ performs worse than $\hat{\Sigma}_T$ while a positive value means otherwise. RI is more robust to outliers than PRIAL.

Often an estimator of the precision matrix Σ^{-1} is of more interest than that of Σ itself, so we also compute RI for the inverse covariance matrix; that is, we compute the median of

$$1 - \frac{\|\tilde{\Sigma}^{-1} - \Sigma^{-1}\|_F^2}{\|\hat{\Sigma}_T^{-1} - \Sigma^{-1}\|_F^2}.$$

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. The weights of the minimum variance portfolio are given by

$$w_{MVP} = \frac{\Sigma^{-1}\iota_n}{\iota_n^\intercal \Sigma^{-1}\iota_n}, \tag{6.2}$$

[9]The Matlab code for the direct nonlinear shrinkage estimator is downloaded from the website of Professor Michael Wolf from the Department of Economics at the University of Zurich. We are grateful for this.

[10]It is defined as

$$\text{PRIAL} = 1 - \frac{\mathbb{E}\|\tilde{\Sigma} - \Sigma\|_F^2}{\mathbb{E}\|\hat{\Sigma}_T - \Sigma\|_F^2}.$$

	n	4	8	16	32	64	128	256
RI-1	QMLE	0.227	0.529	0.714	0.820	0.892	0.929	0.950
	MD1	0.345	0.632	0.789	0.862	0.897	0.909	0.618
	MD2	0.339	0.631	0.785	0.858	0.896	0.908	0.616
	LW2017	0.020	0.027	0.046	0.063	0.087	0.106	0.127
RI-2	QMLE	0.323	0.615	0.805	0.914	0.973	0.995	1.000
	MD1	0.354	0.632	0.771	0.752	0.665	0.588	0.837
	MD2	0.344	0.643	0.790	0.796	0.714	0.628	0.846
	LW2017	0.136	0.181	0.235	0.351	0.521	0.756	0.991
VR	QMLE	0.999	0.995	0.980	0.953	0.899	0.770	0.389
	MD1	0.999	0.993	0.979	0.953	0.900	0.774	0.401
	MD2	0.999	0.993	0.979	0.954	0.899	0.774	0.400
	LW2017	1.000	0.999	0.998	0.993	0.975	0.912	0.544

Table 1: The baseline setting. QMLE, MD1, MD2 and LW2017 stand for the quasi-maximum likelihood estimator of the Kronecker product model, the minimum distance estimator (weighted by W_1) of the Kronecker product model, the minimum distance estimator (weighted by W_2) of the Kronecker product model, and the Ledoit and Wolf (2017)'s direct nonlinear shrinkage estimator, respectively. RI-1 and RI-2 are RI criteria in terms of Σ and Σ^{-1}, respectively. VR is the median of the ratio of the standard deviation of the MVP using the estimator to that using the sample covariance matrix out of sample. The sample size is fixed at $T = 300$.

where $\iota_n = (1, 1, \dots, 1)^\intercal$ is of dimension n (see Ledoit and Wolf (2003), Chan, Karceski, and Lakonishok (1999) etc). The first MVP weights are constructed using the sample covariance matrix $\hat{\Sigma}_T$ while the second MVP weights are constructed using a generic estimator of $\tilde{\Sigma}$. These two minimum variance portfolios are then evaluated by calculating their standard deviations in the out-of-sample data (y_t) generated using the same mechanism. The out-of-sample size is set to $T' = 21$. The ratio of the standard deviation of the minimum variance portfolio constructed from $\tilde{\Sigma}$ over that of the minimum variance portfolio constructed from $\hat{\Sigma}_T$ is calculated. We report its median (VR) over Monte Carlo simulations. Note that VR $\in [0, +\infty)$: A value greater than one means $\tilde{\Sigma}$ performs worse than $\hat{\Sigma}_T$ while a value less than one means otherwise.

Table 1 reports RI-1 (RI in terms of Σ), RI-2 (RI in terms of Σ^{-1}) and VR for various n. We observe the following patterns. First, we see that all our estimators QMLE, MD1, MD2 outperform the sample covariance matrix in all dimensional cases including both the small-dimensional cases (e.g., $n = 4$) and the large-dimensional cases (e.g., $n = 256$). Note that in the large dimensional case like $n = 256, T = 300$, the ratio n/T is close to 1 - a case not really covered by Assumption 3.3. This perhaps illustrates that Assumption 3.3 is a sufficient but not necessary condition for theoretical analysis of our proposed methodology. Second, such a phenomenon holds in terms of RI-1, RI-2 and VR. The superiority of our estimators over the sample covariance matrix increases when n/T increases. Third, the QMLE outperforms the MD estimators whenever n/T is close to one, while the opposite holds when n/T is small. Fourth, the LW2017 estimator also beats the sample covariance matrix but its RI-1 margin is thin. This is perhaps not surprising as the LW2017 estimator does not utilise the Kronecker product structure of the data generating process. Overall, the QMLE is the best estimator in this baseline setting.

As robustness checks, we consider two modifications of our baseline data generating process:

	n	4	8	16	32	64	128	256
RI-1	QMLE	0.219	0.514	0.712	0.821	0.887	0.928	0.951
	MD1	0.321	0.611	0.775	0.849	0.880	0.889	0.798
	MD2	0.310	0.611	0.770	0.844	0.877	0.890	0.796
	LW2017	0.025	0.032	0.049	0.065	0.093	0.117	0.155
RI-2	QMLE	0.320	0.654	0.824	0.932	0.980	0.997	1.000
	MD1	0.338	0.639	0.737	0.691	0.593	0.517	0.822
	MD2	0.347	0.652	0.775	0.753	0.657	0.571	0.839
	LW2017	0.220	0.292	0.429	0.634	0.818	0.939	0.997
VR	QMLE	0.998	0.988	0.975	0.927	0.860	0.728	0.383
	MD1	0.997	0.987	0.973	0.925	0.862	0.733	0.406
	MD2	0.997	0.987	0.973	0.924	0.862	0.732	0.406
	LW2017	1.000	0.999	0.997	0.990	0.970	0.907	0.568

Table 2: Modification (i). QMLE, MD1, MD2 and LW2017 stand for the quasi-maximum likelihood estimator of the Kronecker product model, the minimum distance estimator (weighted by W_1) of the Kronecker product model, the minimum distance estimator (weighted by W_2) of the Kronecker product model, and the Ledoit and Wolf (2017)'s direct nonlinear shrinkage estimator, respectively. RI-1 and RI-2 are RI criteria in terms of Σ and Σ^{-1}, respectively. VR is the median of the ratio of the standard deviation of the MVP using the estimator to that using the sample covariance matrix out of sample. The sample size is fixed at $T = 300$.

(i) Time series y_t is still generated as in (6.1) but the actual data are w_t:

$$\begin{aligned} w_1 &= y_1 \\ w_t &= a_w w_{t-1} + \sqrt{1 - a_w^2} y_t, \qquad t = 2, \ldots, T. \end{aligned}$$

The parameter a_w is set to be 0.5 to capture the temporal dependence.

(ii) Same as modification (i), but y_t is drawn from a multivariate t distribution of 5 degrees of freedom with Σ as its correlation matrix.

In modification (i), w_t is serially correlated given any non-zero autoregressive scalar a_w but its covariance matrix is still Σ. A choice of $a_w = 0.5$ is consistent with Assumption 3.2. Our simulation results are reasonably robust to the choice of a_w. In modification (ii), in addition to the serial dependence, we add heavy-tailed features to the data which might be a better reflection of reality. Heavy-tailed data are not covered by Assumption 3.1, so this modification serves as a robustness check for our theoretical findings.

The results of modification (i) are reported in Table 2. Those four observations we made from the baseline setting (Table 1) still hold when we relax the independence assumption of the data. Modification (ii) are reported in Table 3. When we switch on both temporal dependence and heavy tails, all estimators - ours and the LW2017 estimator - are adversely affected to a certain extent. In particular, in terms of RI-2, both the QMLE and LW2017 estimators fare worse than the sample covariance matrix in small dimensions. Overall, the identity weighted MD estimator is the best estimator in modification (ii). That the MD estimator trumps the QMLE in heavy-tailed data is intuitive because the MD estimator is derived not based on a particular distributional assumption.

	n	4	8	16	32	64	128	256
RI-1	QMLE	0.021	0.105	0.203	0.320	0.442	0.564	0.690
	MD1	0.071	0.211	0.348	0.492	0.621	0.719	0.605
	MD2	0.084	0.242	0.378	0.510	0.626	0.712	0.581
	LW2017	−0.023	−0.001	0.029	0.069	0.118	0.158	0.220
RI-2	QMLE	−0.035	−0.139	−0.357	−0.831	−0.202	0.867	0.999
	MD1	0.006	0.035	0.111	0.385	0.896	0.636	0.829
	MD2	−0.006	−0.009	0.032	0.255	0.894	0.724	0.854
	LW2017	−0.103	−0.206	−0.428	−0.847	−0.279	0.825	0.997
VR	QMLE	0.996	0.982	0.956	0.923	0.842	0.708	0.379
	MD1	0.994	0.982	0.955	0.921	0.840	0.720	0.432
	MD2	0.994	0.982	0.955	0.920	0.840	0.719	0.429
	LW2017	1.000	0.999	0.995	0.989	0.968	0.906	0.577

Table 3: Modification (ii). QMLE, MD1, MD2 and LW2017 stand for the quasi-maximum likelihood estimator of the Kronecker product model, the minimum distance estimator (weighted by W_1) of the Kronecker product model, the minimum distance estimator (weighted by W_2) of the Kronecker product model, and the Ledoit and Wolf (2017)'s direct nonlinear shrinkage estimator, respectively. RI-1 and RI-2 are RI criteria in terms of Σ and Σ^{-1}, respectively. VR is the median of the ratio of the standard deviation of the MVP using the estimator to that using the sample covariance matrix out of sample. The sample size is fixed at $T = 300$.

6.2 An Application

We now consider estimation of the covariance matrix of $n' = 441$ stock returns (y_t) in the S&P 500 index. We have daily observations from January 3, 2005 to November 6, 2015. The number of trading days is $T = 2732$. Since the underlying data might not have a multiplicative structure giving rise to a Kronecker product - or if they do but we are unaware of it - a Kronecker product model in this application is inherently misspecified. In other words, we are exploiting Kronecker product models' approximating feature to a general covariance matrix.

We have proved in Appendix A.2 that in a given Kronecker product model there exists a member which is closest to the true covariance matrix. However, in order for this closest "distance" to be small, the chosen Kronecker product model needs to be versatile enough to capture various data patterns. In this sense, a parsimonious model, say, $441 = 3 \times 3 \times 7 \times 7$, is likely to be inferior to a less parsimonious model, say, $441 = 21 \times 21$.

We add an 3×1 dimensional pseudo random vector z_t which is $N(0, I_3)$ distributed and independent over t. The dimension of the final system is $n = 441 + 3 = 444$. Again we fit Kronecker product models to the correlation matrix of the final system and recover an estimator for the covariance matrix of the final system via left and right multiplication of the estimated correlation matrix of the final system by $\hat{D}_T^{1/2}$. Last, we extract the 441×441 upper-left block of the estimated covariance matrix of the final system to form our Kronecker product estimator of the covariance matrix of y_t. The dimension of the added pseudo random vector should not be too large to avoid introducing additional noise, which could adversely affect the performance of the Kronecker product models. We choose the dimension of the final system to be 444 because its prime factorization is $2 \times 2 \times 3 \times 37$, and we experiment with several Kronecker product models. We did try other dimensions for the final system and the pattern discussed below remains generally the same.

As we are considering less parsimonious models, the QMLE is computationally intensive and found to perform worse than the MD estimator in preliminary investigations, so we only

	MD ($2 \times 2 \times 3 \times 37$)	MD (4×111)	MD (3×148)	MD (2×222)	LW2004	LW2017
			original ordering of the data			
Impr	0.265	0.379	0.394	0.440	0.459	0.518
Prop	0.811	0.896	0.915	0.953	0.991	0.981
			a random permutation of the data			
Impr	0.259	0.364	0.404	0.431	0.459	0.518
Prop	0.811	0.887	0.915	0.943	0.991	0.981
			a random permutation of the data			
Impr	0.263	0.351	0.366	0.436	0.459	0.518
Prop	0.811	0.887	0.906	0.943	0.991	0.981

Table 4: MD, LW2004 and LW2017 stand for the (identity matrix weighted) minimum distance estimators of the Kronecker product models (factorisations given in parentheses), the Ledoit and Wolf (2004)'s linear shrinkage estimator, and the Ledoit and Wolf (2017)'s direct nonlinear shrinkage estimator, respectively. *Impr* is the median of the 106 quantities calculated based on (6.3) and *Prop* is the proportion of the times (out of 106) that a competitor MVP outperforms the sample covariance MVP (i.e., the proportion of the times when (6.3) is positive). A random permutation of the data means that the order of the 441 stocks is randomly reshuffled.

use the MD estimator. The MD estimator is extremely fast because its formula is just (3.3). We choose the weighting matrix to be the identity matrix.

We follow the approach of Fan et al. (2013) and estimate our 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 Kronecker product estimator of the covariance matrix of y_t is used to construct the minimum variance portfolio (MVP) weights as in (6.2). We also compute the MVP weights using the sample covariance matrix of y_t. These two minimum variance portfolios are then evaluated using the next 21 days (equal to one month's trading days) out-of-sample. In particular, we calculate

$$1 - \frac{\text{sd(a competitor MVP)}}{\text{sd(sample covariance MVP)}}, \tag{6.3}$$

where $\text{sd}(\cdot)$ computes standard deviation. Then the estimation window of 504 days is shifted forward by 21 days. This procedure is repeated until we reach the end of the sample; the total number of out-of-sample evaluations is 106. We consider two evaluation criteria of performance: *Impr* and *Prop*. *Impr* is the median of the 106 quantities calculated based on (6.3). Note that *Impr* $\in (-\infty, 1]$: A negative value means a competitor MVP performs worse than the sample covariance MVP while a positive value means otherwise. *Prop* is the proportion of the times (out of 106) that a competitor MVP outperforms the sample covariance MVP (i.e., the proportion of the times when (6.3) is positive).

For comparison, we consider Ledoit and Wolf (2004)'s linear shrinkage estimator and Ledoit and Wolf (2017)'s direct nonlinear shrinkage estimator. The results are reported in Table 4. We first use the original ordering of the data, i.e. alphabetical, and have the following observations. First, all the Kronecker product MVPs outperform the sample covariance MVP. Second, as we move from the most parsimonious factorisation ($444 = 2 \times 2 \times 3 \times 37$) to the least parsimonious factorisation ($444 = 2 \times 222$), the performance of Kronecker product MVPs monotonically improves. This is intuitive: Since we are using Kronecker product models to approximate a general covariance matrix, a more flexible Kronecker product model could fit the data better. There is no over-fitting at least in this application as we consider out-of-sample evaluation. Third, the performance of the (2×222) Kronecker product MVP is very close to that of a sophisticated estimator like Ledoit and Wolf (2004)'s linear shrinkage estimator. This is commendable because here a Kronecker product model is a misspecified parametric model for a general covariance matrix while the linear shrinkage estimator is in essence a data-driven, nonparametric estimator.

We next randomly reshuffle the order of the 441 stocks twice and use the same Kronecker product models. In these two cases, the rows and columns of the true covariance matrix also get reshuffled. We see that the performances of those Kronecker product models are marginally affected by the reshuffle. Ledoit and Wolf (2004)'s and Ledoit and Wolf (2017)'s shrinkage estimators are, as expected, not affected by the ordering of the data.

7 Conclusions

We have established the large sample properties of estimators of Kronecker product models in the large dimensional case. In particular, we obtained norm consistency and the large dimensional CLTs for the MD and one-step estimators. Kronecker product models outperform the sample covariance matrix theoretically, in Monte Carlo simulations, and in an application to portfolio choice. When a Kronecker product model is correctly specified, Monte Carlo simulations show that estimators of it can beat Ledoit and Wolf (2017)'s direct non-linear shrinkage estimator. In the application, when one uses Kronecker product models as an approximating device to a general covariance matrix, a less parsimonious one can perform almost as good as Ledoit and Wolf (2004)'s linear shrinkage estimator. It is possible to extend the framework in various directions to improve performance.

A final motivation for the Kronecker product structure is that it can be used as a component of a super model consisting of several components. 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.

A Appendix

This appendix is organised as follows: Appendix A.1 further discusses this matrix E of the minimum distance estimator in Section 3. Appendix A.2 shows that a Kronecker product model has a best approximation to a general covariance or correlation matrix. Appendix A.3 and A.4 contain proofs of Theorem 3.1 and of Theorem 3.2, respectively. Appendix A.5 contains auxiliary lemmas used in various places of this appendix.

A.1 Matrix E

The proof of the following theorem gives a concrete formula for the matrix E of the minimum distance estimator.

Theorem 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 $n = n_1 \times n_2 \times \cdots \times n_v$. 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

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

unrestricted log parameters; let $\theta^* \in \mathbb{R}^s$ denote these. Then there exists a $n(n+1)/2 \times s$ full column rank matrix E such that

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

Proof. Note that

$$\begin{aligned}\operatorname{vec}(\log \Theta^*) = \operatorname{vec}(\log \Theta_1^* \otimes I_{n_2} \otimes \cdots \otimes I_{n_v}) + \operatorname{vec}(I_{n_1} \otimes \log \Theta_2^* \otimes I_{n_3} \otimes \cdots \otimes I_{n_v}) + \cdots \\ + \operatorname{vec}(I_{n_1} \otimes I_{n_2} \otimes \cdots \otimes \log \Theta_v^*).\end{aligned}$$

If

$$\operatorname{vec}(I_{n_1} \otimes \log \Theta_i^* \otimes I_{n_3} \otimes \cdots \otimes I_{n_v}) = E_i \operatorname{vech}(\log \Theta_i^*)$$

for some $n^2 \times n_i(n_i+1)/2$ matrix E_i for $i = 1, \ldots, v$, then we have

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

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

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

(identifiable) log parameters; let $\theta^* \in \mathbb{R}^s$ 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

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

and $E_{i,(-1)}$ stands for matrix E_i with its first column removed. We now determine the formula for E_i. We first consider $\operatorname{vec}(\log \Theta_1^* \otimes I_{n_2} \otimes \cdots \otimes I_{n_v})$.

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

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

$$E_1 := \left(I_{n_1} \otimes K_{n/n_1,n_1} \otimes I_{n/n_1}\right)\left(I_{n_1^2} \otimes \operatorname{vec} I_{n/n_1}\right) D_{n_1}.$$

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

$$\begin{aligned}&\operatorname{vec}(I_{n_1} \otimes \cdots \otimes \log \Theta_i^* \otimes \cdots \otimes I_{n_v}) = \operatorname{vec}\left[K_{n_1\cdots n_{i-1},n/(n_1\cdots n_{i-1})}\left(\log \Theta_i^* \otimes I_{n/n_i}\right) K_{n/(n_1\cdots n_{i-1}),n_1\cdots n_{i-1}}\right] \\ &= \left[K^\intercal_{n/(n_1\cdots n_{i-1}),n_1\cdots n_{i-1}} \otimes K_{n_1\cdots n_{i-1},n/(n_1\cdots n_{i-1})}\right] \operatorname{vec}\left(\log \Theta_i^* \otimes I_{n/n_i}\right) \\ &= \left[K_{n_1\cdots n_{i-1},n/(n_1\cdots n_{i-1})} \otimes K_{n_1\cdots n_{i-1},n/(n_1\cdots n_{i-1})}\right]\left(I_{n_i} \otimes K_{n/n_i,n_i} \otimes I_{n/n_i}\right)\left(I_{n_i^2} \otimes \operatorname{vec} I_{n/n_i}\right) D_{n_i} \operatorname{vech}(\log \Theta_i^*),\end{aligned}$$

where the first equality is due to the identity $B \otimes A = K_{p,m}(A \otimes B)K_{m,p}$ for A $(m \times m)$ and B $(p \times p)$. Thus

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

for $i = 2, \ldots, v$. $\square$

Lemma A.1. *Given that $n = n_1 \times n_2 \times \cdots \times n_v$, 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^* has appeared in $\operatorname{vech}(\log \Theta^*)$. The value depends on to which $\log \Theta_j^*$ the ith parameter in θ^*, θ_i^*, belongs to. For instance, suppose θ_i^* is a parameter belonging to $\log \Theta_3^*$, 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^*, θ_i^*, and kth parameter in θ^*, θ_k^*, have appeared together (as summands) in an entry of $\operatorname{vech}(\log \Theta^*)$. The value depends on to which $\log \Theta_j^*$ the ith parameter in θ^*, θ_i^*, and kth parameter in θ^*, θ_k^*, belong to. For instance, suppose θ_i^* is a parameter belonging to $\log \Theta_3^*$ and θ_k^* is a parameter belonging to $\log \Theta_5^*$, then

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

However, the formula in the preceding display is overridden for the following two cases. If both θ_i^* and θ_k^* belong to the same $\log \Theta_j^*$, then $(E^\intercal E)_{i,k} = (E^\intercal E)_{k,i} = 0$. Also note that when θ_i^* is an off-diagonal entry of some $\log \Theta_j^*$, 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. $\square$

We here give a concrete example to illustrate Lemma A.1.

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

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

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

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

Then we can invoke Lemma A.1 to write down $E^\intercal E$ without even using Matlab 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 Best Approximation

In this section of the 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 in 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 $n = n_1 \times \cdots \times n_v$ is specified.

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. Space $\mathcal{M}_n$ with inner product $\langle \cdot, \cdot \rangle_W$ can be identified by $\mathbb{R}^{n(n+1)/2}$ with the usual Euclidean inner product. Moreover, since, for finite n, $\mathbb{R}^{n(n+1)/2}$ with the usual Euclidean inner product is a Hilbert space, 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 $j = 1, \ldots, v$. Note that $\mathcal{D}_n$ 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. We have $\log \Sigma \in \mathcal{M}_n$. 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

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

where $\Sigma^0 := \exp L^0$. 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 products. Note that 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 the 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 The Proof of Theorem 3.1

In this subsection, we give a proof for Theorem 3.1. We will first give some preliminary lemmas leading to the proof of this theorem.

The following lemma characterises the relationship between an exponential-type moment assumption and an exponential tail probability.

Lemma A.2. *Suppose that a random variable X satisfies the exponential-type tail condition, i.e., there exist absolute constants $K_1 > 1, K_2 > 0, r_1 > 0$ such that*

$$\mathbb{E}\left[\exp\left(K_2|X|^{r_1}\right)\right] \leq K_1.$$

(i) Then for every $\epsilon \geq 0$, there exists an absolute constant $b_1 > 0$ such that

$$\mathbb{P}(|X| \geq \epsilon) \leq \exp\left[1 - (\epsilon/b_1)^{r_1}\right].$$

(ii) We have $\mathbb{E}|X| < \infty$.

(iii) Part (i) implies that for every $\epsilon \geq 0$, there exists an absolute constant $c_1 > 0$ such that

$$\mathbb{P}(|X - \mathbb{E}X| \geq \epsilon) \leq \exp\left[1 - (\epsilon/c_1)^{r_1}\right].$$

(iv) Suppose that another random variable Y satisfies $\mathbb{E}\left[\exp\left(K_2^*|Y|^{r_1^*}\right)\right] \leq K_1^*$ for some absolute constants $K_1^* > 1, K_2^* > 0, r_1^* > 0$. Then for every $\epsilon \geq 0$, there exists an absolute constant $b_2 > 0$ such that

$$\mathbb{P}(|XY| \geq \epsilon) \leq \exp\left[1 - (\epsilon/b_2)^{r_2}\right],$$

where $r_2 \in \left(0, \frac{r_1 r_1^*}{r_1 + r_1^*}\right]$.

Proof. For part (i), choose $C := \log K_1 \vee 1$ and $b_1 := (C/K_2)^{1/r_1}$. If $\epsilon > b_1$, we have

$$\mathbb{P}(|X| \geq \epsilon) \leq \frac{\mathbb{E}\left[\exp(K_2|X|^{r_1})\right]}{\exp(K_2\epsilon^{r_1})} \leq K_1 e^{-K_2\epsilon^{r_1}} = e^{\log K_1 - K_2\epsilon^{r_1}} = e^{\log K_1 - C(\epsilon/b_1)^{r_1}}$$
$$\leq e^{C[1-(\epsilon/b_1)^{r_1}]} \leq e^{1-(\epsilon/b_1)^{r_1}}$$

where the first inequality is due to a variant of Markov's inequality. If $\epsilon \leq b_1$, we have

$$\mathbb{P}(|X| \geq \epsilon) \leq 1 \leq e^{1-(\epsilon/b_1)^{r_1}}.$$

For part (ii),

$$\mathbb{E}|X| = \int_0^\infty \mathbb{P}(|X| \geq t)dt \leq \int_0^\infty e^{1-(t/b_1)^{r_1}}dt = e\int_0^\infty e^{-(t/b_1)^{r_1}}dt = \frac{eb_1}{r_1}\int_0^\infty y^{\frac{1}{r_1}-1}e^{-y}dy$$
$$= \frac{eb_1}{r_1}\Gamma(r_1^{-1}) < \infty,$$

where the first inequality is due to part (i), the third equality is due to change of variable $y = (t/b_1)^{r_1}$, and the last equality is due to recognition of $\int_0^\infty [\Gamma(r_1^{-1})]^{-1} y^{\frac{1}{r_1}-1}e^{-y}dy = 1$ using Gamma distribution. For part (iii),

$$\mathbb{P}(|X - \mathbb{E}X| \geq \epsilon) \leq \mathbb{P}(|X| \geq \epsilon - \mathbb{E}|X|) = \mathbb{P}(|X| \geq \epsilon - \mathbb{E}|X| \wedge \epsilon) \leq \exp\left[1 - \frac{(\epsilon - \mathbb{E}|X| \wedge \epsilon)^{r_1}}{b_1^{r_1}}\right]$$

where the second inequality is due to part (i). First consider the case $0 < r_1 < 1$.

$$\exp\left[1 - \frac{(\epsilon - \mathbb{E}|X| \wedge \epsilon)^{r_1}}{b_1^{r_1}}\right] \leq \exp\left[1 - \frac{\epsilon^{r_1} - (\mathbb{E}|X| \wedge \epsilon)^{r_1}}{b_1^{r_1}}\right] = \exp\left[1 - \frac{\epsilon^{r_1}}{b_1^{r_1}} + \frac{(\mathbb{E}|X| \wedge \epsilon)^{r_1}}{b_1^{r_1}}\right]$$
$$\leq \exp\left[1 - \frac{\epsilon^{r_1}}{b_1^{r_1}} + \frac{(\mathbb{E}|X|)^{r_1}}{b_1^{r_1}}\right] \leq \exp\left[C - \frac{\epsilon^{r_1}}{b_1^{r_1}}\right] = \exp\left[C\left(1 - \frac{\epsilon^{r_1}}{(C^{\frac{1}{r_1}}b_1)^{r_1}}\right)\right] =: \exp\left[C\left(1 - \frac{\epsilon^{r_1}}{c_1^{r_1}}\right)\right]$$

where the first inequality is due to subadditivity of the concave function: $(x+y)^{r_1} - x^{r_1} \leq y^{r_1}$ for $x, y \geq 0$. If $\epsilon > c_1$, we have, via recognising $C > 1$,

$$\mathbb{P}(|X - \mathbb{E}X| \geq \epsilon) \leq \exp\left[C\left(1 - \frac{\epsilon^{r_1}}{c_1^{r_1}}\right)\right] \leq \exp\left[1 - \frac{\epsilon^{r_1}}{c_1^{r_1}}\right].$$

If $\epsilon \leq c_1$, we have

$$\mathbb{P}(|X - \mathbb{E}X| \geq \epsilon) \leq 1 \leq \exp\left[1 - \frac{\epsilon^{r_1}}{c_1^{r_1}}\right].$$

We now consider the case $r_1 \geq 1$. The proof is almost the same: Instead of relying on subadditivity of the concave function, we rely on Loeve's c_r inequality: $|x+y|^{r_1} \leq 2^{r_1-1}(|x|^{r_1} + |y|^{r_1})$ for $r_1 \geq 1$ to get $2^{1-r_1}\epsilon^{r_1} - (\mathbb{E}|X| \wedge \epsilon)^{r_1} \leq (\epsilon - \mathbb{E}|X| \wedge \epsilon)^{r_1}$. c_1 is now defined as $C^{\frac{1}{r_1}}b_1 2^{\frac{r_1-1}{r_1}}$. For part (iv), an original proof could be found in Fan et al. (2011) p3338. Invoke part (i), $\mathbb{P}(|Y| \geq \epsilon) \leq \exp\left[1 - (\epsilon/b_1^*)^{r_1^*}\right]$. We have, for any $\epsilon \geq 0$, $M := \left(\frac{\epsilon (b_1^*)^{(r_1^*/r_1)}}{b_1}\right)^{\frac{r_1}{r_1+r_1^*}}$, $b := b_1 b_1^*$, $r := \frac{r_1 r_1^*}{r_1+r_1^*}$,

$$\mathbb{P}(|XY| \geq \epsilon) \leq \mathbb{P}(|X| \geq \epsilon/M) + \mathbb{P}(|Y| \geq M) \leq \exp\left[1 - \left(\frac{\epsilon/M}{b_1}\right)^{r_1}\right] + \exp\left[1 - \left(\frac{M}{b_1^*}\right)^{r_1^*}\right]$$
$$= 2\exp\left[1 - (\epsilon/b)^r\right].$$

Pick an $r_2 \in (0, r]$ and $b_2 > (1 + \log 2)^{1/r}b$. We consider the case $\epsilon \leq b_2$ first.

$$\mathbb{P}(|XY| \geq \epsilon) \leq 1 \leq \exp\left[1 - (\epsilon/b_2)^{r_2}\right].$$

We now consider the case $\epsilon > b_2$. Define a function $F(\epsilon) := (\epsilon/b)^r - (\epsilon/b_2)^{r_2}$. Using the definition of b_2, we have $F(b_2) > \log 2$. It is also not difficult to show that $F'(\epsilon) > 0$ when $\epsilon > b_2$. Thus we have $F(\epsilon) > F(b_2) > \log 2$ when $\epsilon > b_2$. Thus,

$$\mathbb{P}(|XY| \geq \epsilon) \leq 2\exp\left[1 - (\epsilon/b)^r\right] = \exp\left[\log 2 + 1 - (\epsilon/b)^r\right] \leq \exp[(\epsilon/b)^r - (\epsilon/b_2)^{r_2} + 1 - (\epsilon/b)^r]$$
$$= \exp[1 - (\epsilon/b_2)^{r_2}].$$

□

This following lemma gives a rate of convergence in terms of spectral norm for the sample covariance matrix.

Lemma A.3. *Assume $n, T \to \infty$ simultaneously and $n/T \leq 1$. Suppose Assumptions 3.1(i) and 3.2 hold with $1/r_1 + 1/r_2 > 1$. Then*

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

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

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

We consider the first term on the right hand side of (A.1) first. Invoke Lemma A.11 in Appendix A.5 with $\varepsilon = 1/4$:

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

where $z_{a,t} := y_t^{\intercal} a$. First, given Assumption 3.1(i), invoke Lemma A.2(i) and (iv): For every $\epsilon \geq 0$, there exists an absolute constant $b_2 > 0$ such that

$$\mathbb{P}(|z_{a,t}^2| \geq \epsilon) \leq \exp\left[1 - (\epsilon/b_2)^{r_1/2}\right].$$

Next, invoke Lemma A.2(iii): For every $\epsilon \geq 0$, there exists an absolute constant $c_2 > 0$ such that

$$\mathbb{P}(|z_{a,t}^2 - \mathbb{E} z_{a,t}^2| \geq \epsilon) \leq \exp\left[1 - (\epsilon/c_2)^{r_1/2}\right].$$

Given Assumption 3.2 and the fact that mixing properties are hereditary in the sense that for any measurable function $m(\cdot)$, the process $\{m(y_t)\}$ possesses the mixing property of $\{y_t\}$ (Fan and Yao (2003) p69), $z_{a,t}^2 - \mathbb{E} z_{a,t}^2$ is strong mixing with the same coefficient: $\alpha(h) \leq \exp\left(-K_3 h^{r_2}\right)$. Define r by $1/r := 2/r_1 + 1/r_2$. Using the fact that $2/r_1 + 1/r_2 > 1$, we can invoke a version of Bernstein's inequality for strong mixing time series (Theorem A.2 in Appendix A.5), followed by Lemma A.12 in Appendix A.5:

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

Invoking Lemma A.10 in Appendix A.5, we have $|\mathcal{N}_{1/4}| \leq 9^n$. Thus we have

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

We now consider the second term on the right hand side of (A.1).

$$\|\bar{y}\bar{y}^{\intercal} - \mu\mu^{\intercal}\|_{\ell_2} \leq 2 \max_{a \in \mathcal{N}_{1/4}} \left| a^{\intercal}\left(\bar{y}\bar{y}^{\intercal} - \mu\bar{y}^{\intercal} + \mu\bar{y}^{\intercal} - \mu\mu^{\intercal}\right) a \right| = 2 \max_{a \in \mathcal{N}_{1/4}} \left| a^{\intercal}\left((\bar{y} - \mu)\bar{y}^{\intercal} + \mu(\bar{y} - \mu)^{\intercal}\right) a \right|$$
$$\leq 2 \max_{a \in \mathcal{N}_{1/4}} \left| a^{\intercal}(\bar{y} - \mu)\bar{y}^{\intercal} a \right| + 2 \max_{a \in \mathcal{N}_{1/4}} \left| a^{\intercal}\mu(\bar{y} - \mu)^{\intercal} a \right|$$
$$\leq 2 \max_{a \in \mathcal{N}_{1/4}} \left| a^{\intercal}(\bar{y} - \mu) \right| \max_{a \in \mathcal{N}_{1/4}} \left| \bar{y}^{\intercal} a \right| + 2 \max_{a \in \mathcal{N}_{1/4}} \left| a^{\intercal}\mu \right| \max_{a \in \mathcal{N}_{1/4}} \left| (\bar{y} - \mu)^{\intercal} a \right|,$$

where the first inequality is due to Lemma A.11 in Appendix A.5 with $\varepsilon = 1/4$. We consider $\max_{a\in\mathcal{N}_{1/4}}\left|(\bar{y}-\mu)^\intercal a\right|$ first.

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

Recycling the proof for $\max_{a\in\mathcal{N}_{1/4}}\left|\frac{1}{T}\sum_{t=1}^{T}(z_{a,t}^2-\mathbb{E}z_{a,t}^2)\right| = O_p\left(\sqrt{\frac{n}{T}}\right)$ but with $1/r := 1/r_1 + 1/r_2 > 1$ this time, we have

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

Now let's consider $\max_{a\in\mathcal{N}_{1/4}}\left|a^\intercal\mu\right|$.

$$\max_{a\in\mathcal{N}_{1/4}}\left|a^\intercal\mu\right| := \max_{a\in\mathcal{N}_{1/4}}\left|\mathbb{E}a^\intercal y_t\right| = \max_{a\in\mathcal{N}_{1/4}}\left|\mathbb{E}z_{a,t}\right| \leq \max_{a\in\mathcal{N}_{1/4}}\mathbb{E}|z_{a,t}| = O(1), \tag{A.3}$$

where the last equality is due to Lemma A.2(ii). Next we consider $\max_{a\in\mathcal{N}_{1/4}}\left|a^\intercal\bar{y}\right|$.

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

where the last equality is due to $n \leq T$. Combining (A.2), (A.3) and (A.4), we have

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

□

The following lemma gives a rate of convergence in terms of spectral norm for various quantities involving variances of y_t. The rate $\sqrt{n/T}$ is suboptimal, but there is no need improving it further as these quantities will not be the dominant terms in the proof of Theorem 3.1.

Lemma A.4. *Suppose Assumptions 3.1(i), 3.2, 3.3(i) and 3.4(i) hold with* $1/r_1 + 1/r_2 > 1$. *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 $\sigma_i^2 := \mathbb{E}(y_{t,i} - \sigma_i)^2$ and $\hat{\sigma}_i^2 := \frac{1}{T}\sum_{t=1}^T (y_{t,i} - \bar{y}_i)^2$, where the subscript i denotes the ith component of the corresponding vector. For part (i),

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

where e_i denotes a unit vector whose ith component is 1. Now invoke Lemma A.3 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 3.4(i). The statement about the minimum eigenvalue of $D^{1/2}$ is also true. For part (iii), invoking Lemma A.13 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.14 in Appendix A.5 via parts (ii) and (iii). For part (v),

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

where the second equality is due to parts (iv) and (ii). For part (vi), we have

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

We have shown that in the proof of Lemma A.3 that $z_{a,t}^2$ has an exponential tail for any $\|a\|_2 = 1$. This says that $\mathbb{E}z_{a,t}^2$ is bounded for any $\|a\|_2 = 1$ via Lemma A.2(ii), 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 Lemma A.16 in Appendix A.5, and the last equality is due to parts (ii), (v) and (iv). ∎

To prove part (ii) of Theorem 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.5 (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 Q, $p(Q, x) = x$ 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 $t \geq 0$ not coinciding with eigenvalues of A

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

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) p3. Then the condition of Lemma A.5 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

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

□

We are now ready to give a proof for Theorem 3.1

Proof of Theorem 3.1. For part (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

$$\begin{aligned} &\|\hat{\Theta}_T - \Theta\|_{\ell_2} = \|\hat{D}_T^{-1/2} \hat{\Sigma}_T \hat{D}_T^{-1/2} - \hat{D}_T^{-1/2} \Sigma \hat{D}_T^{-1/2} + \hat{D}_T^{-1/2} \Sigma \hat{D}_T^{-1/2} - D^{-1/2} \Sigma D^{-1/2}\|_{\ell_2} \\ &\leq \|\hat{D}_T^{-1/2}\|_{\ell_2}^2 \|\hat{\Sigma}_T - \Sigma\|_{\ell_2} + \|\hat{D}_T^{-1/2} \Sigma \hat{D}_T^{-1/2} - D^{-1/2} \Sigma D^{-1/2}\|_{\ell_2}. \end{aligned} \tag{A.5}$$

Invoking Lemmas A.3 and A.4(v), we conclude that the first term of (A.5) is $O_p(\sqrt{n/T})$. Let's consider the second term of (A.5). 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 Lemma A.4(ii), (iv), (v) and (vi), we conclude that the second term of (A.5) is $O_p(\sqrt{n/T})$. For part (ii), it follows trivially from Lemma A.5. For part (iii), we have

$$\|\hat{\theta}_T - \theta^0\|_2 = \|(E^\intercal W E)^{-1} E^\intercal W\|_{\ell_2} \|D_n^+\|_{\ell_2} \|\log \hat{\Theta}_T - \log \Theta\|_F \leq$$

$$\|(E^\intercal W E)^{-1} E^\intercal W\|_{\ell_2} \sqrt{n} \|\log \hat{\Theta}_T - \log \Theta\|_{\ell_2} = O(\sqrt{\varpi\kappa(W)/n})\sqrt{n} O_p(\sqrt{n/T}) = O_p\left(\sqrt{\frac{n\varpi\kappa(W)}{T}}\right),$$

where the first inequality is due to (A.8), and the second equality is due to (A.14) and parts (i)-(ii) of this theorem. □

A.4 The Proof of Theorem 3.2

In this subsection, we give a proof for Theorem 3.2. We will first give some preliminary lemmas leading to the proof of this theorem.

The following lemma linearizes the matrix logarithm.

Lemma A.6. *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*

$$\left\|[t(A - I) + I]^{-1} tB\right\|_{\ell_2} \leq C < 1 \tag{A.6}$$

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 lemma 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) Theorem 4.3.1 p181). Lemma A.6 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) (11.10) p272); however, this lemma 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.3 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.6) to invoke Lemma A.15 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 $\text{mineval}[t(A-I)+I]>C'>0$. □

Lemma A.7. *Suppose Assumptions 3.1(i), 3.2, 3.3(i) and 3.4(i) hold with $1/r_1+1/r_2>1$.*

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

(ii) Then $\hat{\Theta}_T$ 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 Lemma A.4(ii) and (vi). Now let's consider the minimum eigenvalue of Θ.

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

where the second equality is due to Rayleigh-Ritz theorem, and the last inequality is due to Assumption 3.4(i) and Lemma A.4(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 Theorem 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 Theorem 3.1(i), we can invoke Lemma A.14 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. □

Recall $\hat{H}_T$ defined in (3.6). The following lemma gives a rate of convergence for $\hat{H}_T$. It is also true when one replaces $\hat{H}_T$ with $\hat{H}_{T,D}$ defined in (3.5).

Lemma A.8. *Let Assumptions 3.1(i), 3.2, 3.3(i) and 3.4(i) be satisfied with* $1/r_1 + 1/r_2 > 1$. *Then we have*

$$\|H\|_{\ell_2} = O(1), \qquad \|\hat{H}_T\|_{\ell_2} = O_p(1), \qquad \|\hat{H}_T - H\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right). \tag{A.7}$$

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

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

where the second equality is due to Lemma A.16 in Appendix A.5, and the last equality is due to Lemma A.7(ii). Now,

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

where the first inequality is due to Jensen's inequality, the third equality is due to special properties of Kronecker product, the fourth equality is due to Lemma A.16 in Appendix A.5, and the last equality is because Lemma A.7 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 Theorem 3.1(i). The lemma then follows after invoking Lemma A.14 in Appendix A.5. $\square$

Lemma A.9. *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*

$$\|D_n^+\|_{\ell_2} = \|D_n^{+\intercal}\|_{\ell_2} = 1, \qquad \|D_n\|_{\ell_2} = \|D_n^\intercal\|_{\ell_2} = 2. \tag{A.8}$$

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

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

$\square$

We are now ready to give a proof for Theorem 3.2.

Proof of Theorem 3.2. We first show that (A.6) 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 Theorem 3.1(i), and the last equality is due to mineval$(t(\Theta - I) + I) > C > 0$ for some absolute constant C (implied by Lemma A.7(i)) and Assumption 3.3(i). Together with Lemma A.7(ii) and Lemma 2.12 in van der Vaart (1998), we can invoke Lemma A.6 stochastically with $A = \Theta$ and $B = \hat{\Theta}_T - \Theta$:

$$\log \hat{\Theta}_T - \log \Theta = \int_0^1 [t(\Theta - I) + I]^{-1} (\hat{\Theta}_T - \Theta)[t(\Theta - I) + I]^{-1} dt + O_p(\|\hat{\Theta}_T - \Theta\|_{\ell_2}^2). \quad \text{(A.9)}$$

(We can invoke Lemma A.6 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 3.3(i).) Note that (A.9) 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

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

Define

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

To prove Theorem 3.2, 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.

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

Define a triangular array of sigma algebras $\{\mathcal{F}_{T,n,t}, t = 0, 1, 2, \ldots, T\}$ by $\mathcal{F}_{T,n,t} := \mathcal{F}_t$ (the only non-standard thing is that this triangular array has one more subscript n). It is easy to see that $U_{D,T,n,t}$ is $\mathcal{F}_{T,n,t}$-measurable. We now show that $\{U_{D,T,n,t}, \mathcal{F}_{T,n,t}\}$ is a martingale difference sequence (i.e., $\mathbb{E}[U_{D,T,n,t}|\mathcal{F}_{T,n,t-1}] = 0$ almost surely for $t = 1, \ldots, T$). It suffices to show for all t

$$\mathbb{E}\left[(y_t - \mu)(y_t - \mu)^\intercal - \mathbb{E}[(y_t - \mu)(y_t - \mu)^\intercal]|\mathcal{F}_{T,n,t-1}\right] = 0 \qquad a.s.. \tag{A.10}$$

This is straightforward via Assumption 3.5. We now check conditions (i)-(iii) of Theorem A.4 in Appendix A.5. We first investigate at what rate the denominator $\sqrt{c^\intercal J_D c}$ goes to zero:

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

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

$$\text{maxeval}(E^\intercal E) \leq \text{tr}(E^\intercal E) \leq s \cdot n, \tag{A.11}$$

where the last inequality is due to Lemma A.1. For future reference

$$\|E\|_{\ell_2} = \|E^\intercal\|_{\ell_2} = \sqrt{\text{maxeval}(E^\intercal E)} \leq \sqrt{sn}. \tag{A.12}$$

Since the minimum eigenvalue of H is bounded away from zero by an absolute constant by Lemma A.7(i), the maximum eigenvalue of D is bounded from above by an absolute constant (Lemma A.4(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, via Assumption 3.6

$$\frac{1}{\sqrt{c^\intercal J_D c}} = O(\sqrt{s \cdot n \cdot \kappa(W)}). \tag{A.13}$$

Also 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 3.4(ii). Thus

$$\|(E^\intercal W E)^{-1} E^\intercal W\|_{\ell_2} = O(\sqrt{\varpi \kappa(W)/n}), \tag{A.14}$$

whence we have

$$\left\|c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ H(D^{-1/2} \otimes D^{-1/2})\right\|_2 = O(\sqrt{\varpi \kappa(W)/n}), \tag{A.15}$$

via (A.7) and Lemma A.4(ii). We now verify (i) and (ii) of Theorem A.4 in Appendix A.5. 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 article. We consider $|U_{D,T,n,t}|$ first.

$$\begin{aligned}
&|U_{D,T,n,t}| = \\
&\left|\frac{T^{-1/2}c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}\left[(y_t-\mu)(y_t-\mu)^\intercal - \mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right]}{\sqrt{c^\intercal J_D c}}\right| \\
&\leq \frac{T^{-1/2}\|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\|_2\|\operatorname{vec}\left[(y_t-\mu)(y_t-\mu)^\intercal - \mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right]\|_2}{\sqrt{c^\intercal J_D c}} \\
&= O\left(\sqrt{\frac{\varpi s\kappa^2(W)}{T}}\right)\left\|(y_t-\mu)(y_t-\mu)^\intercal - \mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right\|_F \\
&\leq O\left(\sqrt{\frac{n^2\varpi s\kappa^2(W)}{T}}\right)\left\|(y_t-\mu)(y_t-\mu)^\intercal - \mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right\|_\infty
\end{aligned}$$

where the second equality is due to (A.13) and (A.15). Consider

$$\begin{aligned}
&\left\|\left\|(y_t-\mu)(y_t-\mu)^\intercal - \mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right\|_\infty\right\|_{\psi_1} = \left\|\max_{1\leq i,j\leq n}\left|(y_{t,i}-\mu_i)(y_{t,j}-\mu_j) - \mathbb{E}(y_{t,i}-\mu_i)(y_{t,j}-\mu_j)\right|\right\|_{\psi_1} \\
&\leq \log(1+n^2)\max_{1\leq i,j\leq n}\left\|(y_{t,i}-\mu_i)(y_{t,j}-\mu_j) - \mathbb{E}(y_{t,i}-\mu_i)(y_{t,j}-\mu_j)\right\|_{\psi_1} \\
&\leq 2\log(1+n^2)\max_{1\leq i,j\leq n}\left\|(y_{t,i}-\mu_i)(y_{t,j}-\mu_j)\right\|_{\psi_1}
\end{aligned}$$

where the first inequality is due to Lemma 2.2.2 in van der Vaart and Wellner (1996). Assumption 3.1(i) with $r_1 = 2$ gives $\mathbb{E}\left[\exp(K_2|y_{t,i}|^2)\right] \leq K_1$ for all i. Then

$$\begin{aligned}
&\mathbb{P}\left(|(y_{t,i}-\mu_i)(y_{t,j}-\mu_j)| \geq \epsilon\right) \leq \mathbb{P}\left(|y_{t,i}-\mu_i| \geq \sqrt{\epsilon}\right) + \mathbb{P}\left(|y_{t,j}-\mu_j| \geq \sqrt{\epsilon}\right) \\
&\leq 2\exp\left[1 - (\sqrt{\epsilon}/c_1)^2\right] =: Ke^{-C\epsilon}
\end{aligned}$$

where the second inequality is due to Lemma A.2(iii). It follows from Lemma 2.2.1 in van der Vaart and Wellner (1996) that $\|(y_{t,i}-\mu_i)(y_{t,j}-\mu_j)\|_{\psi_1} \leq (1+K)/C$ for all i, j, t. Thus

$$\begin{aligned}
&\left\|\max_{1\leq t\leq T}|U_{D,T,n,t}|\right\|_{\psi_1} \leq \log(1+T)\max_{1\leq t\leq T}\left\|U_{D,T,n,t}\right\|_{\psi_1} \\
&= O\left(\log(1+T)\sqrt{\frac{n^2\varpi s\kappa^2(W)}{T}}\right)\max_{1\leq t\leq T}\left\|\left\|(y_t-\mu)(y_t-\mu)^\intercal - \mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right\|_\infty\right\|_{\psi_1} \\
&= O\left(\log(1+T)\log(1+n^2)\sqrt{\frac{n^2\varpi s\kappa^2(W)}{T}}\right)\max_{1\leq t\leq T}\max_{1\leq i,j\leq n}\left\|(y_{t,i}-\mu_i)(y_{t,j}-\mu_j)\right\|_{\psi_1} \\
&= O\left(\log(1+T)\log(1+n^2)\sqrt{\frac{n^2\varpi s\kappa^2(W)}{T}}\right) = O\left(\sqrt{\frac{n^2\varpi s\kappa^2(W)\log^2(1+T)\log^2(1+n^2)}{T}}\right) \\
&= o(1)
\end{aligned}$$

where the last equality is due to Assumption 3.3(ii). Since $\|U\|_{L_r} \leq r!\|U\|_{\psi_1}$ for any random variable U (van der Vaart and Wellner (1996), p95), we conclude that (i) and (ii) of Theorem

A.4 in Appendix A.5 are satisfied. We now verify condition (iii) of Theorem A.4 in Appendix A.5. Since we have already shown in (A.13) that $sn\kappa(W)c^\intercal J_D c$ is bounded away from zero by an absolute constant, it suffices to show

$$sn\kappa(W)\cdot\left|\frac{1}{T}\sum_{t=1}^{T}\left(c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})u_t\right)^2-c^\intercal J_D c\right|=o_p(1),$$

where $u_t:=\text{vec}\left[(y_t-\mu)(y_t-\mu)^\intercal-\mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right]$. Note that

$$\begin{aligned}
&sn\kappa(W)\cdot\left|\frac{1}{T}\sum_{t=1}^{T}\left(c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})u_t\right)^2-c^\intercal J_D c\right|\\
&\leq sn\kappa(W)\left\|\frac{1}{T}\sum_{t=1}^{T}u_tu_t^\intercal-V\right\|_\infty\|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\|_1^2\\
&\leq sn^3\kappa(W)\left\|\frac{1}{T}\sum_{t=1}^{T}u_tu_t^\intercal-V\right\|_\infty\|c^\intercal(E^\intercal WE)^{-1}E^\intercal WD_n^+H(D^{-1/2}\otimes D^{-1/2})\|_2^2\\
&\leq sn^3\kappa(W)\left\|\frac{1}{T}\sum_{t=1}^{T}u_tu_t^\intercal-V\right\|_\infty\|(E^\intercal WE)^{-1}E^\intercal W\|_{\ell_2}^2\|D_n^+\|_{\ell_2}^2\|H\|_{\ell_2}^2\|D^{-1/2}\otimes D^{-1/2}\|_{\ell_2}^2\\
&=O_p(sn^3\kappa(W))\sqrt{\frac{\log n}{T}}\cdot\frac{\varpi\kappa(W)}{n}=O_p\left(\sqrt{\frac{s^2n^4\kappa^4(W)\log n\cdot\varpi^2}{T}}\right)=o_p(1)
\end{aligned}$$

where the first equality is due to Lemma A.4(ii), Lemma A.16 in Appendix A.5, (A.7), (A.14), (A.8), and the fact that $\left\|T^{-1}\sum_{t=1}^{T}u_tu_t^\intercal-V\right\|_\infty=O_p(\sqrt{\frac{\log n}{T}})$, which can be deduced from the proof of Lemma 8.2 in SM 8.3,[11] the last equality is due to Assumption 3.3(ii). Thus condition (iii) of Theorem A.4 in Appendix A.5 is verified and $t_{D,1}\xrightarrow{d}N(0,1)$.

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.

$$t_{D,1}-\hat{t}_{D,1}=\frac{A_D}{\sqrt{c^\intercal J_D c}}-\frac{\hat{A}_D}{\sqrt{c^\intercal\hat{J}_{T,D}c}}=\frac{\sqrt{sn\kappa(W)}A_D}{\sqrt{sn\kappa(W)c^\intercal J_D c}}-\frac{\sqrt{sn\kappa(W)}\hat{A}_D}{\sqrt{sn\kappa(W)c^\intercal\hat{J}_{T,D}c}}.$$

Since we have already shown in (A.13) that $sn\kappa(W)c^\intercal J_D c$ 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.

[11] To see this, write

$$\begin{aligned}
\frac{1}{T}u_tu_t^\intercal-V&=\frac{1}{T}\sum_{t=1}^{T}\left[(y_t-\mu)(y_t-\mu)^\intercal\otimes(y_t-\mu)(y_t-\mu)^\intercal\right]-\mathbb{E}\left[(y_t-\mu)(y_t-\mu)^\intercal\otimes(y_t-\mu)(y_t-\mu)^\intercal\right]\\
&-\frac{1}{T}\sum_{t=1}^{T}\left[(y_t-\mu)\otimes(y_t-\mu)\right]\cdot\mathbb{E}\left[(y_t-\mu)^\intercal\otimes(y_t-\mu)^\intercal\right]+\mathbb{E}\left[(y_t-\mu)\otimes(y_t-\mu)\right]\cdot\mathbb{E}\left[(y_t-\mu)^\intercal\otimes(y_t-\mu)^\intercal\right]\\
&-\mathbb{E}\left[(y_t-\mu)\otimes(y_t-\mu)\right]\cdot\frac{1}{T}\sum_{t=1}^{T}\left[(y_t-\mu)^\intercal\otimes(y_t-\mu)^\intercal\right]+\mathbb{E}\left[(y_t-\mu)\otimes(y_t-\mu)\right]\cdot\mathbb{E}\left[(y_t-\mu)^\intercal\otimes(y_t-\mu)^\intercal\right].
\end{aligned}$$

Then many parts of the proof of Lemma 8.2 in SM 8.3 could be recycled.

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.,

$$sn\kappa(W)|c^\intercal \hat{J}_{T,D}c - c^\intercal J_D c| = o_p(1).$$

Define

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

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

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

where $\|\cdot\|_\infty$ denotes the absolute elementwise maximum, the third equality is due to Lemma A.4(ii), Lemma A.16 in Appendix A.5, (A.7), (A.14), and (A.8), the second last equality is due to Lemma 8.2 in SM 8.3, and the last equality is due to Assumption 3.3(ii). We now prove $sn\kappa(W)|c^\intercal \tilde{J}_{T,D}c - c^\intercal J_D c| = o_p(1)$.

$$\begin{aligned}
&sn\kappa(W)|c^\intercal \tilde{J}_{T,D}c - c^\intercal J_D c| \\
&= sn\kappa(W)|c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+ \hat{H}_{T,D}(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})\hat{H}_{T,D}D_n^{+\intercal}WE(E^\intercal WE)^{-1}c \\
&\qquad - c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+ H(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})HD_n^{+\intercal}WE(E^\intercal WE)^{-1}c| \\
&\leq sn\kappa(W)\left|\text{maxeval}\left[(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})\right]\right| \|(\hat{H}_{T,D} - H)D_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2^2 \\
&\quad + 2sn\kappa(W)\|(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})HD_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2 \\
&\qquad \cdot \|(\hat{H}_{T,D} - H)D_n^{+\intercal}WE(E^\intercal WE)^{-1}c\|_2
\end{aligned} \tag{A.16}$$

where the inequality is due to Lemma A.17 in Appendix A.5. We consider the first term of (A.16) first.

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

where the second last equality is due to (A.7), (A.8), and (A.14), and the last equality is due to Assumption 3.3(ii). We now consider the second term of (A.16).

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

where the first equality is due to (A.7), (A.8), and (A.14), and the last equality is due to Assumption 3.3(ii). We have proved $|sn\kappa(W)c^\intercal \tilde{J}_{T,D}c - sn\kappa(W)c^\intercal J_D c| = o_p(1)$ and hence $|sn\kappa(W)c^\intercal \hat{J}_{T,D}c - sn\kappa(W)c^\intercal J_D c| = o_p(1)$.

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.

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

where the third equality is due to (A.7), (A.8), and (A.14), the third last equality is due to (8.23) in SM 8.3, and the last equality is due to Assumption 3.3(ii).

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

Write

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

Since the denominator of the preceding equation is bounded away from zero by an absolute constant with probability approaching one by (A.13) and that $|sn\kappa(W)c^\intercal \hat{J}_{T,D}c - sn\kappa(W)c^\intercal J_D c| = o_p(1)$, it suffices to show

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

This is straightforward:

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

where the last equality is due to Assumption 3.3(ii). □

A.5 Auxiliary Lemmas

This subsection of Appendix contains auxiliary lemmas which have been used in other subsections of Appendix. We first review definitions of nets and covering numbers.

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 denoted $|\mathcal{N}_\varepsilon|$ and is called the covering number of T (at scale ε). Equivalently, $|\mathcal{N}_\varepsilon|$ is the minimal number of balls of radius ε and with centers in T needed to cover T.

Lemma A.10. *The unit Euclidean sphere $\{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ equipped with the Euclidean metric satisfies for every $\varepsilon > 0$ that*

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

Proof. See Vershynin (2011) Lemma 5.2 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.11. *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) Lemma 5.4 p8. □

The following theorem is a version of Bernstein's inequality which accommodates strong mixing time series.

Theorem A.2 (Theorem 1 of Merlevede, Peligrad, and Rio (2011))**.** *Let $\{X_t\}_{t\in\mathbb{Z}}$ be a sequence of centered real-valued random variables. Suppose that for every $\epsilon \geq 0$, there exist absolute constants $\gamma_2 \in (0, +\infty]$ and $b \in (0, +\infty)$ such that*

$$\sup_{t\geq 1} \mathbb{P}(|X_t| \geq \epsilon) \leq \exp\left[1 - (\epsilon/b)^{\gamma_2}\right].$$

Moreover, assume its alpha mixing coefficient $\alpha(h)$ satisfies

$$\alpha(h) \leq \exp(-ch^{\gamma_1}), \qquad h \in \mathbb{N}$$

for absolute constants $c > 0$ and $\gamma_1 > 0$. Define γ by $1/\gamma := 1/\gamma_1 + 1/\gamma_2$; constants γ_1 and γ_2 need to be restricted to make sure $\gamma < 1$. Then, for any $T \geq 4$, there exist positive constants C_1, C_2, C_3, C_4, C_5 depending only on b, c, γ_1, γ_2 such that, for every $\epsilon \geq 0$,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T X_t\right| \geq \epsilon\right) \leq T\exp\left(-\frac{(T\epsilon)^\gamma}{C_1}\right) + \exp\left(-\frac{(T\epsilon)^2}{C_2(1 + C_3T)}\right) + \exp\left[-\frac{(T\epsilon)^2}{C_4T}\exp\left(\frac{(T\epsilon)^{\gamma(1-\gamma)}}{C_5(\log(T\epsilon))^\gamma}\right)\right].$$

We can use the preceding theorem to establish a rate for the maximum.

Lemma A.12. *Suppose that we have for $1 \leq i \leq n$, for every $\epsilon \geq 0$,*

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T X_{t,i}\right| \geq \epsilon\right) \leq T\exp\left(-\frac{(T\epsilon)^\gamma}{C_1}\right) + \exp\left(-\frac{(T\epsilon)^2}{C_2(1 + C_3T)}\right) + \exp\left[-\frac{(T\epsilon)^2}{C_4T}\exp\left(\frac{(T\epsilon)^{\gamma(1-\gamma)}}{C_5(\log(T\epsilon))^\gamma}\right)\right].$$

Suppose $\log n = o(T^{\frac{\gamma}{2-\gamma}})$ if $n > T$. Then

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

Proof.

$$
\begin{aligned}
&\mathbb{P}\left(\max_{1\le i\le n}\left|\frac{1}{T}\sum_{t=1}^{T}X_{t,i}\right|\ge\epsilon\right)\le\sum_{i=1}^{n}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}X_{t,i}\right|\ge\epsilon\right)\\
&\le nT\exp\left(-\frac{(T\epsilon)^{\gamma}}{C_1}\right)+n\exp\left(-\frac{(T\epsilon)^2}{C_2(1+C_3T)}\right)+n\exp\left[-\frac{(T\epsilon)^2}{C_4T}\exp\left(\frac{(T\epsilon)^{\gamma(1-\gamma)}}{C_5(\log(T\epsilon))^{\gamma}}\right)\right]
\end{aligned}
$$

We shall choose $\epsilon=C\sqrt{\log n/T}$ for some $C>0$ and consider the three terms on the right side of inequality separately. We consider the first term for the case $n\le T$

$$
\begin{aligned}
&nT\exp\left(-\frac{(T\epsilon)^{\gamma}}{C_1}\right)=\exp\left(\log(nT)-\frac{C^{\gamma}}{C_1}(T\log n)^{\gamma/2}\right)=\exp\left[(T\log n)^{\gamma/2}\left(\frac{\log(nT)}{(T\log n)^{\gamma/2}}-\frac{C^{\gamma}}{C_1}\right)\right]\\
&\le\exp\left[(T\log n)^{\gamma/2}\left(\frac{2\log T}{(T\log n)^{\gamma/2}}-\frac{C^{\gamma}}{C_1}\right)\right]=\exp\left[(T\log n)^{\gamma/2}\left(o(1)-\frac{C^{\gamma}}{C_1}\right)\right]=o(1),
\end{aligned}
$$

for large enough C. We next consider the first term for the case $n>T$

$$
\begin{aligned}
&nT\exp\left(-\frac{(T\epsilon)^{\gamma}}{C_1}\right)=\exp\left(\log(nT)-\frac{C^{\gamma}}{C_1}(T\log n)^{\gamma/2}\right)=\exp\left[(T\log n)^{\gamma/2}\left(\frac{\log(nT)}{(T\log n)^{\gamma/2}}-\frac{C^{\gamma}}{C_1}\right)\right]\\
&\le\exp\left[(T\log n)^{\gamma/2}\left(\frac{2\log n}{(T\log n)^{\gamma/2}}-\frac{C^{\gamma}}{C_1}\right)\right]=\exp\left[(T\log n)^{\gamma/2}\left(o(1)-\frac{C^{\gamma}}{C_1}\right)\right]=o(1),
\end{aligned}
$$

for large enough C given the assumption $\log n=o(T^{\frac{\gamma}{2-\gamma}})$. We consider the second term.

$$
\begin{aligned}
&n\exp\left(-\frac{(T\epsilon)^2}{C_2(1+C_3T)}\right)=\exp\left(\log n-\frac{C^2\log n}{C_2/T+C_2C_3}\right)=\exp\left[\log n\left(1-\frac{C^2}{C_2/T+C_2C_3}\right)\right]\\
&=o(1)
\end{aligned}
$$

for large enough C. We consider the third term.

$$
\begin{aligned}
&n\exp\left[-\frac{(T\epsilon)^2}{C_4T}\exp\left(\frac{(T\epsilon)^{\gamma(1-\gamma)}}{C_5(\log(T\epsilon))^{\gamma}}\right)\right]\le n\exp\left[-\frac{(T\epsilon)^2}{C_4T}\exp\left(\frac{(T\epsilon)^{\gamma(1-\gamma)}}{C_5(T\epsilon)^{\gamma}}\right)\right]\\
&=n\exp\left[-\frac{(T\epsilon)^2}{C_4T}\exp\left(\frac{1}{C_5(T\epsilon)^{\gamma^2}}\right)\right]=n\exp\left[-\frac{(T\epsilon)^2}{C_4T}(1+o(1))\right]\\
&=\exp\left[\log n-\frac{C^2\log n}{C_4}(1+o(1))\right]=o(1),
\end{aligned}
$$

for large enough C. This yields the result. □

Lemma A.13. *Let A,B be $n\times n$ positive semidefinite matrices and not both singular. Then*

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

Proof. See Horn and Johnson (1985) Problem 17 p410. □

Lemma A.14. *Let $\hat{\Omega}_n$ and Ω_n be invertible (both possibly stochastic) $n\times n$ square matrices whose dimensions could be growing. Let T be the sample size. For any matrix norm, suppose that $\|\Omega_n^{-1}\|=O_p(1)$ and $\|\hat{\Omega}_n-\Omega_n\|=O_p(a_{n,T})$ for some sequence $a_{n,T}$ with $a_{n,T}\to 0$ as $n\to\infty$, $T\to\infty$ simultaneously (joint asymptotics). Then $\|\hat{\Omega}_n^{-1}-\Omega_n^{-1}\|=O_p(a_{n,T})$.*

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

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

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

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

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

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

□

Theorem A.3 (Higham (2008) (11.1) p269; Dieci, Morini, and Papini (1996))**.** *For $A \in \mathbb{C}^{n\times n}$ with no eigenvalues lying on the closed negative real axis $(-\infty, 0]$,*

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

Lemma A.15. *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*

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

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

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

Then

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

where the first and third inequalities are due to the submultiplicative property of a matrix norm, the second inequality is due to the triangular inequality. Since A is 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. □

Lemma A.16. *Consider real matrices A $(m \times n)$ and B $(p \times q)$. Then*

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

Proof.

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

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

Lemma A.17. *Let A be a $p\times p$ symmetric matrix and $\hat{v}, v\in\mathbb{R}^p$. Then*

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

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

Theorem A.4 (McLeish (1974))**.** *Let $\{X_{n,i}, i=1,...,k_n\}$ be a martingale difference array with respect to the triangular array of σ-algebras $\{\mathcal{F}_{n,i}, i=0,...,k_n\}$ (i.e., $X_{n,i}$ is $\mathcal{F}_{n,i}$-measurable and $\mathbb{E}[X_{n,i}|\mathcal{F}_{n,i-1}]=0$ almost surely for all n and i) satisfying $\mathcal{F}_{n,i-1}\subset\mathcal{F}_{n,i}$ for all $n\geq 1$. Assume,*

(i) $\max_{i\leq k_n}|X_{n,i}|$ *is uniformly (in n) bounded in L_2 norm,*

(ii) $\max_{i\leq k_n}|X_{n,i}|\xrightarrow{p}0$, *and*

(iii) $\sum_{i=1}^{k_n}X_{n,i}^2\xrightarrow{p}1$.

Then, $S_n=\sum_{i=1}^{k_n}X_{n,i}\xrightarrow{d}N(0,1)$ as $n\to\infty$.

Acknowledgements We thank the editors, two anonymous referees, 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. The article has been presented at University of Köln, University of Messina, University of Bonn, University of Santander, SoFie 2017 in New York, CFE'16 in Sevilla, and Shanghai University of Finance and Economics, and we are grateful for discussions and comments of seminar participants. Linton gratefully acknowledges the financial support from the ERC. Tang gratefully acknowledges the financial support from Shanghai 2018 CHEN GUANG Project (project code 18CG06).

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*.

Andersen, T. G. and B. E. Sørensen (1996): "GMM Estimation of a Stochastic Volatility Model: A Monte Carlo Study," *Journal of Business & Economic Statistics*, 14, 328–352.

Anderson, T. W. (1984): *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons.

ARCHAKOV, I. AND P. R. HANSEN (2018): "A New Parametrization of Correlation Matrices," *Working Paper.*

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.

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.

CHANG, J., Y. QIU, Q. YAO, AND T. ZOU (2018): "Confidence Regions for Entries of a Large Precision Matrix," *Working Paper.*

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.

ENGLE, R. F. AND K. SHEPPARD (2001): "Theoretical and Empirical properties of Dynamic Conditional Correlation Multivariate GARCH," *Working Paper.*

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 (2011): "High-Dimensional Covariance Matrix Estimation in Approximate Factor Models," *The Annals of Statistics*, 39, 3320–3356.

——— (2013): "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 603–680.

FAN, J. AND Q. YAO (2003): *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer.

FOSDICK, B. K. AND P. D. HOFF (2014): "Separable Factor Analysis with Applications to Mortality Data," *The Annals of Applied Statistics*, 8, 120–147.

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

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, 1169–1193.

——— (2016): "Equivariant and Scale-Free Tucker Decomposition Models," *Bayesian Analysis*, 11, 627–648.

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

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

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

Johnstone, I. M. and A. Onatskiy (2018): "Testing in High-Dimensional Spiked Models," *The Annals of Statistics*, forthcoming.

Krijnen, W. P. (2004): "Convergence in Mean Square of Factor Predictors," *British Journal of Mathematical and Statistical Psychology*, 57, 311–326.

Ledoit, O. and M. Wolf (2003): "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection," *Journal of Empirical Finance*, 10, 603–621.

——— (2004): "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices," *Journal of Multivariate Analysis*, 88, 365–411.

——— (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.

——— (2017): "Direct Nonlinear Shrinkage Estimation of Large-Dimensional Covariance Matrices," *Working Paper.*

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.*

Loh, W.-L. and T.-K. Lam (2000): "Estimating Structured Correlation Matrices in Smooth Gaussian Random Field Models," *The Annals of Statistics*, 28, 880–904.

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.

McLeish, D. L. (1974): "Dependent Central Limit Theorems and Invariance Principles," *The Annals of Probability*, 2, 620–628.

Merlevede, F., M. Peligrad, and E. Rio (2011): "A Bernstein Type Inequality and Moderate Deviations for Weakly Dependent Sequences," *Probability Theory and Related Fields*, 151, 435–474.

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.

Park, S., S. Y. Hong, and O. Linton (2016): "Estimating the Quadratic Covariation Matrix for Asynchronously Observed High Frequency Stock Returns Corrupted by Additive Measurement Error," *Journal of Econometrics*, 191, 325–347.

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, 2512–2523.

Swain, A. J. (1975): "Analysis of Parametric Structures for Variance Matrices," *PhD Thesis, University of Adelaide.*

van de Geer, S., P. Buhlmann, Y. Ritov, and R. Dezeure (2014): "On Asymptotically Optimal Confidence Regions and Tests for High-dimensional Models," *The Annals of Statistics*, 42, 1166–1202.

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

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

van der Weide, R. (2002): "GO-GARCH: A Multivariate Generalized Orthogonal GARCH Model," *Journal of Applied Econometrics*, 17, 549–564.

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

Verhees, J. and T. J. Wansbeek (1990): "A Multimode Direct Product Model for Covariance Structure Analysis," *British Journal of Mathematical and Statistical Psychology*, 43, 231–240.

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

——— (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.

Wu, W.-B. and Y. N. Wu (2016): "Performance Bounds for Parameter Estimates of High-dimensional Linear Models with Correlated Errors," *Electronic Journal of Statistics*, 10, 352–379.

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

Yin, J. and H. Li (2012): "Model Selection and Estimation in the Matrix Normal Graphical Model," *Journal of Multivariate Analysis*, 107, 119–140.

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

April 16, 2019

8 Supplementary Material

This section contains supplementary materials to the main article. SM 8.1 contains additional materials related to the Kronecker product (models). SM 8.2 outlines a shrinkage approach via minimum distance to make the estimated $\exp(\log \Theta_j^0)$ indeed a correlation matrix for $j = 1, \ldots, v$. SM 8.3 gives a lemma characterising a rate for $\|\hat{V}_T - V\|_\infty$, which is used in the proofs of limiting distributions of our estimators. SM 8.4, SM 8.5, and SM 8.6 provide proofs of Theorem 3.3, Theorem 4.1, and Theorem 4.2, respectively. SM 8.7 gives proofs of Theorem 3.4 and Corollary 3.3. SM 8.8 contains miscellaneous results.

8.1 Additional Materials Related to the Kronecker Product

The following lemma proves a property of Kronecker products.

Lemma 8.1. *Suppose $v = 2, 3, \ldots$ and 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*

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

Proof. We prove by mathematical induction. We first give a proof for $v = 2$; that is,

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

Since A_1, A_2 are real symmetric, they can be orthogonally diagonalized: $A_i = U_i^\intercal \Lambda_i U_i$ for $i = 1, 2$, where U_i is orthogonal, and $\Lambda_i = \mathrm{diag}(\lambda_{i,1}, \ldots, \lambda_{i,a_i})$ is a diagonal matrix containing those a_i eigenvalues of A_i. Positive definiteness of A_1, A_2 ensures that their Kronecker product is positive definite. Then the logarithm of $A_1 \otimes A_2$ is:

$$\log(A_1 \otimes A_2) = \log[(U_1 \otimes U_2)^\intercal (\Lambda_1 \otimes \Lambda_2)(U_1 \otimes U_2)] = (U_1 \otimes U_2)^\intercal \log(\Lambda_1 \otimes \Lambda_2)(U_1 \otimes U_2), \tag{8.1}$$

[*]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.

[‡]Corresponding author. Fanhai International School of Finance and School of Economics, Fudan University, 220 Handan Road, Yangpu District, Shanghai, 200433, China. Email: hhtang@fudan.edu.cn.

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. Next,

$$\begin{aligned}
&\log(\Lambda_1 \otimes \Lambda_2) = \operatorname{diag}(\log(\lambda_{1,1}\Lambda_2), \ldots, \log(\lambda_{1,a_1}\Lambda_2)) = \operatorname{diag}(\log(\lambda_{1,1}I_{a_2}\Lambda_2), \ldots, \log(\lambda_{1,a_1}I_{a_2}\Lambda_2)) \\
&= \operatorname{diag}(\log(\lambda_{1,1}I_{a_2}) + \log(\Lambda_2), \ldots, \log(\lambda_{1,a_1}I_{a_2}) + \log(\Lambda_2)) \\
&= \operatorname{diag}(\log(\lambda_{1,1}I_{a_2}), \ldots, \log(\lambda_{1,a_1}I_{a_2})) + \operatorname{diag}(\log(\Lambda_2), \ldots, \log(\Lambda_2)) \\
&= \log(\Lambda_1) \otimes I_{a_2} + I_{a_1} \otimes \log(\Lambda_2), \qquad (8.2)
\end{aligned}$$

where the third equality holds only because $\lambda_{1,j}I_{a_2}$ and Λ_2 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):

$$\begin{aligned}
&\log(A_1 \otimes A_2) = (U_1 \otimes U_2)^\intercal \log(\Lambda_1 \otimes \Lambda_2)(U_1 \otimes U_2) = (U_1 \otimes U_2)^\intercal(\log\Lambda_1 \otimes I_{a_2} + I_{a_1} \otimes \log\Lambda_2)(U_1 \otimes U_2) \\
&= (U_1 \otimes U_2)^\intercal(\log\Lambda_1 \otimes I_{a_2})(U_1 \otimes U_2) + (U_1 \otimes U_2)^\intercal(I_{a_1} \otimes \log\Lambda_2)(U_1 \otimes U_2) \\
&= \log A_1 \otimes I_{a_2} + I_{a_1} \otimes \log A_2.
\end{aligned}$$

We now assume that this lemma is true for $v = k$. That is,

$$\begin{aligned}
&\log(A_1 \otimes A_2 \otimes \cdots \otimes A_k) \\
&= \log A_1 \otimes I_{a_2} \otimes \cdots \otimes I_{a_k} + I_{a_1} \otimes \log A_2 \otimes I_{a_3} \otimes \cdots \otimes I_{a_k} + \cdots + I_{a_1} \otimes I_{a_2} \otimes \cdots \otimes \log A_k. \qquad (8.3)
\end{aligned}$$

We prove that the lemma holds for $v = k + 1$. Let $A_{1-k} := A_1 \otimes \cdots \otimes A_k$ and $I_{a_1 \cdots a_k} := I_{a_1} \otimes \cdots \otimes I_{a_k}$.

$$\begin{aligned}
&\log(A_1 \otimes A_2 \otimes \cdots \otimes A_k \otimes A_{k+1}) = \log(A_{1-k} \otimes A_{k+1}) = \log A_{1-k} \otimes I_{a_{k+1}} + I_{a_1 \cdots a_k} \otimes \log A_{k+1} \\
&= \log A_1 \otimes I_{a_2} \otimes \cdots \otimes I_{a_k} \otimes I_{a_{k+1}} + I_{a_1} \otimes \log A_2 \otimes I_{a_3} \otimes \cdots \otimes I_{a_k} \otimes I_{a_{k+1}} + \cdots + \\
&\qquad I_{a_1} \otimes I_{a_2} \otimes \cdots \otimes \log A_k \otimes I_{a_{k+1}} + I_{a_1} \otimes \cdots \otimes I_{a_k} \otimes \log A_{k+1},
\end{aligned}$$

where the third equality is due to (8.3). Thus the lemma holds for $v = k + 1$. By induction, the lemma is true for $v = 2, 3, \ldots$. □

Next we provide two examples to illustrate the necessity of an identification restriction in order to separately identify log parameters.

Example 8.1. *Suppose that $n_1, n_2 = 2$. We have*

$$\log\Theta_1^* = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \qquad \log\Theta_2^* = \begin{pmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{pmatrix}$$

Then we can calculate

$$\log\Theta^* = \log\Theta_1^* \otimes I_2 + I_2 \otimes \log\Theta_2^* = \begin{pmatrix} a_{11}+b_{11} & b_{12} & a_{12} & 0 \\ b_{12} & a_{11}+b_{22} & 0 & a_{12} \\ a_{12} & 0 & a_{22}+b_{11} & b_{12} \\ 0 & a_{12} & b_{12} & a_{22}+b_{22} \end{pmatrix}.$$

Log parameters a_{12}, b_{12} can be separately identified from the off-diagonal entries of $\log\Theta^*$ because they appear separately. We now examine whether log parameters $a_{11}, b_{11}, a_{22}, b_{22}$ can be separately identified from diagonal entries of $\log\Theta^*$. The answer is no. We have the following linear system

$$Ax := \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{22} \\ b_{11} \\ b_{22} \end{pmatrix} = \begin{pmatrix} [\log\Theta^*]_{11} \\ [\log\Theta^*]_{22} \\ [\log\Theta^*]_{33} \\ [\log\Theta^*]_{44} \end{pmatrix} =: d.$$

Note that the rank of A is 3. There are three effective equations and four unknowns; the linear system has infinitely many solutions for x. Hence one identification restriction is needed to separately identify log parameters $a_{11}, b_{11}, a_{22}, b_{22}$. We choose to set $a_{11} = 0$.

Example 8.2. *Suppose that $n_1, n_2, n_3 = 2$. We have*

$$\log \Theta_1^* = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \qquad \log \Theta_2^* = \begin{pmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{pmatrix} \qquad \log \Theta_3^* = \begin{pmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{pmatrix}$$

Then we can calculate

$$\log \Theta^* = \log \Theta_1^* \otimes I_2 \otimes I_2 + I_2 \otimes \log \Theta_2^* \otimes I_2 + I_2 \otimes I_2 \otimes \log \Theta_3^* =$$

$$\begin{pmatrix} a_{11}+b_{11}+c_{11} & c_{12} & b_{12} & 0 & a_{12} & 0 & 0 & 0 \\ c_{12} & a_{11}+b_{11}+c_{22} & 0 & b_{12} & 0 & a_{12} & 0 & 0 \\ b_{12} & 0 & a_{11}+b_{22}+c_{11} & c_{12} & 0 & 0 & a_{12} & 0 \\ 0 & b_{12} & c_{12} & a_{11}+b_{22}+c_{22} & 0 & 0 & 0 & a_{12} \\ a_{12} & 0 & 0 & 0 & a_{22}+b_{11}+c_{11} & c_{12} & b_{12} & 0 \\ 0 & a_{12} & 0 & 0 & c_{12} & a_{22}+b_{11}+c_{22} & 0 & b_{12} \\ 0 & 0 & a_{12} & 0 & b_{12} & 0 & a_{22}+b_{22}+c_{11} & c_{12} \\ 0 & 0 & 0 & a_{12} & 0 & b_{12} & c_{12} & a_{22}+b_{22}+c_{22} \end{pmatrix}.$$

Log parameters a_{12}, b_{12}, c_{12} can be separately identified from off-diagonal entries of $\log \Theta^*$ *because they appear separately. We now examine whether log parameters $a_{11}, b_{11}, c_{11}, a_{22}, b_{22}, c_{22}$ can be separately identified from diagonal entries of* $\log \Theta^*$. *The answer is no. We have the following linear system*

$$Ax := \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{22} \\ b_{11} \\ b_{22} \\ c_{11} \\ c_{22} \end{pmatrix} = \begin{pmatrix} [\log \Theta^*]_{11} \\ [\log \Theta^*]_{22} \\ [\log \Theta^*]_{33} \\ [\log \Theta^*]_{44} \\ [\log \Theta^*]_{55} \\ [\log \Theta^*]_{66} \\ [\log \Theta^*]_{77} \\ [\log \Theta^*]_{88} \end{pmatrix} =: d.$$

Note that the rank of A is 4. There are four effective equations and six unknowns; the linear system has infinitely many solutions for x. Hence two identification restrictions are needed to separately identify log parameters $a_{11}, b_{11}, c_{11}, a_{22}, b_{22}, c_{22}$. We choose to set $a_{11} = b_{11} = 0$.

8.2 Shrinkage via Minimum Distance

Recall that in the fill and shrink method, there is no guarantee that the estimated $\exp(\log \Theta^0)$ will be a correlation matrix. However, the estimated $D^{1/2} \exp(\log \Theta^0) D^{1/2}$ will be a covariance matrix. As mentioned in the main article, one can re-normalise the estimated covariance matrix to obtain a correlation matrix. The alternative method would be to shrink $\exp(\log \Theta_j^0)$ to a correlation matrix for $j = 1, \ldots, v$.

This is easy for the $n = 2^v$ case. Consider the 2×2 submatrix Θ_1^0, with $\log \Theta_1^0$ containing log parameters θ_1^0. Given that Θ_1^0 is a correlation matrix, then we have

$$\log \Theta_1^0 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \lambda_{1,1} & 0 \\ 0 & \lambda_{1,1} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \frac{1}{2} = \begin{pmatrix} \frac{1}{2}\lambda_{1,1} + \frac{1}{2}\lambda_{1,2} & \frac{1}{2}\lambda_{1,1} - \frac{1}{2}\lambda_{1,2} \\ \frac{1}{2}\lambda_{1,1} - \frac{1}{2}\lambda_{1,2} & \frac{1}{2}\lambda_{1,1} + \frac{1}{2}\lambda_{1,2} \end{pmatrix}$$

which implies that

$$\theta_1^0 := \begin{pmatrix} \theta_{1,1}^0 \\ \theta_{1,2}^0 \\ \theta_{1,3}^0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_{1,1} \\ \lambda_{1,2} \end{pmatrix} =: C \begin{pmatrix} \lambda_{1,1} \\ \lambda_{1,2} \end{pmatrix}.$$

Further, we have

$$\Theta_1^0 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \exp(\lambda_{1,1}) & 0 \\ 0 & \exp(\lambda_{1,2}) \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \frac{1}{2}$$
$$= \begin{pmatrix} \frac{1}{2}\exp(\lambda_{1,1}) + \frac{1}{2}\exp(\lambda_{1,2}) & \frac{1}{2}\exp(\lambda_{1,1}) - \frac{1}{2}\exp(\lambda_{1,2}) \\ \frac{1}{2}\exp(\lambda_{1,1}) - \frac{1}{2}\exp(\lambda_{1,2}) & \frac{1}{2}\exp(\lambda_{1,1}) + \frac{1}{2}\exp(\lambda_{1,2}) \end{pmatrix}. \tag{8.4}$$

By observing the diagonal elements of (8.4), we must have $\frac{1}{2}\exp(\lambda_{1,1}) + \frac{1}{2}\exp(\lambda_{1,2}) = 1$ or equivalently $\lambda_{1,1} = \log\big(2 - \exp(\lambda_{1,2})\big)$. Also, we have

$$\exp(\lambda_{1,1}) - \exp(\lambda_{1,2}) = 2 - 2\exp(\lambda_{1,2}) \in [-2, 2], \tag{8.5}$$

by observing the off-diagonal elements of (8.4). From (8.5), we have $-\infty < \lambda_{1,2} \leq \log 2$.

We now consider shrinkage. Given $\theta_1^0 \in \mathbb{R}^3$ we define $\lambda_{1,2}$ as the solution of the following population objective function

$$\min_{t \in (-\infty, \log 2]} \left\| \theta_1^0 - C \begin{pmatrix} \log(2 - \exp(t)) \\ t \end{pmatrix} \right\|_2$$

Thus define the estimator $\hat{\lambda}_{1,2}$ to be the solution of the following sample objective function

$$\min_{t \in (-\infty, \log 2]} \left\| \hat{\theta}_1 - C \begin{pmatrix} \log(2 - \exp(t)) \\ t \end{pmatrix} \right\|_2,$$

where $\hat{\theta}_1$ is some fill and shrink estimator of θ_1^0. Then we calculate $\hat{\lambda}_{1,1} = \log(2 - \exp(\hat{\lambda}_{1,2}))$. This ensures that $\hat{\Theta}_{1,S}^0 := \Theta_1^0(\hat{\lambda}_{1,1}, \hat{\lambda}_{1,2})$ is a correlation matrix. We can repeat this procedure for other sub-matrices $\{\Theta_j^0\}_{j=2}^v$. The final estimate

$$\hat{\Theta}_S^0 = \hat{\Theta}_{1,S}^0 \otimes \cdots \otimes \hat{\Theta}_{v,S}^0$$

will be a correlation matrix. We acknowledge that for higher dimensional sub-matrices, this approach starts to get problematic. We leave it for future research.

8.3 A Rate for $\|\hat{V}_T - V\|_\infty$

The following lemma characterises a rate for $\|\hat{V}_T - V\|_\infty$, which is used in the proofs of limiting distributions of our estimators.

Lemma 8.2. *Let Assumptions 3.1(i) and 3.2 be satisfied with $1/\gamma := 1/r_1 + 1/r_2 > 1$. Suppose $\log n = o(T^{\frac{\gamma}{2-\gamma}})$ if $n > T$. Then*

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

Proof. Let $\tilde{y}_{t,i}$ denote $y_{t,i} - \bar{y}_i$, similarly for $\tilde{y}_{t,j}, \tilde{y}_{t,k}, \tilde{y}_{t,\ell}$, where $i, j, k, \ell = 1, \ldots, n$. Let $\dot{y}_{t,i}$

denote $y_{t,i} - \mu_i$, similarly for $\dot{y}_{t,j}, \dot{y}_{t,k}, \dot{y}_{t,\ell}$ where $i, j, k, \ell = 1, \dots, n$.

$$\|\hat{V}_T - V\|_\infty := \max_{1\leq a,b\leq n^2} |\hat{V}_{T,a,b} - V_{a,b}| = \max_{1\leq i,j,k,\ell\leq n} |\hat{V}_{T,i,j,k,\ell} - V_{i,j,k,\ell}|$$

$$\leq \max_{1\leq i,j,k,\ell\leq n} \left| \frac{1}{T}\sum_{t=1}^{T} \tilde{y}_{t,i}\tilde{y}_{t,j}\tilde{y}_{t,k}\tilde{y}_{t,\ell} - \frac{1}{T}\sum_{t=1}^{T} \dot{y}_{t,i}\dot{y}_{t,j}\dot{y}_{t,k}\dot{y}_{t,\ell} \right| \tag{8.6}$$

$$+ \max_{1\leq i,j,k,\ell\leq n} \left| \frac{1}{T}\sum_{t=1}^{T} \dot{y}_{t,i}\dot{y}_{t,j}\dot{y}_{t,k}\dot{y}_{t,\ell} - \mathbb{E}[\dot{y}_{t,i}\dot{y}_{t,j}\dot{y}_{t,k}\dot{y}_{t,\ell}] \right| \tag{8.7}$$

$$+ \max_{1\leq i,j,k,\ell\leq n} \left| \left(\frac{1}{T}\sum_{t=1}^{T} \tilde{y}_{t,i}\tilde{y}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T} \tilde{y}_{t,k}\tilde{y}_{t,\ell}\right) - \left(\frac{1}{T}\sum_{t=1}^{T} \dot{y}_{t,i}\dot{y}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T} \dot{y}_{t,k}\dot{y}_{t,\ell}\right) \right| \tag{8.8}$$

$$+ \max_{1\leq i,j,k,\ell\leq n} \left| \left(\frac{1}{T}\sum_{t=1}^{T} \dot{y}_{t,i}\dot{y}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T} \dot{y}_{t,k}\dot{y}_{t,\ell}\right) - \mathbb{E}[\dot{y}_{t,i}\dot{y}_{t,j}]\mathbb{E}[\dot{y}_{t,k}\dot{y}_{t,\ell}] \right| \tag{8.9}$$

Display (8.7)

Assumption 3.1(i) says that for all t, there exist absolute constants $K_1 > 1, K_2 > 0, r_1 > 0$ such that

$$\mathbb{E}\left[\exp\left(K_2|y_{t,i}|^{r_1}\right)\right] \leq K_1 \qquad \text{for all } i = 1, \dots, n.$$

By repeated using Lemma A.2 in Appendix A.3, we have for all $i, j, k, \ell = 1, 2, \dots, n$, every $\epsilon \geq 0$, absolute constants $b_1, c_1, b_2, c_2, b_3, c_3 > 0$ such that

$$\begin{aligned}
\mathbb{P}(|y_{t,i}| \geq \epsilon) &\leq \exp\left[1 - (\epsilon/b_1)^{r_1}\right] \\
\mathbb{P}(|\dot{y}_{t,i}| \geq \epsilon) &\leq \exp\left[1 - (\epsilon/c_1)^{r_1}\right] \\
\mathbb{P}(|\dot{y}_{t,i}\dot{y}_{t,j}| \geq \epsilon) &\leq \exp\left[1 - (\epsilon/b_2)^{r_3}\right] \\
\mathbb{P}(|\dot{y}_{t,i}\dot{y}_{t,j} - \mathbb{E}[\dot{y}_{t,i}\dot{y}_{t,j}]| \geq \epsilon) &\leq \exp\left[1 - (\epsilon/c_2)^{r_3}\right] \\
\mathbb{P}(|\dot{y}_{t,i}\dot{y}_{t,j}\dot{y}_{t,k}\dot{y}_{t,\ell}| \geq \epsilon) &\leq \exp\left[1 - (\epsilon/b_3)^{r_4}\right] \\
\mathbb{P}(|\dot{y}_{t,i}\dot{y}_{t,j}\dot{y}_{t,k}\dot{y}_{t,\ell} - \mathbb{E}[\dot{y}_{t,i}\dot{y}_{t,j}\dot{y}_{t,k}\dot{y}_{t,\ell}]| \geq \epsilon) &\leq \exp\left[1 - (\epsilon/c_3)^{r_4}\right]
\end{aligned}$$

where $r_3 \in (0, r_1/2]$ and $r_4 \in (0, r_1/4]$. Use the assumption $1/r_1 + 1/r_2 > 1$ to invoke Theorem A.2 followed by Lemma A.12 in Appendix A.5 to get

$$\max_{1\leq i,j,k,\ell\leq n} \left| \frac{1}{T}\sum_{t=1}^{T} \dot{y}_{t,i}\dot{y}_{t,j}\dot{y}_{t,k}\dot{y}_{t,\ell} - \mathbb{E}\dot{y}_{t,i}\dot{y}_{t,j}\dot{y}_{t,k}\dot{y}_{t,\ell} \right| = O_p\left(\sqrt{\frac{\log n}{T}}\right). \tag{8.10}$$

Display (8.9)

We now consider (8.9).

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

$$\leq \max_{1\leq i,j,k,\ell\leq n} \left| \left(\frac{1}{T}\sum_{t=1}^{T} \dot{y}_{t,i}\dot{y}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T} \dot{y}_{t,k}\dot{y}_{t,\ell} - \mathbb{E}[\dot{y}_{t,k}\dot{y}_{t,\ell}]\right) \right| \tag{8.11}$$

$$+ \max_{1\leq i,j,k,\ell\leq n} \left| \mathbb{E}[\dot{y}_{t,k}\dot{y}_{t,\ell}]\left(\frac{1}{T}\sum_{t=1}^{T} \dot{y}_{t,i}\dot{y}_{t,j} - \mathbb{E}[\dot{y}_{t,i}\dot{y}_{t,j}]\right) \right|. \tag{8.12}$$

Consider (8.11).

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

where the first equality is due to Lemma A.2(ii) in Appendix A.3, Theorem A.2 and Lemma A.12 in Appendix A.5. Now consider (8.12).

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

where the equality is due to Lemma A.2(ii) in Appendix A.3, Theorem A.2 and Lemma A.12 in Appendix A.5. Thus

$$
\max_{1\leq i,j,k,\ell\leq n}\left|\left(\frac{1}{T}\sum_{t=1}^{T}\dot{y}_{t,i}\dot{y}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\dot{y}_{t,k}\dot{y}_{t,\ell}\right)-\mathbb{E}[\dot{y}_{t,i}\dot{y}_{t,j}]\mathbb{E}[\dot{y}_{t,k}\dot{y}_{t,\ell}]\right|=O_p\left(\sqrt{\frac{\log n}{T}}\right). \tag{8.13}
$$

Display (8.6)

We first give a rate for $\max_{1\leq i\leq n}|\bar{y}_i-\mu_i|$. The index i is arbitrary and could be replaced with j,k,ℓ. Invoking Lemma A.12 in Appendix A.5, we have

$$
\max_{1\leq i\leq n}|\bar{y}_i-\mu_i|=\max_{1\leq i\leq n}\left|\frac{1}{T}\sum_{t=1}^{T}(y_{t,i}-\mu_i)\right|=O_p\left(\sqrt{\frac{\log n}{T}}\right). \tag{8.14}
$$

Then we also have

$$
\max_{1\leq i\leq n}|\bar{y}_i|=\max_{1\leq i\leq n}|\bar{y}_i-\mu_i+\mu_i|\leq \max_{1\leq i\leq n}|\bar{y}_i-\mu_i|+\max_{1\leq i\leq n}|\mu_i|=O_p\left(\sqrt{\frac{\log n}{T}}\right)+O(1)=O_p(1). \tag{8.15}
$$

We now consider (8.6):

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

With expansion, simplification and recognition that the indices i,j,k,ℓ are completely symmetric, we can bound (8.6) by

$$
\max_{1\leq i,j,k,\ell\leq n}\left|\bar{y}_i\bar{y}_j\bar{y}_k\bar{y}_\ell-\mu_i\mu_j\mu_k\mu_\ell\right| \tag{8.16}
$$

$$
+4\max_{1\leq i,j,k,\ell\leq n}\left|\bar{y}_i\left(\bar{y}_j\bar{y}_k\bar{y}_\ell-\mu_j\mu_k\mu_\ell\right)\right| \tag{8.17}
$$

$$
+6\max_{1\leq i,j,k,\ell\leq n}\left|\left(\frac{1}{T}\sum_{t=1}^{T}y_{t,i}y_{t,j}\right)\left(\bar{y}_k\bar{y}_\ell-\mu_k\mu_\ell\right)\right| \tag{8.18}
$$

$$
+4\max_{1\leq i,j,k,\ell\leq n}\left|\left(\frac{1}{T}\sum_{t=1}^{T}y_{t,i}y_{t,j}y_{t,k}\right)\left(\bar{y}_\ell-\mu_\ell\right)\right|. \tag{8.19}
$$

We consider (8.16) first. (8.16) can be bounded by repeatedly invoking triangular inequalities (e.g., inserting terms like $\mu_i \bar{y}_j \bar{y}_k \bar{y}_\ell$) using Lemma A.2(ii) in Appendix A.3, (8.15) and (8.14). (8.16) is of order $O_p(\sqrt{\log n/T})$. (8.17) is of order $O_p(\sqrt{\log n/T})$ by a similar argument. (8.18) and (8.19) are of the same order $O_p(\sqrt{\log n/T})$ using a similar argument provided that both $\max_{1\leq i,j\leq n} |\sum_{t=1}^T y_{t,i}y_{t,j}|/T$ and $\max_{1\leq i,j,k\leq n} |\sum_{t=1}^T y_{t,i}y_{t,j}y_{t,k}|/T$ are $O_p(1)$; these follow from Lemma A.2(ii) in Appendix A.3, Theorem A.2 and Lemma A.12 in Appendix A.5. Thus

$$\max_{1\leq i,j,k,\ell\leq n}\left|\frac{1}{T}\sum_{t=1}^T \tilde{y}_{t,i}\tilde{y}_{t,j}\tilde{y}_{t,k}\tilde{y}_{t,\ell} - \frac{1}{T}\sum_{t=1}^T \dot{y}_{t,i}\dot{y}_{t,j}\dot{y}_{t,k}\dot{y}_{t,\ell}\right| = O_p(\sqrt{\log n/T}). \tag{8.20}$$

Display (8.8)

We now consider (8.8).

$$\max_{1\leq i,j,k,\ell\leq n}\left|\left(\frac{1}{T}\sum_{t=1}^T \tilde{y}_{t,i}\tilde{y}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^T \tilde{y}_{t,k}\tilde{y}_{t,\ell}\right) - \left(\frac{1}{T}\sum_{t=1}^T \dot{y}_{t,i}\dot{y}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^T \dot{y}_{t,k}\dot{y}_{t,\ell}\right)\right|$$
$$\leq \max_{1\leq i,j,k,\ell\leq n}\left|\left(\frac{1}{T}\sum_{t=1}^T \tilde{y}_{t,i}\tilde{y}_{t,j}\right)\left(\frac{1}{T}\sum_{t=1}^T (\tilde{y}_{t,k}\tilde{y}_{t,\ell} - \dot{y}_{t,k}\dot{y}_{t,\ell})\right)\right| \tag{8.21}$$
$$+ \max_{1\leq i,j,k,\ell\leq n}\left|\left(\frac{1}{T}\sum_{t=1}^T \dot{y}_{t,k}\dot{y}_{t,\ell}\right)\left(\frac{1}{T}\sum_{t=1}^T (\tilde{y}_{t,i}\tilde{y}_{t,j} - \dot{y}_{t,i}\dot{y}_{t,j})\right)\right| \tag{8.22}$$

It suffices to give a bound for (8.21) as the bound for (8.22) is of the same order and follows through similarly. First, it is easy to show that $\max_{1\leq i,j\leq n} |\frac{1}{T}\sum_{t=1}^T \tilde{y}_{t,i}\tilde{y}_{t,j}| = \max_{1\leq i,j\leq n} |\frac{1}{T}\sum_{t=1}^T y_{t,i}y_{t,j} - \bar{y}_i\bar{y}_j| = O_p(1)$ (using Lemma A.2(ii) in Appendix A.3 and Lemma A.12 in Appendix A.5). Next

$$\max_{1\leq k,\ell\leq n}\left|\frac{1}{T}\sum_{t=1}^T (\tilde{y}_{t,k}\tilde{y}_{t,\ell} - \dot{y}_{t,k}\dot{y}_{t,\ell})\right| = \max_{1\leq k,\ell\leq n}\left|-(\bar{y}_k-\mu_k)(\bar{y}_\ell-\mu_\ell)\right| = O_p\left(\frac{\log n}{T}\right). \tag{8.23}$$

The lemma follows after summing up the rates for (8.10), (8.13), (8.20) and (8.23). □

8.4 Proof of Theorem 3.3

In this subsection, we give a proof for Theorem 3.3. We will first give a preliminary lemma leading to the proof of this theorem.

Lemma 8.3. *Let Assumptions 3.1(i), 3.2, 3.3(i) and 3.4(i) hold with* $1/r_1 + 1/r_2 > 1$. *Then we have*

$$\|P\|_{\ell_2} = O(1), \qquad \|\hat{P}_T\|_{\ell_2} = O_p(1), \qquad \|\hat{P}_T - P\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right). \tag{8.24}$$

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.

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

where the second equality is due to (A.8) and Lemma A.16 in Appendix A.5, and last equality is due to Lemma A.7(ii). Now,

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

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

We are now ready to give a poof for Theorem 3.3.

Proof of Theorem 3.3. We write

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

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

Define

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

To prove Theorem 3.3, 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 3.2, so we will be concise for the parts which are almost identical to those of Theorem 3.2.

8.4.1 $t_1\xrightarrow{d}N(0,1)$

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

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

Again it is straightforward to show that $\{U_{T,n,t},\mathcal{F}_{T,n,t}\}$ is a martingale difference sequence. We first investigate at what rate the denominator $\sqrt{c^\intercal Jc}$ goes to zero:

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

where the second inequality is due to Assumption 3.7(ii). Using (A.11), we have

$$\frac{1}{\sqrt{c^{\intercal}Jc}} = O(\sqrt{s^2 \cdot n \cdot \kappa^2(W) \cdot \varpi}). \tag{8.25}$$

Verification of conditions (i)-(iii) of Theorem A.4 in Appendix A.5 will be exactly the same as those in Section A.4.1, so we omit the details in the interest of space.

8.4.2 $t_1 - \hat{t}_1 = o_p(1)$

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.

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

Since we have already shown in (8.25) that $s^2 n\kappa^2(W)\varpi c^{\intercal}Jc$ 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.

8.4.3 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.,

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

Define

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

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

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

where $\|\cdot\|_\infty$ denotes the absolute elementwise maximum, the third equality is due to Lemma A.4(v), Lemma A.16 in Appendix A.5, (A.7), (A.14), (A.8) and (8.24), the second last equality is due to Lemma 8.2 in SM 8.3, and the last equality is due to Assumption 3.3(ii).

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

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

We use triangular inequality again

$$s^2 n\kappa^2(W)\varpi|c^\intercal \tilde{J}_T c - c^\intercal J c| \leq$$
$$s^2 n\kappa^2(W)\varpi|c^\intercal \tilde{J}_T c - c^\intercal \tilde{J}_{T,a} c| + s^2 n\kappa^2(W)\varpi|c^\intercal \tilde{J}_{T,a} c - c^\intercal \tilde{J}_{T,b} c| + s^2 n\kappa^2(W)\varpi|c^\intercal \tilde{J}_{T,b} c - c^\intercal J c|. \tag{8.26}$$

We consider the first term on the right side of (8.26).

$$\begin{aligned}
&s^2 n\kappa^2(W)\varpi|c^\intercal \tilde{J}_T c - c^\intercal \tilde{J}_{T,a} c| = \\
&s^2 n\kappa^2(W)\varpi|c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ \hat{H}_T \hat{P}_T (\hat{D}_T^{-1/2} \otimes \hat{D}_T^{-1/2}) V (\hat{D}_T^{-1/2} \otimes \hat{D}_T^{-1/2}) \hat{P}_T^\intercal \hat{H}_T D_n^{+\intercal} W E (E^\intercal W E)^{-1} c \\
&\quad - c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ \hat{H}_T \hat{P}_T (D^{-1/2} \otimes D^{-1/2}) V (D^{-1/2} \otimes D^{-1/2}) \hat{P}_T^\intercal \hat{H}_T D_n^{+\intercal} W E (E^\intercal W E)^{-1} c| \\
&\leq s^2 n\kappa^2(W)\varpi\big|\text{maxeval}(V)\big| \, \|(\hat{D}_T^{-1/2} \otimes \hat{D}_T^{-1/2} - D^{-1/2} \otimes D^{-1/2}) \hat{P}_T^\intercal \hat{H}_T D_n^{+\intercal} W E (E^\intercal W E)^{-1} c\|_2^2 \\
&\quad + s^2 n\kappa^2(W)\varpi \|V (D^{-1/2} \otimes D^{-1/2}) \hat{P}_T^\intercal \hat{H}_T D_n^{+\intercal} W E (E^\intercal W E)^{-1} c\|_2 \\
&\quad\quad \cdot \|(\hat{D}_T^{-1/2} \otimes \hat{D}_T^{-1/2} - D^{-1/2} \otimes D^{-1/2}) \hat{P}_T^\intercal \hat{H}_T D_n^{+\intercal} W E (E^\intercal W E)^{-1} c\|_2
\end{aligned} \tag{8.27}$$

where the inequality is due to Lemma A.17 in Appendix A.5. We consider the first term of (8.27) first.

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

where the second last equality is due to (A.7), (A.8), (A.14), (8.24) and Lemma A.4(vii), and the last equality is due to Assumption 3.3(ii).

We now consider the second term of (8.27).

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

where the first equality is due to (A.7), (A.8), (A.14), (8.24) and Lemma A.4(vii), and the last equality is due to Assumption 3.3(ii). We have proved $s^2 n\kappa^2(W)\varpi|c^\intercal \tilde{J}_T c - c^\intercal \tilde{J}_{T,a} c| = o_p(1)$.

We consider the second term on the right hand side of (8.26).

$$\begin{aligned}
&s^2 n\kappa^2(W)\varpi|c^\intercal \tilde{J}_{T,a} c - c^\intercal \tilde{J}_{T,b} c| = \\
&s^2 n\kappa^2(W)\varpi|c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ \hat{H}_T \hat{P}_T (D^{-1/2} \otimes D^{-1/2}) V (D^{-1/2} \otimes D^{-1/2}) \hat{P}_T^\intercal \hat{H}_T D_n^{+\intercal} W E (E^\intercal W E)^{-1} c \\
&\quad - c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ \hat{H}_T P (D^{-1/2} \otimes D^{-1/2}) V (D^{-1/2} \otimes D^{-1/2}) P^\intercal \hat{H}_T D_n^{+\intercal} W E (E^\intercal W E)^{-1} c| \\
&\leq s^2 n\kappa^2(W)\varpi\big|\text{maxeval}[(D^{-1/2} \otimes D^{-1/2}) V (D^{-1/2} \otimes D^{-1/2})]\big| \, \|(\hat{P}_T - P)^\intercal \hat{H}_T D_n^{+\intercal} W E (E^\intercal W E)^{-1} c\|_2^2 \\
&\quad + 2s^2 n\kappa^2(W)\varpi \|(D^{-1/2} \otimes D^{-1/2}) V (D^{-1/2} \otimes D^{-1/2}) P^\intercal \hat{H}_T D_n^{+\intercal} W E (E^\intercal W E)^{-1} c\|_2 \\
&\quad\quad \cdot \|(\hat{P}_T - P)^\intercal \hat{H}_T D_n^{+\intercal} W E (E^\intercal W E)^{-1} c\|_2
\end{aligned} \tag{8.28}$$

where the inequality is due to Lemma A.17 in Appendix A.5. We consider the first term of (8.28) first.

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

where the second last equality is due to (A.7), (A.8), (A.14), and (8.24), and the last equality is due to Assumption 3.3(ii).

We now consider the second term of (8.28).

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

where the first equality is due to (A.7), (A.8), (A.14), and (8.24), and the last equality is due to Assumption 3.3(ii). We have proved $s^2n\kappa^2(W)\varpi|c^\intercal \tilde{J}_{T,a}c - c^\intercal \tilde{J}_{T,b}c| = o_p(1)$.

We consider the third term on the right hand side of (8.26).

$$
\begin{aligned}
&s^2n\kappa^2(W)\varpi|c^\intercal \tilde{J}_{T,b}c - c^\intercal Jc| = \\
&s^2n\kappa^2(W)\varpi|c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+\hat{H}_T P(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal \hat{H}_T D_n^{+\intercal} WE(E^\intercal WE)^{-1}c \\
&\qquad - c^\intercal (E^\intercal WE)^{-1}E^\intercal WD_n^+ HTP(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal HD_n^{+\intercal} WE(E^\intercal WE)^{-1}c| \\
&\leq s^2n\kappa^2(W)\varpi\big|\text{maxeval}[P(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal]\big| \,\|(\hat{H}_T - H)D_n^{+\intercal} WE(E^\intercal WE)^{-1}c\|_2^2 \\
&\quad + 2s^2n\kappa^2(W)\varpi\|P(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})P^\intercal HD_n^{+\intercal} WE(E^\intercal WE)^{-1}c\|_2 \\
&\qquad \cdot \|(\hat{H}_T - H)D_n^{+\intercal} WE(E^\intercal WE)^{-1}c\|_2
\end{aligned}
\tag{8.29}
$$

where the inequality is due to Lemma A.17 in Appendix A.5. We consider the first term of (8.29) first.

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

where the second last equality is due to (A.7), (A.8), and (A.14), and the last equality is due to Assumption 3.3(ii).

We now consider the second term of (8.29).

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

where the first equality is due to (A.7), (A.8), and (A.14), and the last equality is due to Assumption 3.3(ii). We have proved $s^2n\kappa^2(W)\varpi|c^\intercal \tilde{J}_{T,b}c - c^\intercal Jc| = o_p(1)$. Hence we have proved $s^2n\kappa^2(W)\varpi|c^\intercal \tilde{J}_T c - c^\intercal Jc| = o_p(1)$.

8.4.4 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.,

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

Note that

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

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

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

where the second equality is due to Assumption 3.7(i), the third equality is due to Lemma A.3, and final equality is due to Assumption 3.3(ii).

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

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

where the third last equality is due to (8.23), the last equality is due to Assumption 3.3(ii), and the second equality is due to (A.7), (A.8), (A.14), and the fact that

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

8.4.5 $\hat{t}_2 = o_p(1)$

Write

$$\hat{t}_2 = \frac{\sqrt{T}\sqrt{s^2 n\kappa^2(W)\varpi} c^\intercal (E^\intercal W E)^{-1} E^\intercal W D_n^+ \operatorname{vec} O_p(\|\hat{\Theta}_T - \Theta\|_{\ell_2}^2)}{\sqrt{s^2 n\kappa^2(W)\varpi c^\intercal \hat{J}_T c}}.$$

Since the denominator of the preceding equation is bounded away from zero by an absolute constant with probability approaching one by (8.25) and that $s^2 n\kappa^2(W)\varpi|c^\intercal \hat{J}_T c - c^\intercal J c| = o_p(1)$, 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 3.3(ii). □

8.5 Proof of Theorem 4.1

In this subsection, we give a proof for Theorem 4.1. We first give a useful lemma which is used in the proof of Theorem 4.1.

Lemma 8.4 (Magnus and Neudecker (2007) p218)**.** *Let ϕ be a twice differentiable real-valued function of an $n \times q$ matrix X. Then the following two relationships hold between the second differential and the Hessian matrix of ϕ at X:*

$$d^2\phi(X) = tr\left[B(dX)^\intercal C dX\right] \quad \Longleftrightarrow \quad \frac{\partial^2\phi(X)}{\partial(\operatorname{vec} X)\partial(\operatorname{vec} X)^\intercal} = \frac{1}{2}(B^\intercal \otimes C + B \otimes C^\intercal)$$

and

$$d^2\phi(X) = tr\left[B(dX) C dX\right] \quad \Longleftrightarrow \quad \frac{\partial^2\phi(X)}{\partial(\operatorname{vec} X)\partial(\operatorname{vec} X)^\intercal} = \frac{1}{2}K_{qn}(B^\intercal \otimes C + C^\intercal \otimes B).$$

We are now ready to give the proof of Theorem 4.1.

Proof of Theorem 4.1. For part (i), letting A denote $D^{-1/2}\tilde{\Sigma}_T D^{-1/2}$, we take the first differential of $\ell_{T,D}(\theta,\mu)$ with respect to $\Omega(\theta)$:

$$
\begin{aligned}
d\ell_{T,D}(\theta,\mu) &= -\frac{T}{2}d\log|e^{\Omega}| - \frac{1}{2}d\sum_{t=1}^{T}\operatorname{tr}\left[(y_t-\mu)^{\intercal}D^{-1/2}e^{-\Omega}D^{-1/2}(y_t-\mu)\right]\\
&= -\frac{T}{2}d\log|e^{\Omega}| - \frac{T}{2}d\operatorname{tr}\left[D^{-1/2}\frac{1}{T}\sum_{t=1}^{T}(y_t-\mu)(y_t-\mu)^{\intercal}D^{-1/2}e^{-\Omega}\right]\\
&= -\frac{T}{2}d\log|e^{\Omega}| - \frac{T}{2}d\operatorname{tr}\left[Ae^{-\Omega}\right] = -\frac{T}{2}\operatorname{tr}(e^{-\Omega}de^{\Omega}) - \frac{T}{2}\operatorname{tr}\left(Ade^{-\Omega}\right)\\
&= -\frac{T}{2}\operatorname{tr}(e^{-\Omega}de^{\Omega}) + \frac{T}{2}\operatorname{tr}\left(Ae^{-\Omega}(de^{\Omega})e^{-\Omega}\right)\\
&= -\frac{T}{2}\operatorname{tr}(e^{-\Omega}de^{\Omega}) + \frac{T}{2}\operatorname{tr}\left(e^{-\Omega}Ae^{-\Omega}de^{\Omega}\right) \qquad (8.30)\\
&= \frac{T}{2}\operatorname{tr}\left[\left(e^{-\Omega}Ae^{-\Omega}-e^{-\Omega}\right)de^{\Omega}\right] = \frac{T}{2}\left(\operatorname{vec}\left[\left(e^{-\Omega}Ae^{-\Omega}-e^{-\Omega}\right)^{\intercal}\right]\right)^{\intercal}\operatorname{vec}de^{\Omega}\\
&= \frac{T}{2}\left(\operatorname{vec}\left[e^{-\Omega}Ae^{-\Omega}-e^{-\Omega}\right]\right)^{\intercal}\operatorname{vec}\left[\int_0^1 e^{(1-t)\Omega}(d\Omega)e^{t\Omega}dt\right]\\
&= \frac{T}{2}\left(\operatorname{vec}\left[e^{-\Omega}Ae^{-\Omega}-e^{-\Omega}\right]\right)^{\intercal}\left[\int_0^1 e^{t\Omega}\otimes e^{(1-t)\Omega}dt\right]d\operatorname{vec}\Omega\\
&= \frac{T}{2}\left(\operatorname{vec}\left[e^{-\Omega}Ae^{-\Omega}-e^{-\Omega}\right]\right)^{\intercal}\left[\int_0^1 e^{t\Omega}\otimes e^{(1-t)\Omega}dt\right]D_nEd\theta
\end{aligned}
$$

where the fourth equality is due to that $d\log|X| = \operatorname{tr}(X^{-1}dX)$ for any square matrix X, the fifth equality is due to that $dX^{-1} = -X^{-1}(dX)X^{-1}$, the six equality is due to the cyclic property of trace operator, the eighth equality is due to that $\operatorname{tr}(AB) = (\operatorname{vec}[A^{\intercal}])^{\intercal}\operatorname{vec}B$, the ninth equality is due to that $de^{\Omega} = \int_0^1 e^{(1-t)\Omega}(d\Omega)e^{t\Omega}dt$ (c.f. (10.15) in Higham (2008) p238), the second last equality is due to that $\operatorname{vec}(ABC) = (C^{\intercal}\otimes A)\operatorname{vec}B$, and the last equality is due to $\operatorname{vec}\Omega = D_n\operatorname{vech}\Omega = D_nE\theta$. Thus, we conclude that

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

For part (ii), the $s\times s$ block of the Hessian matrix of (4.3) corresponding to θ is more difficult to derive. There are two approaches; they give the same Hessian but sometimes it is difficult to see the equivalence because of the presence of Kronecker products, duplication matrices etc. The first approach is to differentiate the score function with respect to θ again. The second approach is to start from (8.30), take differential again, manipulate the final result into the canonical form, and extract the Hessian from the canonical form. The second approach is due to Magnus and Neudecker (2007); Minka (2000) provided an easily accessible introduction to this approach. We shall use the second approach to derive the Hessian matrix.

There are two terms in (8.30). The first term could be simplified into

$$
\begin{aligned}
&-\frac{T}{2}\operatorname{tr}(e^{-\Omega}de^{\Omega}) = -\frac{T}{2}\operatorname{tr}\left(e^{-\Omega}\int_0^1 e^{(1-t)\Omega}(d\Omega)e^{t\Omega}dt\right) = -\frac{T}{2}\int_0^1\operatorname{tr}\left(e^{-\Omega}e^{(1-t)\Omega}(d\Omega)e^{t\Omega}\right)dt\\
&= -\frac{T}{2}\int_0^1\operatorname{tr}\left(e^{-t\Omega}(d\Omega)e^{t\Omega}\right)dt = -\frac{T}{2}\int_0^1\operatorname{tr}\left(d\Omega\right)dt = -\frac{T}{2}\operatorname{tr}\left(d\Omega\right)
\end{aligned}
$$

whence we see that it is not a function of Ω ($d\Omega$ is not a function of Ω). Thus taking differential of (8.30) will cause this term drop out. We now take the differential of the second term in

(8.30):

$$
\begin{aligned}
&d\frac{T}{2}\operatorname{tr}\left(e^{-\Omega}Ae^{-\Omega}de^{\Omega}\right) = d\frac{T}{2}\operatorname{tr}\left(e^{-\Omega}Ae^{-\Omega}\int_0^1 e^{(1-t)\Omega}(d\Omega)e^{t\Omega}dt\right)\\
&= d\frac{T}{2}\int_0^1 \operatorname{tr}\left(e^{(t-1)\Omega}Ae^{-t\Omega}d\Omega\right)dt = \frac{T}{2}\int_0^1 \operatorname{tr}\left((de^{(t-1)\Omega})Ae^{-t\Omega}d\Omega + e^{(t-1)\Omega}A(de^{-t\Omega})d\Omega\right)dt\\
&= \frac{T}{2}\int_0^1 \operatorname{tr}\left(\int_0^1 e^{(1-s)(t-1)\Omega}(d(t-1)\Omega)e^{s(t-1)\Omega}dsAe^{-t\Omega}d\Omega\right)dt\\
&\qquad + \frac{T}{2}\int_0^1 \operatorname{tr}\left(e^{(t-1)\Omega}A\int_0^1 e^{-(1-s)t\Omega}(d(-t)\Omega)e^{-st\Omega}dsd\Omega\right)dt\\
&= -\frac{T}{2}\int_0^1\int_0^1 \operatorname{tr}\left(e^{-(1-s)(1-t)\Omega}(d\Omega)e^{-s(1-t)\Omega}Ae^{-t\Omega}d\Omega\right)ds\cdot(1-t)dt\\
&\qquad - \frac{T}{2}\int_0^1\int_0^1 \operatorname{tr}\left(e^{-(1-t)\Omega}Ae^{-(1-s)t\Omega}(d\Omega)e^{-st\Omega}d\Omega\right)ds\cdot tdt.
\end{aligned}
$$

We next invoke Lemma 8.4 to get

$$
\begin{aligned}
&\frac{\partial^2\ell_{T,D}(\theta,\mu)}{\partial\operatorname{vec}\Omega\partial(\operatorname{vec}\Omega)^\intercal} =\\
&-\frac{T}{2}\int_0^1\int_0^1 \frac{1}{2}K_{n,n}\left(e^{-(1-s)(1-t)\Omega}\otimes e^{-s(1-t)\Omega}Ae^{-t\Omega} + e^{-t\Omega}Ae^{-s(1-t)\Omega}\otimes e^{-(1-s)(1-t)\Omega}\right)ds\cdot(1-t)dt\\
&\qquad -\frac{T}{2}\int_0^1\int_0^1 \frac{1}{2}K_{n,n}\left(e^{-(1-s)t\Omega}Ae^{-(1-t)\Omega}\otimes e^{-st\Omega} + e^{-st\Omega}\otimes e^{-(1-t)\Omega}Ae^{-(1-s)t\Omega}\right)ds\cdot tdt\\
&= -\frac{T}{2}\int_0^1\int_0^1 \frac{1}{2}K_{n,n}\left(e^{-st\Omega}\otimes e^{-(1-s)t\Omega}Ae^{-(1-t)\Omega} + e^{-(1-t)\Omega}Ae^{-(1-s)t\Omega}\otimes e^{-st\Omega}\right)ds\cdot tdt\\
&\qquad -\frac{T}{2}\int_0^1\int_0^1 \frac{1}{2}K_{n,n}\left(e^{-(1-s)t\Omega}Ae^{-(1-t)\Omega}\otimes e^{-st\Omega} + e^{-st\Omega}\otimes e^{-(1-t)\Omega}Ae^{-(1-s)t\Omega}\right)ds\cdot tdt
\end{aligned}
$$

where the second equality is due to change of variables $1-t\mapsto t$ and $1-s\mapsto s$ for the first term only. Note that although we have used symmetry of Ω throughout the derivation, we have not yet incorporated this fact into the Hessian. In our case, there is no need to incorporate symmetry of Ω into the Hessian because our ultimate goal is to get the Hessian in terms of the unique elements of Ω, θ (see Minka (2000) for more explanations of this). Thus the final Hessian in terms of θ is

$$
\begin{aligned}
&\frac{\partial^2\ell_{T,D}(\theta,\mu)}{\partial\theta\partial\theta^\intercal} =\\
&-\frac{T}{2}\int_0^1\int_0^1 \frac{1}{2}E^\intercal D_n^\intercal K_{n,n}\left(e^{-st\Omega}\otimes e^{-(1-s)t\Omega}Ae^{-(1-t)\Omega} + e^{-(1-t)\Omega}Ae^{-(1-s)t\Omega}\otimes e^{-st\Omega}\right)ds\cdot tdtD_nE\\
&\qquad -\frac{T}{2}\int_0^1\int_0^1 \frac{1}{2}E^\intercal D_n^\intercal K_{n,n}\left(e^{-(1-s)t\Omega}Ae^{-(1-t)\Omega}\otimes e^{-st\Omega} + e^{-st\Omega}\otimes e^{-(1-t)\Omega}Ae^{-(1-s)t\Omega}\right)ds\cdot tdtD_nE\\
&= -\frac{T}{4}E^\intercal D_n^\intercal\int_0^1\int_0^1 \left(e^{-st\Omega}\otimes e^{-(1-s)t\Omega}Ae^{-(1-t)\Omega} + e^{-(1-t)\Omega}Ae^{-(1-s)t\Omega}\otimes e^{-st\Omega}\right)ds\cdot tdtD_nE\\
&\qquad -\frac{T}{4}E^\intercal D_n^\intercal\int_0^1\int_0^1 \left(e^{-(1-s)t\Omega}Ae^{-(1-t)\Omega}\otimes e^{-st\Omega} + e^{-st\Omega}\otimes e^{-(1-t)\Omega}Ae^{-(1-s)t\Omega}\right)ds\cdot tdtD_nE
\end{aligned}
$$

where the second equality is due to that $K_{n,n}D_n = D_n$ and symmetry of $K_{n,n}$ (see (52) of Magnus and Neudecker (1986)).

For part (iii), note that $\mathbb{E}[A] = \mathbb{E}[D^{-1/2}\tilde{\Sigma}_T D^{-1/2}] = \Theta = e^{\Omega}$. Then by merging terms, we have

$$\Upsilon_D = \frac{1}{2} E^\intercal D_n^\intercal \int_0^1 \int_0^1 \left(e^{-st\Omega} \otimes e^{st\Omega} + e^{st\Omega} \otimes e^{-st\Omega}\right) ds \cdot tdt D_n E.$$

To prove the equivalence between (4.4) and (4.5), it suffices to show

$$\int_0^1 \int_0^1 \left(e^{-st\Omega} \otimes e^{st\Omega} + e^{st\Omega} \otimes e^{-st\Omega}\right) ds \cdot tdt = \int_0^1 \int_0^1 e^{(t+s-1)\Omega} \otimes e^{(1-t-s)\Omega} dsdt. \tag{8.31}$$

Suppose $\Theta = e^\Omega = Q^\intercal \operatorname{diag}(\lambda_1, \ldots, \lambda_n) Q$ (orthogonal diagonalization). The eigenvalues λ_js are all positive but need not be distinct. We first consider the first term of (8.31). By definition of matrix function, we have

$$e^{-st\Omega} = Q^\intercal \operatorname{diag}(\lambda_1^{-st}, \ldots, \lambda_n^{-st}) Q \qquad e^{st\Omega} = Q^\intercal \operatorname{diag}(\lambda_1^{st}, \ldots, \lambda_n^{st}) Q$$

$$\begin{aligned}
&e^{-st\Omega} \otimes e^{st\Omega} + e^{st\Omega} \otimes e^{-st\Omega} = \\
&(Q \otimes Q)^\intercal \left[\operatorname{diag}(\lambda_1^{-st}, \ldots, \lambda_n^{-st}) \otimes \operatorname{diag}(\lambda_1^{st}, \ldots, \lambda_n^{st}) + \operatorname{diag}(\lambda_1^{st}, \ldots, \lambda_n^{st}) \otimes \operatorname{diag}(\lambda_1^{-st}, \ldots, \lambda_n^{-st})\right] (Q \otimes Q) \\
&=: (Q \otimes Q)^\intercal M_1 (Q \otimes Q),
\end{aligned}$$

where M_1 is an $n^2 \times n^2$ diagonal matrix whose $[(i-1)n + j]$th diagonal entry is $\left(\frac{\lambda_j}{\lambda_i}\right)^{st} + \left(\frac{\lambda_i}{\lambda_j}\right)^{st}$ for $i, j = 1, \ldots, n$. Thus

$$\int_0^1 \int_0^1 \left(e^{-st\Omega} \otimes e^{st\Omega} + e^{st\Omega} \otimes e^{-st\Omega}\right) ds \cdot tdt = (Q \otimes Q)^\intercal \int_0^1 \int_0^1 M_1 ds \cdot tdt (Q \otimes Q),$$

where $\int_0^1 \int_0^1 M_1 tdsdt$ is an $n^2 \times n^2$ diagonal matrix whose $[(i-1)n + j]$th diagonal entry is

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

for $i, j = 1, \ldots, n$. To see this,

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

Similarly

$$\int_0^1 \int_0^1 \left(\frac{\lambda_i}{\lambda_j}\right)^{st} tdsdt = \frac{1}{\left[\log\left(\frac{\lambda_i}{\lambda_j}\right)\right]^2} \left(\frac{\lambda_i}{\lambda_j} - 1 - \log\left(\frac{\lambda_i}{\lambda_j}\right)\right),$$

whence we have

$$\int_0^1 \int_0^1 \left[\left(\frac{\lambda_j}{\lambda_i}\right)^{st} + \left(\frac{\lambda_i}{\lambda_j}\right)^{st}\right] tdsdt = \frac{1}{\left[\log\left(\frac{\lambda_i}{\lambda_j}\right)\right]^2} \left[\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} - 2\right].$$

We now consider the second term of (8.31). By definition of matrix function, we have

$$e^{(t+s-1)\Omega} = Q^\intercal \mathrm{diag}(\lambda_1^{(t+s-1)}, \ldots, \lambda_n^{(t+s-1)})Q \qquad e^{(1-t-s)\Omega} = Q^\intercal \mathrm{diag}(\lambda_1^{(1-t-s)}, \ldots, \lambda_n^{(1-t-s)})Q$$
$$e^{(t+s-1)\Omega} \otimes e^{(1-t-s)\Omega} = (Q \otimes Q)^\intercal \left[\mathrm{diag}(\lambda_1^{(t+s-1)}, \ldots, \lambda_n^{(t+s-1)}) \otimes \mathrm{diag}(\lambda_1^{(1-t-s)}, \ldots, \lambda_n^{(1-t-s)})\right](Q \otimes Q)$$
$$=: (Q \otimes Q)^\intercal M_2 (Q \otimes Q),$$

where M_2 is an $n^2 \times n^2$ diagonal matrix whose $[(i-1)n+j]$th diagonal entry is $\left(\frac{\lambda_i}{\lambda_j}\right)^{s+t-1}$ for $i, j = 1, \ldots, n$. Thus

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

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

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

for $i, j = 1, \ldots, n$. To see this,

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

Comparing $\int_0^1 \int_0^1 M_1 t ds dt$ with $\int_0^1 \int_0^1 M_2 ds dt$, we realise (8.31) hold.

For part (iv), using the expression for $\frac{\partial \ell_{T,D}(\theta,\mu)}{\partial \theta^\intercal}$ and the fact that it has zero expectation, we have

$$
\begin{aligned}
&\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] = \frac{T}{4}E^\intercal D_n^\intercal \Psi \mathrm{var}\left(\mathrm{vec}\left(e^{-\Omega}D^{-1/2}\tilde{\Sigma}_T D^{-1/2}e^{-\Omega}\right)\right)\Psi D_n E\\
&= \frac{T}{4}E^\intercal D_n^\intercal \Psi\left(e^{-\Omega}\otimes e^{-\Omega}\right)(D^{-1/2}\otimes D^{-1/2})\mathrm{var}\left(\mathrm{vec}\left[\frac{1}{T}\sum_{t=1}^{T}(y_t-\mu)(y_t-\mu)^\intercal\right]\right)\\
&\qquad\cdot(D^{-1/2}\otimes D^{-1/2})\left(e^{-\Omega}\otimes e^{-\Omega}\right)\Psi D_n E\\
&= \frac{1}{4}E^\intercal D_n^\intercal \Psi\left(e^{-\Omega}\otimes e^{-\Omega}\right)(D^{-1/2}\otimes D^{-1/2})\mathrm{var}\left(\mathrm{vec}\left[(y_t-\mu)(y_t-\mu)^\intercal\right]\right)\\
&\qquad\cdot(D^{-1/2}\otimes D^{-1/2})\left(e^{-\Omega}\otimes e^{-\Omega}\right)\Psi D_n E\\
&= \frac{1}{4}E^\intercal D_n^\intercal \Psi\left(e^{-\Omega}\otimes e^{-\Omega}\right)(D^{-1/2}\otimes D^{-1/2})2D_nD_n^+(\Sigma\otimes\Sigma)(D^{-1/2}\otimes D^{-1/2})\left(e^{-\Omega}\otimes e^{-\Omega}\right)\Psi D_n E\\
&= \frac{1}{4}E^\intercal D_n^\intercal \Psi\left(e^{-\Omega}\otimes e^{-\Omega}\right)(D^{-1/2}\otimes D^{-1/2})(I_{n^2}+K_{n,n})(\Sigma\otimes\Sigma)(D^{-1/2}\otimes D^{-1/2})\left(e^{-\Omega}\otimes e^{-\Omega}\right)\Psi D_n E\\
&= \frac{1}{4}E^\intercal D_n^\intercal \Psi\left(e^{-\Omega}\otimes e^{-\Omega}\right)(D^{-1/2}\otimes D^{-1/2})(\Sigma\otimes\Sigma)(D^{-1/2}\otimes D^{-1/2})\left(e^{-\Omega}\otimes e^{-\Omega}\right)\Psi D_n E\\
&\qquad+\frac{1}{4}E^\intercal D_n^\intercal \Psi\left(e^{-\Omega}\otimes e^{-\Omega}\right)(D^{-1/2}\otimes D^{-1/2})K_{n,n}(\Sigma\otimes\Sigma)(D^{-1/2}\otimes D^{-1/2})\left(e^{-\Omega}\otimes e^{-\Omega}\right)\Psi D_n E\\
&= \frac{1}{4}E^\intercal D_n^\intercal \Psi\left(e^{-\Omega}\otimes e^{-\Omega}\right)(D^{-1/2}\otimes D^{-1/2})(\Sigma\otimes\Sigma)(D^{-1/2}\otimes D^{-1/2})\left(e^{-\Omega}\otimes e^{-\Omega}\right)\Psi D_n E\\
&\qquad+\frac{1}{4}E^\intercal D_n^\intercal \Psi\left(e^{-\Omega}\otimes e^{-\Omega}\right)(D^{-1/2}\otimes D^{-1/2})(\Sigma\otimes\Sigma)(D^{-1/2}\otimes D^{-1/2})\left(e^{-\Omega}\otimes e^{-\Omega}\right)K_{n,n}\Psi D_n E\\
&= \frac{1}{2}E^\intercal D_n^\intercal \Psi\left(e^{-\Omega}\otimes e^{-\Omega}\right)(D^{-1/2}\otimes D^{-1/2})(\Sigma\otimes\Sigma)(D^{-1/2}\otimes D^{-1/2})\left(e^{-\Omega}\otimes e^{-\Omega}\right)\Psi D_n E\\
&= \frac{1}{2}E^\intercal D_n^\intercal \Psi\left(e^{-\Omega}\otimes e^{-\Omega}\right)\Psi D_n E,
\end{aligned}
$$

where the third equality is due to weak stationarity of y_t and (A.10) via Assumption 3.5, the fifth equality is due to that $2D_nD_n^+ = I_{n^2}+K_{n,n}$, the seventh equality is due to that $K_{n,n}(A\otimes B) = (B\otimes A)K_{n,n}$ for arbitrary $n\times n$ matrices A and B, and the second last equality is due to

$$K_{n,n}\Psi = \int_0^1 K_{n,n}\left(e^{t\Omega}\otimes e^{(1-t)\Omega}\right)dt = \int_0^1 e^{(1-t)\Omega}\otimes e^{t\Omega}dt = \int_0^1 e^{s\Omega}\otimes e^{(1-s)\Omega}dt = \Psi,$$

via change of variable $1-t\mapsto s$. □

8.6 Proof of Theorem 4.2

In this subsection, we give a proof for Theorem 4.2. We will first give some preliminary lemmas leading to the proof of this theorem.

Lemma 8.5. *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) Corollary 6.2.32 p430. □

Define

$$\Xi := \int_0^1\int_0^1 \Theta^{t+s-1}\otimes\Theta^{1-t-s}dtds \qquad \hat{\Xi}_{T,D} := \int_0^1\int_0^1 \hat{\Theta}_{T,D}^{t+s-1}\otimes\hat{\Theta}_{T,D}^{1-t-s}dtds$$

such that Υ_D and $\hat{\Upsilon}_{T,D}$ could be denoted $\frac{1}{2}E^\intercal D_n^\intercal \Xi D_n E$ and $\frac{1}{2}E^\intercal D_n^\intercal \hat{\Xi}_{T,D} D_n E$, respectively.

Lemma 8.6. *Suppose Assumptions 3.1(i), 3.2, 3.3(i) and 3.4(i) hold with $1/r_1 + 1/r_2 > 1$. Then*

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

(ii) $\hat{\Xi}_{T,D}$ *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)

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

(iv)

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

Proof. The proofs for the first two parts are the same, so we only give one for part (i). Under assumptions of this lemma, we can invoke Lemma A.7(i) in Appendix A.4 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. We have already shown in the proof of Theorem 4.1 in SM 8.5 that eigenvalues of Ξ are

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

for $i, j = 1, \ldots, n$. This concludes the proof.

For part (iii), we have

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

First, note that for any $t, s \in [0, 1]$, $\|\hat{\Theta}_{T,D}^{t+s-1}\|_{\ell_2}$ 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 result would then follow if we show that

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

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

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

where the first inequality is due to Lemma 8.5, 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 Theorem 3.1 to get $\|\log \hat{\Theta}_{T,D} - \log \Theta\|_{\ell_2} = O_p\left(\sqrt{\frac{n}{T}}\right)$. The result follows after recognising $\exp(o_p(1)) = O_p(1)$.

The proof for part (iv) is very similar to the one which we gave in the proof of Theorem 4.1 in SM 8.5. Since $\Theta = Q^\intercal \text{diag}(\lambda_1, \dots, \lambda_n) Q$, we have $\Theta^t = Q^\intercal \text{diag}(\lambda_1^t, \dots, \lambda_n^t) Q$ and $\Theta^{1-t} = Q^\intercal \text{diag}(\lambda_1^{1-t}, \dots, \lambda_n^{1-t}) Q$. Then

$$\Theta^t \otimes \Theta^{1-t} = (Q \otimes Q)^\intercal \left[\text{diag}(\lambda_1^t, \dots, \lambda_n^t) \otimes \text{diag}(\lambda_1^{1-t}, \dots, \lambda_n^{1-t})\right] (Q \otimes Q) =: (Q \otimes Q)^\intercal M_3 (Q \otimes Q),$$

where M_3 is an $n^2 \times n^2$ diagonal matrix whose $[(i-1)n+j]$th diagonal entry is $\lambda_j \left(\frac{\lambda_i}{\lambda_j}\right)^t$ for $i, j = 1, \dots, n$. Thus

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

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

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

for $i, j = 1, \dots, n$. To see this,

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

□

Lemma 8.7. *Suppose Assumptions 3.1(i), 3.2, 3.3(i) and 3.4 hold with $1/r_1 + 1/r_2 > 1$. 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),

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

where the second equality is due to (A.8), and the last equality is due to (A.12) and Lemma 8.6(iii).

For part (ii),

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

where the second last equality is due to (8.32). □

We are now ready to give a proof for Theorem 4.2.

Proof of Theorem 4.2. We first show that $\hat{\Upsilon}_{T,D}$ is invertible with probability approaching 1, so that our estimator $\tilde{\theta}_{T,D} := \hat{\theta}_{T,D} - \hat{\Upsilon}_{T,D}^{-1}\frac{\partial \ell_{T,D}(\hat{\theta}_{T,D},\bar{y})}{\partial \theta^\intercal}/T$ 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.

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

for some absolute positive constant C with probability approaching one, where the second inequality is due to Lemma 8.6(ii), Assumption 3.4(ii), and that $D_n^\intercal D_n$ is a diagonal matrix with diagonal entries either 1 or 2. 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

$$\|\hat{\Upsilon}_{T,D}^{-1}\|_{\ell_2} = \frac{1}{\text{mineval}(\hat{\Upsilon}_{T,D})} = O_p\left(\frac{\varpi}{n}\right) \qquad \|\Upsilon_D^{-1}\|_{\ell_2} = \frac{1}{\text{mineval}(\Upsilon_D)} = O\left(\frac{\varpi}{n}\right). \tag{8.32}$$

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

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

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

$$\sqrt{T}\frac{a^\intercal}{\|a\|_2}(\tilde{\theta}_{T,D} - \theta) = \sqrt{T}\frac{a^\intercal}{\|a\|_2}\hat{\Upsilon}_{T,D}^{-1}(\hat{\Upsilon}_{T,D} - \Upsilon_D)(\hat{\theta}_{T,D} - \theta) - \frac{a^\intercal}{\|a\|_2}\hat{\Upsilon}_{T,D}^{-1}\sqrt{T}\frac{1}{T}\frac{\partial \ell_{T,D}(\theta,\bar{y})}{\partial \theta^\intercal} + \frac{o_p(1)}{\|a\|_2}.$$

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, for any $c \in \mathbb{R}^s$ with $\|c\|_2 = 1$,

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

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

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

where the first equality is due to (8.32), Lemma 8.7(i) and the rate of convergence for the minimum distance estimator $\hat{\theta}_T$ ($\hat{\theta}_{T,D}$). Thus

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

whence, if we divide by $\sqrt{c^\intercal \hat{\Upsilon}_{T,D}^{-1} c}$, we have

$$\frac{\sqrt{T} c^\intercal (\tilde{\theta}_{T,D} - \theta)}{\sqrt{c^\intercal \hat{\Upsilon}_{T,D}^{-1} c}} = \frac{-c^\intercal \hat{\Upsilon}_{T,D}^{-1} \sqrt{T} \frac{\partial \ell_{T,D}(\theta, \bar{y})}{\partial \theta^\intercal}/T}{\sqrt{c^\intercal \hat{\Upsilon}_{T,D}^{-1} c}} + \frac{\text{rem}}{\sqrt{c^\intercal \hat{\Upsilon}_{T,D}^{-1} c}} =: \hat{t}_{os,D,1} + t_{os,D,2}.$$

Define

$$t_{os,D,1} := \frac{-c^\intercal \Upsilon_D^{-1} \sqrt{T} \frac{\partial \ell_{T,D}(\theta, \mu)}{\partial \theta^\intercal}/T}{\sqrt{c^\intercal \Upsilon_D^{-1} c}}.$$

To prove Theorem 4.2, 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 $t_{os,D,2} = o_p(1)$.

8.6.1 $t_{os,D,1} \xrightarrow{d} N(0,1)$

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

$$t_{os,D,1} := \frac{-c^\intercal \Upsilon_D^{-1} \sqrt{T} \frac{\partial \ell_{T,D}(\theta, \mu)}{\partial \theta^\intercal}/T}{\sqrt{c^\intercal \Upsilon_D^{-1} c}} =$$

$$\sum_{t=1}^{T} \frac{-\frac{1}{2} c^\intercal \Upsilon_D^{-1} E^\intercal D_n^\intercal \Psi (\Theta^{-1} \otimes \Theta^{-1})(D^{-1/2} \otimes D^{-1/2}) T^{-1/2} \operatorname{vec}\left[(y_t - \mu)(y_t - \mu)^\intercal - \mathbb{E}(y_t - \mu)(y_t - \mu)^\intercal\right]}{\sqrt{c^\intercal \Upsilon_D^{-1} c}}$$

$$=: \sum_{t=1}^{T} U_{os,D,T,n,t}.$$

The proof is very similar to that of $t_{D,1} \xrightarrow{d} N(0,1)$ in Section A.4.1. It is straightforward to show that $\{U_{os,D,T,n,t}, \mathcal{F}_{T,n,t}\}$ is a martingale difference sequence. We first investigate that at what rate the denominator $\sqrt{c^\intercal \Upsilon_D^{-1} c}$ goes to zero.

$$c^\intercal \Upsilon_D^{-1} c = 2c^\intercal \left(E^\intercal D_n^\intercal \Xi D_n E\right)^{-1} c \geq 2\text{mineval}\left(\left(E^\intercal D_n^\intercal \Xi D_n E\right)^{-1}\right) = \frac{2}{\text{maxeval}\left(E^\intercal D_n^\intercal \Xi D_n E\right)}.$$

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 because of Lemma 8.6(i), (A.11) and that $D_n^\intercal D_n$ is a diagonal matrix with diagonal entries either 1 or 2. Thus we have

$$\frac{1}{\sqrt{c^\intercal \Upsilon_D^{-1} c}} = O(\sqrt{sn}). \tag{8.33}$$

We now verify (i) and (ii) of Theorem A.4 in Appendix A.5. We consider $|U_{os,D,T,n,t}|$ first.

$$
\begin{aligned}
&|U_{os,D,T,n,t}| = \\
&\left|\frac{\frac{1}{2}c^\intercal \Upsilon_D^{-1} E^\intercal D_n^\intercal \Psi(\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2})T^{-1/2}\operatorname{vec}\left[(y_t-\mu)(y_t-\mu)^\intercal - \mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right]}{\sqrt{c^\intercal \Upsilon_D^{-1} c}}\right| \\
&\leq \frac{\frac{1}{2}T^{-1/2}\left\|c^\intercal \Upsilon_D^{-1} E^\intercal D_n^\intercal \Psi(\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2})\right\|_2 \left\|\operatorname{vec}\left[(y_t-\mu)(y_t-\mu)^\intercal - \mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right]\right\|_2}{\sqrt{c^\intercal \Upsilon_D^{-1} c}} \\
&= O\left(\sqrt{\frac{s^2\varpi^2}{T}}\right)\left\|(y_t-\mu)(y_t-\mu)^\intercal - \mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right\|_F \\
&\leq O\left(\sqrt{\frac{n^2 s^2\varpi^2}{T}}\right)\left\|(y_t-\mu)(y_t-\mu)^\intercal - \mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right\|_\infty,
\end{aligned}
$$

where the second equality is due to (8.33) and that

$$
\begin{aligned}
&\left\|c^\intercal \Upsilon_D^{-1} E^\intercal D_n^\intercal \Psi(\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2})\right\|_2 \\
&\leq \|\Upsilon_D^{-1}\|_{\ell_2}\|E^\intercal\|_{\ell_2}\|D_n^\intercal\|_{\ell_2}\|\Psi\|_{\ell_2}\|\Theta^{-1}\otimes\Theta^{-1}\|_{\ell_2}\|D^{-1/2}\otimes D^{-1/2}\|_{\ell_2} = O\left(\frac{\varpi}{n}\right)\sqrt{sn} = O\left(\sqrt{\frac{s\varpi^2}{n}}\right)
\end{aligned}
$$

via (8.32) and (A.12). Next, using a similar argument which we explained in detail in Section A.4.1, we have

$$
\begin{aligned}
&\left\|\max_{1\leq t\leq T}|U_{os,D,T,n,t}|\right\|_{\psi_1} \leq \log(1+T)\max_{1\leq t\leq T}\left\|U_{os,D,T,n,t}\right\|_{\psi_1} \\
&= \log(1+T)O\left(\sqrt{\frac{n^2s^2\varpi^2}{T}}\right)\max_{1\leq t\leq T}\left\|\left\|(y_t-\mu)(y_t-\mu)^\intercal - \mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right\|_\infty\right\|_{\psi_1} \\
&= \log(1+T)\log(1+n^2)O\left(\sqrt{\frac{n^2s^2\varpi^2}{T}}\right)\max_{1\leq t\leq T}\max_{1\leq i,j\leq n}\left\|(y_{t,i}-\mu_i)(y_{t,j}-\mu_j)\right\|_{\psi_1} \\
&= O\left(\sqrt{\frac{n^2s^2\varpi^2\log^2(1+T)\log^2(1+n^2)}{T}}\right) = o(1)
\end{aligned}
$$

where the last equality is due to Assumption 3.3(iii). Since $\|U\|_{L_r}\leq r!\|U\|_{\psi_1}$ for any random variable U (van der Vaart and Wellner (1996), p95), we conclude that (i) and (ii) of Theorem A.4 in Appendix A.5 are satisfied.

We now verify condition (iii) of Theorem A.4 in Appendix A.5. Since we have already shown that $snc^\intercal \Upsilon_D^{-1} c$ is bounded away from zero by an absolute constant, it suffices to show

$$
sn\left|\frac{1}{T}\sum_{t=1}^T\left(\frac{1}{2}c^\intercal \Upsilon_D^{-1} E^\intercal D_n^\intercal \Psi(\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2})u_t\right)^2 - c^\intercal \Upsilon_D^{-1} c\right| = o_p(1),
$$

where $u_t := \operatorname{vec}\left[(y_t-\mu)(y_t-\mu)^\intercal - \mathbb{E}(y_t-\mu)(y_t-\mu)^\intercal\right]$. Under Assumptions 3.1(ii) and 3.5, we have already shown in the proof of part (iv) of Theorem 4.1 that

$$
\begin{aligned}
&c^\intercal \Upsilon_D^{-1} c = c^\intercal \Upsilon_D^{-1}\Upsilon_D\Upsilon_D^{-1} c = c^\intercal \Upsilon_D^{-1}\left(\frac{1}{2}E^\intercal D_n^\intercal \Psi(\Theta^{-1}\otimes\Theta^{-1})\Psi D_n E\right)\Upsilon_D^{-1} c \\
&= \frac{1}{4}c^\intercal \Upsilon_D^{-1} E^\intercal D_n^\intercal \Psi(\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})(\Theta^{-1}\otimes\Theta^{-1})\Psi D_n E\Upsilon_D^{-1} c.
\end{aligned}
$$

Thus

$$
\begin{aligned}
&sn\left|\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2}c^\intercal \Upsilon_D^{-1}E^\intercal D_n^\intercal \Psi(\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2})u_t\right)^2 - c^\intercal \Upsilon_D^{-1}c\right| \\
&\leq \frac{1}{4}sn\left\|\frac{1}{T}\sum_{t=1}^{T}u_t u_t^\intercal - V\right\|_\infty \left\|(D^{-1/2}\otimes D^{-1/2})(\Theta^{-1}\otimes\Theta^{-1})\Psi D_n E\Upsilon_D^{-1}c\right\|_1^2 \\
&\leq \frac{1}{4}sn^3\left\|\frac{1}{T}\sum_{t=1}^{T}u_t u_t^\intercal - V\right\|_\infty \left\|(D^{-1/2}\otimes D^{-1/2})(\Theta^{-1}\otimes\Theta^{-1})\Psi D_n E\Upsilon_D^{-1}c\right\|_2^2 \\
&\leq \frac{1}{4}sn^3\left\|\frac{1}{T}\sum_{t=1}^{T}u_t u_t^\intercal - V\right\|_\infty \|D^{-1/2}\otimes D^{-1/2}\|_{\ell_2}^2\|\Theta^{-1}\otimes\Theta^{-1}\|_{\ell_2}^2\|\Psi\|_{\ell_2}^2\|D_n\|_{\ell_2}^2\|E\|_{\ell_2}^2\|\Upsilon_D^{-1}\|_{\ell_2}^2 \\
&= O_p(sn^3)\sqrt{\frac{\log n}{T}}\cdot sn\cdot\frac{\varpi^2}{n^2} = O_p\left(\sqrt{\frac{n^4\cdot\log n\cdot\varpi^4\cdot\log^4 n}{T}}\right) = o_p(1)
\end{aligned}
$$

where the first equality is due to (8.32), (A.12) and the fact that $\left\|T^{-1}\sum_{t=1}^{T}u_t u_t^\intercal - V\right\|_\infty = O_p(\sqrt{\frac{\log n}{T}})$, which can be deduced from the proof of Lemma 8.2 in SM 8.3, and the last equality is due to Assumption 3.3(iii).

8.6.2 $\hat{t}_{os,D,1} - t_{os,D,1} = o_p(1)$

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.

$$
\hat{t}_{os,D,1} - t_{os,D,1} = \frac{\hat{A}_{os,D}}{\sqrt{c^\intercal \hat{\Upsilon}_{T,D}^{-1}c}} - \frac{A_{os,D}}{\sqrt{c^\intercal \Upsilon_D^{-1}c}} = \frac{\sqrt{sn}\hat{A}_{os,D}}{\sqrt{snc^\intercal \hat{\Upsilon}_{T,D}^{-1}c}} - \frac{\sqrt{sn}A_{os,D}}{\sqrt{snc^\intercal \Upsilon_D^{-1}c}}
$$

Since we have already shown in (8.33) that $snc^\intercal \Upsilon_D^{-1}c$ 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.

8.6.3 Denominators of $\hat{t}_{os,D,1}$ and $t_{os,D,1}$

We need to show

$$
sn|c^\intercal(\hat{\Upsilon}_{T,D}^{-1} - \Upsilon_D^{-1})c| = o_p(1).
$$

This is straightforward.

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

where the last equality is due to Assumption 3.3(iii).

8.6.4 Numerators of $\hat{t}_{os,D,1}$ and $t_{os,D,1}$

We now show

$$
\sqrt{sn}\left|c^\intercal \hat{\Upsilon}_{T,D}^{-1}\sqrt{T}\frac{\partial \ell_{T,D}(\theta,\bar{y})}{\partial\theta^\intercal}/T - c^\intercal \Upsilon_D^{-1}\sqrt{T}\frac{\partial \ell_{T,D}(\theta,\mu)}{\partial\theta^\intercal}/T\right| = o_p(1).
$$

Using triangular inequality, we have

$$
\begin{aligned}
&\sqrt{sn}\left|c^\intercal\hat{\Upsilon}_{T,D}^{-1}\sqrt{T}\frac{\partial\ell_{T,D}(\theta,\bar{y})}{\partial\theta^\intercal}/T-c^\intercal\Upsilon_D^{-1}\sqrt{T}\frac{\partial\ell_{T,D}(\theta,\mu)}{\partial\theta^\intercal}/T\right|\\
&\le\sqrt{sn}\left|c^\intercal\hat{\Upsilon}_{T,D}^{-1}\sqrt{T}\frac{\partial\ell_{T,D}(\theta,\bar{y})}{\partial\theta^\intercal}/T-c^\intercal\Upsilon_D^{-1}\sqrt{T}\frac{\partial\ell_{T,D}(\theta,\bar{y})}{\partial\theta^\intercal}/T\right|\\
&\quad+\sqrt{sn}\left|c^\intercal\Upsilon_D^{-1}\sqrt{T}\frac{\partial\ell_{T,D}(\theta,\bar{y})}{\partial\theta^\intercal}/T-c^\intercal\Upsilon_D^{-1}\sqrt{T}\frac{\partial\ell_{T,D}(\theta,\mu)}{\partial\theta^\intercal}/T\right| \qquad (8.34)
\end{aligned}
$$

We first show that the first term of (8.34) is $o_p(1)$.

$$
\begin{aligned}
&\sqrt{sn}\left|c^\intercal(\hat{\Upsilon}_{T,D}^{-1}-\Upsilon_D^{-1})\sqrt{T}\frac{\partial\ell_{T,D}(\theta,\bar{y})}{\partial\theta^\intercal}/T\right|\\
&=\sqrt{sn}\left|c^\intercal(\hat{\Upsilon}_{T,D}^{-1}-\Upsilon_D^{-1})\sqrt{T}\frac{1}{2}E^\intercal D_n^\intercal\Psi(\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}(\hat{\Sigma}_T-\Sigma)\right|\\
&=O(\sqrt{sn})\|\hat{\Upsilon}_{T,D}^{-1}-\Upsilon_D^{-1}\|_{\ell_2}\sqrt{T}\|E^\intercal\|_{\ell_2}\|\hat{\Sigma}_T-\Sigma\|_F=O(\sqrt{sn})\varpi^2s\sqrt{1/(nT)}\sqrt{T}\sqrt{sn}\sqrt{n}\|\hat{\Sigma}_T-\Sigma\|_{\ell_2}\\
&=O(\sqrt{sn})\varpi^2s\sqrt{1/(nT)}\sqrt{T}\sqrt{sn}\sqrt{n}\sqrt{n/T}=O_p\left(\sqrt{\frac{n^3s^4\varpi^4}{T}}\right)=o_p(1),
\end{aligned}
$$

where the last equality is due to Assumption 3.3(iii).

We now show that the second term of (8.34) is $o_p(1)$.

$$
\begin{aligned}
&\sqrt{sn}\left|c^\intercal\Upsilon_D^{-1}\sqrt{T}\left(\frac{\partial\ell_{T,D}(\theta,\bar{y})}{\partial\theta^\intercal}/T-\frac{\partial\ell_{T,D}(\theta,\mu)}{\partial\theta^\intercal}/T\right)\right|\\
&=\sqrt{sn}\left|c^\intercal\Upsilon_D^{-1}\sqrt{T}\frac{1}{2}E^\intercal D_n^\intercal\Psi(\Theta^{-1}\otimes\Theta^{-1})(D^{-1/2}\otimes D^{-1/2})\operatorname{vec}(\hat{\Sigma}_T-\tilde{\Sigma}_T)\right|\\
&=O(\sqrt{sn})\|\Upsilon_D^{-1}\|_{\ell_2}\sqrt{T}\|E\|_{\ell_2}\|\hat{\Sigma}_T-\tilde{\Sigma}_T\|_F=O_p(\sqrt{sn})\frac{\varpi}{n}\sqrt{T}\sqrt{sn}n\frac{\log n}{T}=O_p\left(\sqrt{\frac{\log^4 n\cdot n^2\varpi^2}{T}}\right)=o_p(1),
\end{aligned}
$$

where the third last equality is due to (8.23), and the last equality is due to Assumption 3.3(iii).

8.6.5 $t_{os,D,2}=o_p(1)$

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 3.3(iii). □

8.7 Proof of Theorem 3.4 and Corollary 3.3

In this subsection, we give proofs of Theorem 3.4 and Corollary 3.3.

Proof of Theorem 3.4. 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 3.2 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

$$
\sqrt{T}g_{T,D}(\theta)\xrightarrow{d}N(0,S),\qquad S:=D_n^+H(D^{-1/2}\otimes D^{-1/2})V(D^{-1/2}\otimes D^{-1/2})HD_n^{+\intercal}, \qquad (8.35)
$$

where S is positive definite given the assumptions of this theorem. 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)

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

where the second equality is due to (8.36). Using (8.35), we have

$$\sqrt{T} g_{T,D}(\hat{\theta}_{T,D}) \xrightarrow{d} N\left(0, \left[I_{n(n+1)/2} - E(E^\intercal W E)^{-1} E^\intercal W\right] S \left[I_{n(n+1)/2} - E(E^\intercal W E)^{-1} E^\intercal W\right]^\intercal\right).$$

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.7) and Lemma 8.2, which hold under the assumptions of this theorem. The asymptotic covariance matrix in the preceding display is idempotent and has rank $n(n+1)/2 - s$. Thus, under H_0,

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

□

To prove Corollary 3.3, we give the following two auxiliary lemmas.

Lemma 8.8 (van der Vaart (1998) p27).

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

as $k \to \infty$.

Lemma 8.9 (van der Vaart (2010) p41). *For $T, n \in \mathbb{N}$ let $X_{T,n}$ be random vectors such that $X_{T,n} \xrightarrow{d} X_n$ as $T \to \infty$ for every fixed n such that $X_n \xrightarrow{d} X$ as $n \to \infty$. Then there exists a sequence $n_T \to \infty$ such that $X_{T,n_T} \xrightarrow{d} X$ as $T \to \infty$.*

Now we are ready to give a proof for Corollary 3.3.

Proof of Corollary 3.3. We only give a proof for part (i), as that for part (ii) is similar. From (3.7) 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 8.8

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

as $n \to \infty$ under H_0. Next invoke Lemma 8.9, there exists a sequence $n = n_T$ such that

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

as $T \to \infty$. □

8.8 Miscellaneous Results

This subsection contains miscellaneous results of the article.

Proof of Corollary 3.1. Note that Theorem 3.2 and a result we proved before, namely,

$$|c^\intercal \hat{J}_{T,D} c - c^\intercal J_D c| = o_p\left(\frac{1}{sn\kappa(W)}\right), \tag{8.39}$$

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.11) 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 $j = 1, \ldots, k$

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

where the equality is due to $\|A\|_{\ell_2} = O(\sqrt{sn\kappa(W)})$. 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

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

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

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

□

Lemma 8.10. *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 8.3. *In the special case of normality, $V = 2D_n D_n^+(\Sigma \otimes \Sigma)$ (Magnus and Neudecker (1986) Lemma 9). Then $c^\intercal J_D c$ could be simplified into*

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

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

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

References

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

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

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.

Minka, T. P. (2000): "Old and New Matrix Algebra Useful for Statistics," *Working Paper*.

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.