Rate-optimal estimation of a high-dimensional semiparametric time series model

Fang Han[*], Sheng Xu[†], and Han Liu[‡]

Abstract

An emerging literature in high-dimensional statistics focuses on studying high-dimensional time series. However, there is little fundamental study on optimal estimation. This paper founds, for the first time in the literature, such a result. For this, we focus on the copula-based time series model. It includes the Gaussian vector autoregressive model as a sub-class. The temporal dependence of this model is fully characterized by the transition matrix $\mathbf{A}$, which we aim to estimate. Under a low-rank assumption on $\mathbf{A}$, we derive sharp upper and lower bounds for estimation. A key step to establishing the lower bound is through a novel analysis of the log determinant term in calculating the Kullback-Leibler divergence. For this, we observe a clear distinction from the analysis of the independent data, where the log determinant term is typically ignorable.

Keywords: high-dimensional time series; transition matrix estimation; α-mixing; constrained ℓ_*-minimization; minimax lower bound.

1 Introduction

The multivariate time series analysis plays a fundamental role in modelling and analyzing many types of datasets of temporal correlatedness. For example, it is critically useful for the analysis of stock market data (Fan et al., 2011b), time course genomic data (Michailidis, 2012), and task-based/resting state functional magnetic resonance image (fMRI) data (Lindquist, 2008; Smith, 2012). A common feature through these datasets is that the time series dimension d is un-ignorable compared to the time series length T. This characteristic motivates regularized estimation (Bickel et al., 2009; Negahban and Wainwright, 2011).

There exists an emerging literature in studying high-dimensional time series. Focused on the vector autoregressive (VAR) model (Lütkepohl, 2007), Shojaie and Michailidis (2010), Negahban and Wainwright (2011), and Han et al. (2015) studied estimating the large transition matrix $\mathbf{A}$ of an order one VAR model (VAR(1)). Song and Bickel (2011) and Davis et al. (2016) studied VAR models of order $p > 1$ (VAR(p)). Very recently, Basu and Michailidis (2015) provided an

[*] Department of Statistics, University of Washington, Seattle, WA 98195, USA; e-mail: fanghan@uw.edu

[†] Department of Statistics, Yale University, New Haven, CT 06511, USA; e-mail: sheng.xu@yale.edu

[‡] Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: hanliu@princeton.edu

analysis of stationary Gaussian VAR(p) models and revealed the role of spectral density functions in measuring the parameter estimation accuracy. Related statistical problems of interest include stochastic regression (Loh and Wainwright, 2012; Chang et al., 2015) and covariance/precision matrix estimation (Fan et al., 2011a; Xiao and Wu, 2012; Chen et al., 2013).

Although there have been a variety of methods for studying high-dimensional VAR models, it is still unclear whether they are minimax optimal. This is because the minimax lower bound, serving as an important benchmark for evaluating the estimators' performance, is still unknown in the time series literature. In addition, we mention two constraints in the literature: (i) most aforementioned methods require the data to be Gaussian distributed; (ii) most results focus on the linear temporal system, while the analysis for nonlinear systems is largely unexplored.

This paper focuses on studying a general semiparametric time series model as an order one Markov chain[1]. Specifically, let $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ be a multivariate time series with $\boldsymbol{X}_t \in \mathbb{R}^d$. We focus on the following time series model.

- **Semiparametric meta-elliptical-based stationary time series.** Assume there exists a set of unknown strictly increasing functions, $\boldsymbol{f} := \{f_1, \ldots, f_d\}$, such that
$$\boldsymbol{f}(\boldsymbol{X}_t) = \mathbf{A}\boldsymbol{f}(\boldsymbol{X}_{t-1}) + \boldsymbol{E}_t \text{ and } \boldsymbol{Z}_t := \boldsymbol{f}(\boldsymbol{X}_t) \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \xi), \quad \text{for any } t \in \mathbb{Z}. \tag{1.1}$$
Here $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ is of elements identically and continuously elliptically distributed[2].

Model (1.1) is related to a growing literature in modelling the possibly nonlinear temporal dependence (Hsing and Wu, 2004; Beare, 2010; Patton, 2012a,b; Wang and Xia, 2015). In particular, Model (1.1) is a multivariate extension to the univariate order one Gaussian copula Markov chain introduced in Chen and Fan (2006) and Chen et al. (2009). It is also obvious that Model (1.1) is a strict extension to the stationary Gaussian VAR(1) model considered in Negahban and Wainwright (2011), Han et al. (2015), and Basu and Michailidis (2015).

This paper aims to estimate the transition matrix $\mathbf{A}$ in Model (1.1) under a low-rank assumption. Here the assumption, $r := \text{rank}(\mathbf{A}) < d$, is regular in the related literature (Christiano et al., 1999; Bai, 2003; Pan and Yao, 2008; Bańbura et al., 2010; Negahban and Wainwright, 2011; Lam et al., 2011; Lam and Yao, 2012; Basu, 2014). In particular, it is strongly motivated from a latent factor model, where a few latent factors drive the main movement of the multivariate time series, and the marginal transformation $\boldsymbol{f}$ could be pictured as contamination (Chen et al., 2009).

We establish the minimax optimal rate of convergence for transition matrix estimation within Model (1.1). Here, on one hand, for parameter estimation and, in particular, for handling the possibly nonlinear temporal system, we introduce a novel algorithm based on the *robust sign transformation*, whose intrinsic idea comes from the one-to-one map between the Pearson's correlation and Kendall's tau under the Gaussian copula model (Kruskal, 1958).

On the other hand, regarding the optimality, we build a rate-sharp minimax lower bound. Conventional techniques, designed for the independent data, are not well suited for studying the time series. We overcome this issue via a novel analysis focused on the log determinant term in calculating the Kullback-Leibler divergence (Tsybakov, 2009). The log determinant term reflects the

[1] We will discuss extensions to the order $p > 1$ case in Section 5.

[2] Mathematical definition of elliptical distribution will be provided in the later sections.

impact of data dependence on estimation, and is regularly ignorable under the data independence assumption.

1.1 Other related works

Our work on optimal estimation of large transition matrix is closely related to estimating large covariance matrices in the time series. For this, we refer the readers to: (i) Fan et al. (2011a), Bai and Liao (2016), and Han and Liu (2013) for results under different mixing conditions; (ii) Xiao and Wu (2012) and Chen et al. (2013) for results under the physical dependence condition (Wu, 2005); and (iii) Sancetta (2008) and Fan et al. (2012) for results under the weak dependence condition (Doukhan and Louhichi, 1999).

Our work is also related to a vast literature on low-rank matrix estimation under the univariate/multivariate regression (Izenman, 1975; Reinsel and Velu, 1998; Yuan et al., 2007; Bunea et al., 2011, 2012) and matrix completion (Candès and Recht, 2009; Candès et al., 2011; Koltchinskii et al., 2011) settings. In particular, under the multivariate regression setting, Yuan et al. (2007) and Bunea et al. (2011) advocated using singular value ℓ_1 (nuclear-) and ℓ_0 (rank-) norms to penalize the loss function. Ever since the analysis of high-dimensional time series is distinct from the analysis of the independent data, our results are complementary to those in Yuan et al. (2007) and Bunea et al. (2011).

1.2 Notation

Throughout the paper, let $\mathbb{Z}$ and $\mathbb{R}$ represent the sets of integers and real numbers. Let $\boldsymbol{v} = (v_1, \ldots, v_m)^{\mathsf{T}}$ and $\mathbf{M} = [\mathbf{M}_{jk}] \in \mathbb{R}^{n\times m}$ be an n-dimensional real vector and an n by m real matrix. For sets $I \subset \{1, \ldots, n\}$ and $J \subset \{1, \ldots, m\}$, let $\boldsymbol{v}_J$ be the subvector of $\boldsymbol{v}$ with entries indexed by J, and $\mathbf{M}_{I,J}$ be the submatrix of $\mathbf{M}$ with entries indexed by I and J. For any two matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{n\times m}$, we define $\langle \mathbf{M}, \mathbf{N}\rangle := \mathrm{Tr}(\mathbf{M}^{\mathsf{T}}\mathbf{N})$, where $\mathrm{Tr}(\cdot)$ represents the trace for square matrices. For $0 < q < \infty$, we define the vector ℓ_0, ℓ_q, and ℓ_∞ (pseudo-)norms of $\boldsymbol{v}$ to be $\|\boldsymbol{v}\|_0 := \mathrm{card}(\{j : v_j \neq 0\})$, $\|\boldsymbol{v}\|_q := (\sum_{i=1}^m |v_i|^q)^{1/q}$, and $\|\boldsymbol{v}\|_\infty := \max_{1\le i\le m} |v_i|$. We define the matrix element-wise supremum ($\ell_{\max}$), operator (ℓ_q), Frobenius (ℓ_{F}), and nuclear (ℓ_*) norms as $\|\mathbf{M}\|_{\max} := \max_{j,k} |\mathbf{M}_{jk}|$, $\|\mathbf{M}\|_q := \max_{\boldsymbol{v}} \|\mathbf{M}\boldsymbol{v}\|_q/\|\boldsymbol{v}\|_q$, $\|\mathbf{M}\|_{\mathsf{F}} := (\sum \mathbf{M}_{jk}^2)^{1/2}$, and $\|\mathbf{M}\|_* := \sum \sigma_j(\mathbf{M})$. Here $\sigma_j(\mathbf{M})$ represents the j-th largest singular value of $\mathbf{M}$. For any univariate function $f(\cdot) : \mathbb{R} \to \mathbb{R}$, define $f(\mathbf{M}) := [f(\mathbf{M}_{jk})] \in \mathbb{R}^{n\times m}$. For the symmetric real matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, let $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ be the largest and smallest eigenvalues of $\mathbf{M}$. For a set of functions $\boldsymbol{f} := \{f_j\}_{j=1}^m$, define $\boldsymbol{f}(\boldsymbol{v}) := (f_1(v_1), \ldots, f_m(v_m))^{\mathsf{T}}$. Let $\mathrm{diag}(\mathbf{M})$ be the diagonal matrix with the diagonals $\mathbf{M}_{11}, \mathbf{M}_{22}, \ldots, \mathbf{M}_{nn}$. Let $\mathrm{vec}(\mathbf{M})$ be the vectorized version of $\mathbf{M}$. Let $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ be the identity matrix. For any $x \in \mathbb{R}$, we define the sign function $\mathrm{sign}(x) := x/|x|$, where by convention we let $0/0 = 0$, and the floor function $\lfloor x \rfloor$ as the largest integer not greater than x. We let c, C be two generic absolute positive constants, whose actual values might vary at different locations. For any two real sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ if there exist c, C such that $c|b_n| \le |a_n| \le C|b_n|$ for any large enough n.

1.3 Paper organization

We organize the rest of the paper as follows. Section 2 introduces the method. Section 3 studies the theoretical properties of the proposed method. In Section 4, we establish a minimax lower bound of the studied problem. Section 5 gives extensions to Markov chains of order $p > 1$. Section 6 provides finite sample simulations and the real stock market data analysis. Section 7 concludes. Proofs and axillary lemmas are relegated to a supplementary material.

2 Model description and estimation procedure

This section elaborates Model (1.1) in more detail, and introduces the proposed method. In the sequel, suppose we only observe a length T fragment, $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T\}$, of the times series $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$. In addition, assume the time series is centered with median$(\boldsymbol{X}_t) = \mathbf{0}$ for $t \in \mathbb{Z}$. In the sequel, for any two random vectors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^d$, we write $\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{Y}$ if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are identically distributed.

2.1 Model description

For elaborating Model (1.1) in more detail, we first define the elliptical distribution.

Definition 2.1 (Elliptical distribution, Fang et al. (1990))**.** A continuous d-dimensional random vector $\boldsymbol{Z}$ of covariance matrix $\boldsymbol{\Sigma}$ is said to follow an elliptical distribution if and only if there exist a vector $\boldsymbol{\mu} \in \mathbb{R}^d$, a nonnegative random variable $\xi \in \mathbb{R}$ of $\mathbb{P}(\xi = 0) = 0$, a random vector $\boldsymbol{U} \in \mathbb{R}^d$ uniformly distributed in the unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ and independent of ξ, such that

$$\boldsymbol{Z} \overset{\mathsf{d}}{=} \boldsymbol{\mu} + \xi \boldsymbol{\Sigma}^{1/2} \boldsymbol{U}.$$

We then write $\boldsymbol{Z} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$. Of note, the elliptical distribution family contains the Gaussian and multivariate t-distribution families.

Suppose $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$ is the observed time series. It follows Model (1.1) if and only if the following two properties hold.

(P1). There exists a set of unknown strictly increasing functions, $\boldsymbol{f} := \{f_1, \ldots, f_d\}$, such that

$$\boldsymbol{f}(\boldsymbol{X}_t) = \mathbf{A} \boldsymbol{f}(\boldsymbol{X}_{t-1}) + \boldsymbol{E}_t \text{ and } \boldsymbol{Z}_t := \boldsymbol{f}(\boldsymbol{X}_t) \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \xi), \quad \text{for any } t \in \mathbb{Z}.$$

For model identifiability, we assume $\boldsymbol{\Sigma}$ has diagonals all equal to 1.

(P2). The random vector $(\boldsymbol{Z}_1^{\mathsf{T}}, \ldots, \boldsymbol{Z}_T^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{Td}$ is continuous and elliptically distributed for arbitrary integer $T \geq 1$.

In practice, such a time series could be generated through an iterative algorithm presented in Rémillard et al. (2012). Specifically, for each $t \in \mathbb{Z}$, given $\boldsymbol{Z}_{t-1}$, the idiosyncratic error $\boldsymbol{E}_t$ is constructed such that $(\boldsymbol{Z}_{t-1}^{\mathsf{T}}, \boldsymbol{Z}_t^{\mathsf{T}})^{\mathsf{T}}$ is elliptically distributed and $\{\boldsymbol{Z}_{t-1}, \boldsymbol{Z}_t\}$ satisfies (P1)[3]. We then generate $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$ based on $\{\boldsymbol{Z}_t\}_{t \in \mathbb{Z}}$ and the set of strictly increasing functions $\boldsymbol{f} = \{f_j\}_{f=1}^d$.

Due to the uniqueness of $\boldsymbol{E}_t$, we immediately have the following proposition. It guarantees that Property (P2) holds.

[3] Of note, $\boldsymbol{E}_t$ is possibly correlated with $\boldsymbol{Z}_{t-1}$.

Proposition 2.1. Suppose we generate $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ using Algorithm 1 in Rémillard et al. (2012). Then we have $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ is globally elliptically distributed, i.e., $(\boldsymbol{Z}_{t_1}^{\mathsf{T}}, \ldots, \boldsymbol{Z}_{t_m}^{\mathsf{T}})^{\mathsf{T}}$ is elliptically distributed for arbitrary positive integer m and $t_1, \ldots, t_m \in \mathbb{Z}$.

Model (1.1) is very general. Because the Gaussian belongs to the elliptical distribution family, Model (1.1) is a strict extension to the Gaussian-copula-based time series model, which has attracted a lot of attention in Chen and Fan (2006). In addition, when the transition matrix $\mathbf{A} = \mathbf{0}$, we recover the meta-elliptical distribution family introduced in Fang et al. (2002). Our model is also very related to Markovian models equipped with meta-elliptical copulas (Rémillard et al., 2012). Model (1.1) could also be regarded as a possibly nonlinear alternative to the Kalman filter, for which we refer the readers to Roweis and Ghahramani (1999) for a complementary review.

2.2 Estimation procedure

The problem of transition matrix estimation is strongly related to multiple regression. For this, even when the temporal system is linear, under the low-rank assumption on $\mathbf{A}$, the least square estimator,

$$\widehat{\mathbf{A}}^{\mathrm{LSE}} := \operatorname*{argmin}_{\mathbf{Q}\in\mathbb{R}^{d\times d}} \frac{1}{T-1}\sum_{t=2}^{T} \|\boldsymbol{X}_t - \mathbf{Q}\boldsymbol{X}_{t-1}\|_{\mathsf{F}}^2,$$

is not statistically efficient for estimating $\mathbf{A}$ (Izenman, 1975). For improving estimation efficiency, there have been a number of methods introduced in the literature. In particular, Negahban and Wainwright (2011) proposed the following penalized M-estimator:

$$\widehat{\mathbf{A}}_{\lambda}^{\mathrm{NW}} := \operatorname*{argmin}_{\mathbf{Q}\in\mathbb{R}^{d\times d}} \frac{1}{T-1}\sum_{t=2}^{T} \|\boldsymbol{X}_t - \mathbf{Q}\boldsymbol{X}_{t-1}\|_{\mathsf{F}}^2 + \lambda\|\mathbf{Q}\|_{*}. \tag{2.1}$$

Here the nuclear norm is added to induce the sparsity of the estimator's singular values, and hence encourages low-rankness. The obtained estimator is easy to implement, and proves to enjoy good theoretical properties.

However, given (2.1), because transformation functions $f_1, \ldots, f_d$ are unknown, estimating $\mathbf{A}$ in Model (1.1) requires extra efforts. More specifically, let's consider the following general setting. Suppose $\mathbf{M} \in \mathbb{R}^{n\times n}$ and $\mathbf{M}_1 \in \mathbb{R}^{n\times m}$, of dimensions n and $\{n, m\}$, are two real matrices, $\mathbf{M}$ is symmetric and positive definite, and the matrix of interest $\mathbf{A}$ could be written as $\mathbf{A} = \mathbf{M}_1^{\mathsf{T}}\mathbf{M}^{-1} \in \mathbb{R}^{m\times n}$ of rank $r \le \min(m, n)$. Let $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}}_1$ be estimates of $\mathbf{M}$ and $\mathbf{M}_1$, and define

$$\widehat{\mathbf{A}}_{\lambda}^{\mathrm{G}} := \operatorname*{argmin}_{\mathbf{Q}\in\mathbb{R}^{m\times n}} L_{\lambda}(\mathbf{Q}; \widehat{\mathbf{M}}, \widehat{\mathbf{M}}_1), \quad \text{where } L_{\lambda}(\mathbf{Q}; \widehat{\mathbf{M}}, \widehat{\mathbf{M}}_1) := \langle -2\widehat{\mathbf{M}}_1 + \widehat{\mathbf{M}}\mathbf{Q}^{\mathsf{T}}, \mathbf{Q}^{\mathsf{T}}\rangle + \lambda\|\mathbf{Q}\|_{*}. \tag{2.2}$$

The following lemma provides the key result motivating our new method. Its proof is very straightforward given the literature.

Lemma 2.2. Assume $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}}_1$ are the estimates of $\mathbf{M} \in \mathbb{R}^{n\times n}$ and $\mathbf{M}_1 \in \mathbb{R}^{n\times m}$ based on T observations, satisfying

$$\mathbb{P}(\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 \le \delta_1) \ge 1 - \epsilon_1 \quad \text{and} \quad \mathbb{P}(\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2 \le \delta_2) \ge 1 - \epsilon_2. \tag{2.3}$$

Here $\delta_1, \delta_2, \epsilon_1, \epsilon_2$ are functions of (T, m, n), and ϵ_1, ϵ_2 go to zero as (T, m, n) increases to infinity. Further assume

$$\mathbf{M} = \mathbf{M}^{\mathsf{T}},\ \lambda_{\min}(\mathbf{M}) \geq \gamma_{\min},\ \mathbf{A} := \mathbf{M}_1^{\mathsf{T}}\mathbf{M}^{-1},\ \operatorname{rank}(\mathbf{A}) \leq r,\ \|\mathbf{A}\|_2 \leq \gamma_{\max},$$

and

$$\lambda \geq 2(\gamma_{\max}\delta_1 + \delta_2) \ \text{ and } \ \mu \leq \gamma_{\min} - \delta_1,$$

where $\gamma_{\min}$ and $\gamma_{\max}$ are two absolute positive constants. We then have

$$\mathbb{P}\Big(\|\widehat{\mathbf{A}}_{\lambda}^{\mathrm{G}} - \mathbf{A}\|_{\mathsf{F}} \geq \frac{\lambda + 2\sqrt{2}(\gamma_{\max}\delta_1 + \delta_2)}{2\mu}\sqrt{r}\Big) \leq \epsilon_1 + \epsilon_2.$$

Lemma 2.2 implies, as long as $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}}_1$ are consistent estimators of $\mathbf{M}$ and $\mathbf{M}_1$, and the tuning parameter λ is appropriately chosen, $\widehat{\mathbf{A}}_{\lambda}^{G}$ could consistently estimate $\mathbf{A}$. For Model (1.1), if we write

$$\mathbf{\Sigma} = \operatorname{Cov}(\boldsymbol{Z}_t) \ \text{ and } \ \mathbf{\Sigma}_1 = \operatorname{Cov}(\boldsymbol{Z}_t, \boldsymbol{Z}_{t+1})$$

to be the covariance and lag 1 covariance matrices of $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$, by the celebrated Yule-Walker equation, we have $\mathbf{A} = \mathbf{\Sigma}_1^{\mathsf{T}}\mathbf{\Sigma}^{-1}$. Hence, Lemma 2.2 indicates, for estimating the transition matrix $\mathbf{A}$, a major step is to construct efficient estimators of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_1$. Of note, by setting $\widehat{\mathbf{M}} = \sum_{t=1}^{T-1} \boldsymbol{X}_t\boldsymbol{X}_t^{\mathsf{T}}/(T-1)$ and $\widehat{\mathbf{M}}_1 = \sum_{t=1}^{T-1} \boldsymbol{X}_t\boldsymbol{X}_{t+1}^{\mathsf{T}}/(T-1)$, we recover the estimator $\widehat{\mathbf{A}}_{\lambda}^{\mathrm{NW}}$.

However, in Model (1.1), $\{\boldsymbol{Z}_t\}_{1\leq t\leq T}$ is not observable, and hence $\widehat{\mathbf{A}}_{\lambda}^{\mathrm{NW}}$ could be biased. For consistently estimating $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_1$, the key observation is the following lemma, revealing that the latent covariances could be directly estimated via a sign transformation of the original data.

Lemma 2.3 (Kruskal (1958))**.** Suppose $\boldsymbol{Z} \sim EC_d(\mathbf{0}, \mathbf{\Sigma}, \xi)$ is elliptically distributed with the diagonals of $\mathbf{\Sigma}$ all equal to 1. Let $\boldsymbol{S} := \operatorname{sign}(\boldsymbol{Z})$ and $\mathbf{T} := \operatorname{Cov}(\boldsymbol{S})$. We then have $\mathbf{\Sigma} = \sin(\frac{\pi}{2}\mathbf{T})$.

Motivated from Lemmas 2.2 and 2.3, our transition matrix estimation procedure has two steps. The first step is a robust estimation of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_1$ through a new algorithm we call "robust sign transformation". In the second step, we plug the estimators $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Sigma}}_1$ into the general penalization algorithm (2.2). The detailed procedure is as follows.

- **Step 1: Estimating $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_1$.** For $t = 1, \ldots, T$, we define the robust-sign-transformation version of the observed data points $\boldsymbol{X}_t$ as

$$\boldsymbol{S}_t := \operatorname{sign}(\boldsymbol{X}_t) = (\operatorname{sign}(X_{t1}), \operatorname{sign}(X_{t2}), \ldots, \operatorname{sign}(X_{td}))^{\mathsf{T}}. \tag{2.4}$$

  We further calculate

$$\widehat{\mathbf{T}} := \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{S}_t\boldsymbol{S}_t^{\mathsf{T}} \ \text{ and } \ \widehat{\mathbf{T}}_1 := \frac{1}{T-1}\sum_{t=1}^{T-1} \boldsymbol{S}_t\boldsymbol{S}_{t+1}^{\mathsf{T}}.$$

  $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_1$ are then separately estimated using

$$\widehat{\mathbf{\Sigma}} := \sin(\frac{\pi}{2}\widehat{\mathbf{T}}) \ \text{ and } \ \widehat{\mathbf{\Sigma}}_1 := \sin(\frac{\pi}{2}\widehat{\mathbf{T}}_1).$$

- **Step 2: Estimating the transition matrix A.** We plug $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Sigma}}_1$ into the general algorithm (2.2) to obtain the final estimator:

$$\widehat{\mathbf{A}}_\lambda := \underset{\mathbf{Q}\in\mathbb{R}^{d\times d}}{\operatorname{argmin}}\, L_\lambda(\mathbf{Q}; \widehat{\mathbf{\Sigma}}, \widehat{\mathbf{\Sigma}}_1). \tag{2.5}$$

Of note, $\widehat{\mathbf{A}}_\lambda$ directly estimates $\mathbf{A}$ without requiring calculating the transformation functions $f_1, \ldots, f_d$. This is due to Lemma 2.3 and the fact that the sign function is invariant to strictly increasing functions.

Remark 2.4. We compare the robust-sign-transformation approach to the others. On one hand, Chen and Fan (2006) studied the setting where the exact copula function is known, and proposed a likelihood-based approach. However, Model (1.1) does not specify the exact copula function, and hence is more general. On the other hand, Rémillard et al. (2012) studied models like Model (1.1) and proposed a fully nonparametric method. However, a key step there is to estimate the joint distribution function, which is known to be very inefficient in high dimensions.

Remark 2.5. For presentation simplicity, we focus on the centered time series following similar settings as in Negahban and Wainwright (2011), Han et al. (2015), and Basu and Michailidis (2015). However, of note, if $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ is a non-centered time series following Model (1.1), then we can instead apply the developed procedure to $\{\boldsymbol{X}_t - \boldsymbol{X}_{t-\Delta T}\}_{t\in\mathbb{Z}}$. This new time series is centered and corresponds to $\{\boldsymbol{Z}_t - \boldsymbol{Z}_{t-\Delta T}\}_{t\in\mathbb{Z}}$, whose transition matrix approximates $\mathbf{A}$ as the gap Δ_T increases with T. Such a procedure, we call "split-and-conquer", is common in the literature and can settle the non-centeredness issue without hurting the theory. In practice, we could also conduct a pre-processing procedure by deleting the sample median from the time series.

3 Properties of the proposed estimator

We now study the properties of the robust-sign-transformation estimator $\widehat{\mathbf{A}}_\lambda$ introduced in Section 2. To this end, let's first introduce some extra definitions. Given a measurable space $(\Omega, \mathcal{F}, \mathbb{P})$, for any two σ-algebras $\mathcal{A}$ and $\mathcal{B}$ in $\mathcal{F}$, define the corresponding α-mixing coefficient by

$$\alpha(\mathcal{A}, \mathcal{B}) := \sup_{A\in\mathcal{A}, B\in\mathcal{B}} \big|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)\big|.$$

Let $\{X_t\}_{t\in\mathbb{Z}}$ be a sequence of real-valued random variables. We further define

$$\alpha(n; \{X_t\}_{t\in\mathbb{Z}}) := \sup_{t\in\mathbb{Z}} \alpha(\mathcal{G}_{-\infty}^t, \mathcal{G}_{t+n}^\infty),$$

where for arbitrary $j \in \mathbb{Z}$, $\mathcal{G}_{-\infty}^j := \sigma(X_t, t \le j)$ and $\mathcal{G}_j^\infty := \sigma(X_t, t \ge j)$ represent the sigma fields generated by $\{X_t\}_{t\le j}$ and $\{X_t\}_{t\ge j}$ respectively.

We then study the properties of the estimator $\widehat{\mathbf{A}}_\lambda$ under Model (1.1). For this, let's first provide two assumptions on the sequence $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$.

- **Assumption (A1).** Assume $\lambda_{\min}(\mathbf{\Sigma})$ and $\lambda_{\max}(\mathbf{\Sigma})$ are lower and upper bounded by two absolute positive constants, respectively.

- **Assumption (A2).** The sequence $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ satisfies the following strong mixing condition:

$$\alpha(n; \{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}) = \alpha(n; \{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}) \le \exp(-\kappa_1 n^{\gamma_1}), \quad \text{for all } n \ge 1, \tag{3.1}$$

for some absolute positive constants κ_1 and γ_1.

Of note, Assumption **(A1)** is regular in the low-rank matrix estimation literature (Negahban and Wainwright, 2011; Candes and Plan, 2011). Assumption **(A2)** is routine in studying autoregression models (Andrews, 1991; Den Haan and Levin, 1998; Liebscher, 2005; Bandyopadhyay, 2006). In particular, Beare (2010), Rémillard et al. (2012), and Han and Li (2017) verified that the sequence $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ satisfies **(A2)** with $\gamma_1 = 1$ when certain explicitly stated conditions hold.

Proposition 3.1 (Rémillard et al. (2012))**.** Suppose Model (1.1) and Assumption **(A1)** hold. In addition, suppose $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ is stationarily Gaussian or multivariate t-distributed with the rank of $\mathbf{A}$ bounded. We then have (3.1) holds with $\gamma_1 = 1$.

Proposition 3.2 (Kolmogorov and Rozanov (1960) and Han and Li (2017))**.** Suppose Model (1.1) and Assumption **(A1)** hold. In addition, suppose $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ is stationarily Gaussian-distributed with $\|\mathbf{A}\|_2 < C < 1$ for some absolute constant $C > 0$. We then have (3.1) holds with $\gamma_1 = 1$.

Remark 3.3. We first note Proposition 3.1 could be further strengthened. In particular, by checking the proofs of Propositions 2 and 4 in Rémillard et al. (2012), we immediately have, when $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ is stationarily Gaussian or multivariate t-distributed, the time series $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ is ρ-mixing. Longla and Peligrad (2012) showed $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ is also ϕ-mixing. We refer the readers to Bradley (2005) for explicit definitions of these mixing conditions. Secondly, we note Proposition 3.2 is established in Han and Li (2017) and credited to Kolmogorov and Rozanov (1960), which showed that the ρ-mixing coefficient of a Gaussian sequence is determined by the corresponding canonical correlation and hence can be explicitly calculated.

The next theorem studies the approximation error of the proposed estimator $\widehat{\mathbf{A}}_\lambda$ in (2.5) over the class of low-rank matrices $\mathcal{A}_M(r, \gamma_{\max})$, defined as follows:

$$\mathcal{A}_M(r, \gamma_{\max}) := \Big\{\mathbf{M} \in \mathbb{R}^{d\times d} : \operatorname{rank}(\mathbf{M}) \le r, \|\mathbf{M}\|_2 \le \gamma_{\max}\Big\}. \tag{3.2}$$

Here $\gamma_{\max}$ is an absolute positive constant not necessarily smaller than 1.

Theorem 3.4. Assume Model (1.1) and Assumptions **(A1), (A2)** hold, and assume the transition matrix $\mathbf{A} \in \mathcal{A}_M(r, \gamma_{\max})$ for some absolute positive constant $\gamma_{\max}$. Suppose there exist three absolute constants $\mu, C_1, C_2 > 0$ such that

$$\lambda \ge 2\big(C_1\gamma_{\max}\sqrt{d/T} + C_2\sqrt{d/(T-1)}\big), \tag{3.3}$$

and

$$\mu \le \lambda_{\min}(\mathbf{\Sigma}) - C_1\sqrt{d/T}. \tag{3.4}$$

Further assume there exists an absolute constant $\kappa_2 > 0$ such that $T \ge \kappa_2 d^{2/\gamma_0 - 1}$, where $\gamma_0 = \gamma_1/(\gamma_1+1)$ and $T \ge 4$. We then have

$$\mathbb{P}\Big(\|\widehat{\mathbf{A}}_\lambda - \mathbf{A}\|_{\mathsf{F}} \ge \frac{1}{2\mu}\big(\lambda\sqrt{r} + 2\sqrt{2}C_1\gamma_{\max}\sqrt{\frac{dr}{T}} + 2\sqrt{2}C_2\sqrt{\frac{dr}{T-1}}\big)\Big) = O\Big(\exp(-C_3 d) + d^2\exp\Big(-2\sqrt{\frac{T}{d}}\Big)\Big),$$

where $C_3 > 0$ is a constant only depending on $C_1, C_2, \kappa_1, \kappa_2, \gamma_1, \lambda_{\max}(\mathbf{\Sigma})$, and $\lambda_{\min}(\mathbf{\Sigma})$.

Theorem 3.4 shows, when we choose $\lambda \asymp \sqrt{d/T}$ and the theorem's conditions hold, we have

$$\|\widehat{\mathbf{A}}_\lambda - \mathbf{A}\|_{\mathsf{F}} = O_P(\sqrt{dr/T}),$$

which is the minimax optimal rate as in Section 4 we will show. Of note, this result is a strict extension to Corollary 4 in Negahban and Wainwright (2011), which heavily relies on the Gaussian assumption, a linear temporal system, and an additional spectral norm constraint on $\mathbf{A}$: $\|\mathbf{A}\|_2 < 1$. This result is also closely related to Proposition V.2 in Basu (2014), which recovered the same rate for the linear Gaussian family.

Remark 3.5. Theorem 3.4 also reveals an interesting phenomenon on scaling. Specifically, when we allow the dimension d to increase with the time series length T, the temporal dependence strength will have an impact on the scaling requirement, i.e., T is required to be larger than $\kappa_2 d^{2/\gamma_0 - 1}$ for some absolute constant κ_2. When the data are independent, so that $\mathbf{A} = 0$ and $\gamma_1 = \infty$, we attain the efficient scaling $T \geq \kappa_2 d$. However, when $\gamma_1 < \infty$, T is required to be of an order larger than d. In comparison, there is no rate lost in approximation if γ_1 is assumed to be finite, which holds under the conditions in either Proposition 3.1 or Proposition 3.2. Similar phenomena have also been discovered in the literature (e.g., Lemma 3.1 in Fan et al. (2011a) and Theorem 3 in Fan et al. (2012)).

A key step in proving Theorem 3.4 is to establish the convergence rates of $\widehat{\mathbf{T}}$ and $\widehat{\mathbf{T}}_1$ under the spectral norm. The sign subgaussian property of the elliptical distribution proved in Han and Liu (2017) and Barber and Kolar (2015) is used. In particular, for any $\boldsymbol{Z} \sim EC_d(\mathbf{0}, \boldsymbol{\Sigma}, \xi)$, $\xi \in \mathbb{R}$, and $\boldsymbol{v} \in \mathbb{S}^{d-1}$, we have

$$\mathbb{E}\exp(\xi \cdot \boldsymbol{v}^T \mathrm{sign}(\boldsymbol{Z})) \leq \exp\Big(\frac{\lambda_{\max}(\boldsymbol{\Sigma})}{2\lambda_{\min}(\boldsymbol{\Sigma})} \cdot \xi^2\Big). \tag{3.5}$$

Indeed, from the proof of Theorem 3.4, under the mixing condition **(A2)** and using the sign subgaussian property (3.5), we have

$$\|\widehat{\mathbf{T}} - \mathbb{E}\widehat{\mathbf{T}}\|_2 = O_P(\sqrt{d/T}) \quad \text{and} \quad \|\widehat{\mathbf{T}}_1 - \mathbb{E}\widehat{\mathbf{T}}_1\|_2 = O_P(\sqrt{d/T}),$$

which further implies

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 = O_P(\sqrt{d/T}) \quad \text{and} \quad \|\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1\|_2 = O_P(\sqrt{d/T}).$$

The proof of above equations hinges on combining the results in Han and Liu (2017) and Barber and Kolar (2015) with recent developments in concentration inequalities for weakly dependent data (Merlevède et al., 2011).

Remark 3.6. Of note, It is also of interest to compare the rate of convergence derived in Theorem 3.4 (under a low-rank assumption on $\mathbf{A}$) to the one obtained in Basu and Michailidis (2015) (under a sparsity assumption on $\mathbf{A}$). In particular, Proposition 4.1 in Basu and Michailidis (2015) shows there exists an estimator $\widehat{\mathbf{A}}^s$ such that

$$\|\widehat{\mathbf{A}}^s - \mathbf{A}\|_{\mathsf{F}} = O_P(\sqrt{s \log d/n}), \tag{3.6}$$

where s represents the total number of nonzero entries in $\mathbf{A}$. Assume

$$\mathbf{A}_{jj} \neq 0 \quad \text{for} \quad j = 1, \ldots, d,$$

indicating that each covariate is at least correlated with its own history. Equation (3.6) then yields

$$\|\widehat{\mathbf{A}}^s - \mathbf{A}\|_{\mathsf{F}} = O_P(\sqrt{d\log d/n}).$$

This rate could be slower than our derived rate $O_P(rd/n)$ under the low-rank assumption.

4 Minimax lower bound

Theorem 3.4 shows that the proposed estimator $\widehat{\mathbf{A}}_\lambda$ attains the rate of convergence $\sqrt{dr/T}$. This section shows this rate cannot be further improved in the minimax sense.

To this end, let's define the following sets of interest:

$$\mathcal{V}_M := \Big\{\mathbf{M} \in \mathbb{R}^{d\times d} : \mathrm{diag}(\mathbf{M}) = \mathbf{I}_d, \mathbf{M} = \mathbf{M}^\mathsf{T}, \text{and there exist two generic absolute positive constants } c \text{ and } C \text{ such that } 0 < c \le \lambda_{\min}(\mathbf{M}) \le \lambda_{\max}(\mathbf{M}) \le C < \infty\Big\}$$

$$\text{and}\quad \mathcal{V}_f := \Big\{\boldsymbol{f} = \{f_j\}_{j=1}^d : f_j \text{ is strictly increasing}\Big\}.$$

The next theorem characterizes the minimax lower bound in approximating the transition matrix $\mathbf{A}$. It fills a long standing gap on understanding the optimal estimation of high-dimensional time series, and proves the minimax optimality of $\widehat{\mathbf{A}}_\lambda$.

Theorem 4.1. Assume $rd/T = o(1)$ and Model (1.1) holds. Let $\mathbb{P}_{\mathbf{A},\mathbf{\Sigma},\boldsymbol{f}}$ be the probability measure on marginally Gaussian copula distributed $\{\boldsymbol{X}_t\}_{t=1}^T$ whose probability space is uniquely determined by the parameters $\mathbf{A}$, $\mathbf{\Sigma}$, and $\boldsymbol{f}$. Then for arbitrary absolute constant $\beta \in (0,1)$, there exists a sufficiently small absolute constant $C_\beta > 0$, such that the following inequality,

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}\in\mathcal{A}_M(r,\gamma_{\max}),\mathbf{\Sigma}\in\mathcal{V}_M,\boldsymbol{f}\in\mathcal{V}_f} \mathbb{P}_{\mathbf{A},\mathbf{\Sigma},\boldsymbol{f}}\Big(\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathsf{F}}^2 \ge C_\beta \cdot \frac{rd}{T}\Big) > \beta, \tag{4.1}$$

holds for all sufficiently large T. Here $\gamma_{\max}$ can be any absolute positive constant and the infimum is taken over all measurable estimators on the probability space generated by the multivariate time series $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$.

Theorem 4.1, combined with Theorem 3.4, yields the minimax rate-optimality of $\widehat{\mathbf{A}}_\lambda$. The proof of Theorem 4.1 is involved. We provide a sketch here, while deferring more details to Section A1 in a supplementary material.

Proof. We first introduce some additional notation. Let $\alpha \in (0, 1/8)$ be an absolute constant and remind r is the parameter representing the rank of the latent transition matrix $\mathbf{A}$. We define $\mathcal{C}$ to be a subset of d by r real matrices taking values only on $\{0, \gamma\ (r/(dT))^{1/2}\}$:

$$\mathcal{C} := \Big\{\overline{\mathbf{M}} \in \mathbb{R}^{d\times r} : \overline{\mathbf{M}}_{jk} \in \big\{0, \gamma\big(\frac{r}{dT}\big)^{\frac{1}{2}}\big\},\ \forall 1 \le j \le d, 1 \le k \le r\Big\}. \tag{4.2}$$

Here γ is an absolute constant controlling the magnitude of the matrices in $\mathcal{C}$, satisfying:

$$0 \le \gamma \le \sqrt{\frac{\alpha\log 2}{8}}. \tag{4.3}$$

Built on the matrix set $\mathcal{C}$, we further define the following set of matrices:

$$\mathcal{B}(\mathcal{C}) := \Big\{\mathbf{M} = (\overline{\mathbf{M}} \mid \dots \mid \overline{\mathbf{M}} \mid \mathbf{0}) \in \mathbb{R}^{d\times d} : \overline{\mathbf{M}} \in \mathcal{C}\Big\},$$

where $\mathbf{0}$ is a d by $(d - r\lfloor d/r\rfloor)$ matrix with all elements equal to zero. By construction and the condition that $rd/T = o(1)$, for sufficiently large T, any matrix in $\mathcal{B}(\mathcal{C})$ has rank at most r, and has the spectral norm upper bounded by $\gamma_{\max}$, as long as $\gamma_{\max}$ is an absolute positive constant. Thus, $\mathcal{B}(\mathcal{C})$ is a subset of $\mathcal{A}_M(r, \gamma_{\max})$.

We are now ready to prove the theorem. First, we note, using the fact that $\mathbf{I}_d \in \mathcal{V}_M$ and the set of identity functions belongs to $\mathcal{V}_f$, it suffices to show that (4.1) holds when the supremum is taken over the probability space with the specific $\mathbf{\Sigma} = \mathbf{I}_d$ and $\boldsymbol{f} = \boldsymbol{f}^0$, the set of identity functions, and the latent sequence $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ Gaussian distributed. In other words, it is sufficient to study the setting where $\boldsymbol{Z}_t \sim N_d(\mathbf{0}, \mathbf{I}_d)$ follows a standard Gaussian distribution and $\boldsymbol{X}_t = \boldsymbol{Z}_t$.

We then turn to prove (4.1). To this end, we exploit the general framework introduced by Lucien Le Cam and detailed in Tsybakov (2009). There are two steps.

(i) We construct a sufficiently small subset $\mathcal{A}^0$ of $\mathcal{B}(\mathcal{C})$ and calculate its cardinality.

(ii) We show (4.1) holds when the supremum is taken over all $\mathbf{A} \in \mathcal{A}^0$.

In detail, first, similar to Koltchinskii et al. (2011), using Lemma A2.1, we deduce there exists a subset $\mathcal{A}^0 \subset \mathcal{B}(\mathcal{C})$ containing the d by d zero matrix and having the cardinality $\text{card}(\mathcal{A}^0) \geq 2^{dr/8} + 1$. In addition, $\mathcal{A}^0$ satisfies, for any two distinct elements $\mathbf{A}_1 = (\bar{\mathbf{A}}_1 \mid \dots \mid \bar{\mathbf{A}}_1 \mid \mathbf{0})$ and $\mathbf{A}_2 = (\bar{\mathbf{A}}_2 \mid \dots \mid \bar{\mathbf{A}}_2 \mid \mathbf{0})$ of $\mathcal{A}^0$, we have

$$\rho_{\mathsf{H}}(\text{vec}(\bar{\mathbf{A}}_1), \text{vec}(\bar{\mathbf{A}}_2)) \geq \frac{rd}{8},$$

where ρ_{H} denotes the Hamming distance between two vectors.

By the definition of the Frobenius norm, we then have

$$\|\mathbf{A}_1 - \mathbf{A}_2\|_{\mathsf{F}}^2 \geq \frac{rd}{8}\Big(\gamma^2 \frac{r}{dT}\Big)\Big\lfloor \frac{d}{r} \Big\rfloor \geq \frac{\gamma^2 rd}{16T},$$

where the last inequality is due to the fact that $d \geq r$ implies $(d/r) < 2\lfloor d/r\rfloor$.

The second part distinguishes our proof from Koltchinskii et al. (2011). Specifically, focusing on the subset $\mathcal{A}^0$, we aim to bound (4.1) with the supremum taken over all $\mathbf{A} \in \mathcal{A}^0$. To this end, it is sufficient to consider the VAR process $\{\boldsymbol{Z}_t\}_{t=1}^T$ with the transition matrix $\mathbf{A}$. Let $\boldsymbol{Y} := (\boldsymbol{Z}_T^\mathsf{T}, \dots, \boldsymbol{Z}_1^\mathsf{T})^\mathsf{T} \sim N_{Td}(\mathbf{0}, \mathbf{V}_\mathbf{A})$ be the vectorized version of $\{\boldsymbol{Z}_t\}_{t=1}^T$, with the covariance matrix $\mathbf{V}_\mathbf{A}$ defined as

$$\mathbf{V}_\mathbf{A} := \begin{bmatrix} \mathbf{I}_d & \mathbf{A} & \mathbf{A}^2 & \dots & \mathbf{A}^{T-1} \\ \mathbf{A}^\mathsf{T} & \mathbf{I}_d & \mathbf{A} & \dots & \mathbf{A}^{T-2} \\ (\mathbf{A}^\mathsf{T})^2 & \mathbf{A}^\mathsf{T} & \mathbf{I}_d & \dots & \mathbf{A}^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (\mathbf{A}^\mathsf{T})^{T-1} & (\mathbf{A}^\mathsf{T})^{T-2} & (\mathbf{A}^\mathsf{T})^{T-3} & \dots & \mathbf{I}_d \end{bmatrix} \in \mathbb{R}^{Td\times Td}. \tag{4.4}$$

Let $\mathbb{P}_\mathbf{A}$ be the probability measure of $\boldsymbol{Y}$ and $\mathbb{P}_0$ be the probability measure of the Td-dimensional standard Gaussian. For any two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ over a set $\mathcal{X}$ satisfying that $\mathbb{P}_1$ is absolutely continuous with respect to $\mathbb{P}_2$, the Kullback-Leibler divergence from $\mathbb{P}_1$ to $\mathbb{P}_2$ is

$$D_{\mathsf{KL}}(\mathbb{P}_1\|\mathbb{P}_2) := \int_{\mathcal{X}} \log \frac{\mathsf{d}\mathbb{P}_1}{\mathsf{d}\mathbb{P}_2} \mathsf{d}\mathbb{P}_1,$$

where $\mathsf{d}\mathbb{P}_1/\mathsf{d}\mathbb{P}_2$ denotes the Radon-Nikodym derivative of $\mathbb{P}_1$ with respect to $\mathbb{P}_2$. Then, for any $\mathbf{A} \in \mathcal{A}^0$, the Kullback-Leibler divergence $D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}}\|\mathbb{P}_{\mathbf{0}})$ satisfies

$$D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}}\|\mathbb{P}_{\mathbf{0}}) = \frac{1}{2}\Big(\mathrm{Tr}(\mathbf{V}_{\mathbf{A}} - \mathbf{I}_{Td}) - \log\det \mathbf{V}_{\mathbf{A}}\Big),$$

where $\log\det(\cdot)$ represents the log determinant. Because $\mathrm{Tr}(\mathbf{V}_{\mathbf{A}}) = Td$, we further have

$$D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}}\|\mathbb{P}_{\mathbf{0}}) = -\frac{1}{2}\log\det \mathbf{V}_{\mathbf{A}}. \tag{4.5}$$

Of note, this part is 0 when $\mathbf{A} = 0$, corresponding to the case of independent observations, and hence is ignorable in the regular calculation of minimax lower bounds. Some involved calculations (details are in Section A1.2) yield

$$D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}}\|\mathbb{P}_{\mathbf{0}}) \le \gamma^2 rd. \tag{4.6}$$

Combining (4.3) and (4.6), we deduce

$$\frac{1}{\mathrm{card}(\mathcal{A}^0) - 1} \sum_{\mathbf{A}\in\mathcal{A}^0} D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}}\|\mathbb{P}_{\mathbf{0}}) \le \alpha \log(\mathrm{card}(\mathcal{A}^0) - 1). \tag{4.7}$$

Combining (4.7) and Lemma A2.2, we have, for arbitrary absolute constant $\beta \in (0,1)$, there exists a sufficiently small absolute constant C_β such that

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}\in\mathcal{A}^0, \mathbf{\Sigma}=\mathbf{I}_d, \boldsymbol{f}=\boldsymbol{f}^0} \mathbb{P}_{\mathbf{A},\mathbf{\Sigma},\boldsymbol{f}}\Big(\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathsf{F}}^2 \ge C_\beta \cdot \frac{rd}{T}\Big) > \beta.$$

This further yields

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}\in\mathcal{A}_M(r,a), \mathbf{\Sigma}\in\mathcal{V}_M, \boldsymbol{f}\in\mathcal{V}_f} \mathbb{P}_{\mathbf{A},\mathbf{\Sigma},\boldsymbol{f}}\Big(\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\mathsf{F}}^2 \ge C_\beta \cdot \frac{rd}{T}\Big) > \beta,$$

and hence completes the proof. □

5 Extensions to Markov chains of higher orders

For presentation clearness, we focus on the order one Markov chains in the above sections. However, it is very straightforward to extend the results to Markov chains of order $p > 1$. This section discusses such an extension.

Let's consider the following centered time series:

$$\boldsymbol{f}(\boldsymbol{X}_t) = \sum_{j=1}^{p} \mathbf{A}_j \boldsymbol{f}(\boldsymbol{X}_{t-j}) + \boldsymbol{E}_t \text{ and } \boldsymbol{Z}_t := \boldsymbol{f}(\boldsymbol{X}_t) \sim EC_d(\mathbf{0}, \mathbf{\Sigma}, \xi), \quad \text{for any } t \in \mathbb{Z}, \tag{5.1}$$

which satisfies $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ is globally elliptically distributed. By a repeatedly used argument of converting the VAR(p) model to the VAR(1) one (Han et al., 2015; Basu and Michailidis, 2015), we can write

$$\widetilde{\boldsymbol{Z}}_t = \widetilde{\mathbf{A}}\widetilde{\boldsymbol{Z}}_{t-1} + \widetilde{\boldsymbol{E}}_t,$$

where we denote

$$\widetilde{\boldsymbol{Z}}_t = \begin{pmatrix} \boldsymbol{Z}_t \\ \boldsymbol{Z}_{t-1} \\ \vdots \\ \boldsymbol{Z}_{t-p+1} \end{pmatrix}, \quad \widetilde{\mathbf{A}} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \dots & \mathbf{A}_{p-1} & \mathbf{A}_p \\ \mathbf{I}_d & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_d & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \dots & \dots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I}_d & \mathbf{0} \end{pmatrix}, \quad \text{and} \quad \widetilde{\boldsymbol{E}}_t = \begin{pmatrix} \boldsymbol{E}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}.$$

Notice, for any $t \in \mathbb{Z}$, we have

$$\boldsymbol{Z}_t = \underbrace{(\mathbf{A}_1, \dots, \mathbf{A}_p)}_{\mathbf{B} \in \mathbb{R}^{d \times dp}} \cdot \underbrace{(\boldsymbol{Z}_{t-1}^{\mathsf{T}}, \boldsymbol{Z}_{t-2}^{\mathsf{T}}, \dots, \boldsymbol{Z}_{t-p}^{\mathsf{T}})^{\mathsf{T}}}_{\mathbf{Y}_t \in \mathbb{R}^{dp \times 1}} + \boldsymbol{E}_t,$$

which yields

$$\mathbf{B} = \mathbf{\Omega}_1^{\mathsf{T}} \mathbf{\Omega}^{-1}, \quad \text{where} \quad \mathbf{\Omega} := \mathrm{Cov}(\boldsymbol{Y}_t) \in \mathbb{R}^{dp \times dp} \text{ and } \mathbf{\Omega}_1 := \mathrm{Cov}(\boldsymbol{Y}_t, \boldsymbol{Z}_t) \in \mathbb{R}^{dp \times d}.$$

Accordingly, we could exploit Lemma 2.2 again and propose a similar algorithm to estimate $\mathbf{B}$. In detail, reminding $\boldsymbol{S}_t$ is defined in (2.4), we write

$$\widehat{\mathbf{K}} := \frac{1}{T-p+1} \sum_{t=p}^{T} (\boldsymbol{S}_t^{\mathsf{T}}, \boldsymbol{S}_{t-1}^{\mathsf{T}}, \dots, \boldsymbol{S}_{t-p+1}^{\mathsf{T}})^{\mathsf{T}} \cdot (\boldsymbol{S}_t^{\mathsf{T}}, \boldsymbol{S}_{t-1}^{\mathsf{T}}, \dots, \boldsymbol{S}_{t-p+1}^{\mathsf{T}}),$$

$$\widehat{\mathbf{K}}_1 := \frac{1}{T-p} \sum_{t=p}^{T-1} (\boldsymbol{S}_t^{\mathsf{T}}, \boldsymbol{S}_{t-1}^{\mathsf{T}}, \dots, \boldsymbol{S}_{t-p+1}^{\mathsf{T}})^{\mathsf{T}} \cdot \boldsymbol{S}_{t+1}^{\mathsf{T}},$$

$$\text{and} \quad \widehat{\mathbf{\Omega}} := \sin\Big(\frac{\pi}{2}\widehat{\mathbf{K}}\Big), \quad \widehat{\mathbf{\Omega}}_1 := \sin\Big(\frac{\pi}{2}\widehat{\mathbf{K}}_1\Big).$$

We then estimate $\mathbf{B}$ by

$$\widehat{\mathbf{B}}_\lambda := \underset{\mathbf{Q} \in \mathbb{R}^{d \times dp}}{\mathrm{argmin}} \; L_\lambda(\mathbf{Q}; \widehat{\mathbf{\Omega}}, \widehat{\mathbf{\Omega}}_1).$$

To study the properties of $\widehat{\mathbf{B}}_\lambda$, we require Assumption **(A2)** and an additional assumption on the sequence $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$.

- **Assumption (B1).** We assume $\lambda_{\min}(\mathbf{\Omega})$ and $\lambda_{\max}(\mathbf{\Omega})$ are lower and upper bounded by two absolute positive constants respectively.

Here, on one hand, similar to the arguments in Section 3, Assumption **(A2)** holds for Gaussian and multivariate t models satisfying (5.1). On the other hand, Assumption **(B1)** implies $\lambda_{\min}(\mathbf{\Sigma})$ and $\lambda_{\max}(\mathbf{\Sigma})$ are also lower and upper bounded by two absolute positive constants, and hence is a stronger assumption than **(A1)**.

Let's define the following set of matrices:

$$\mathcal{B}_M(r, \gamma_{\max}) := \Big\{\mathbf{M} \in \mathbb{R}^{d \times dp} : \mathrm{rank}(\mathbf{M}) \le r, \|\mathbf{M}\|_2 \le \gamma_{\max}\Big\}.$$

The next theorem characterizes an upper bound of the estimation error $\|\widehat{\mathbf{B}}_\lambda - \mathbf{B}\|_{\mathsf{F}}$.

Theorem 5.1. Assume Model (5.1) and Assumptions **(A2), (B1)** hold. Assume $\mathbf{B} \in \mathcal{B}_M(r; \gamma_{\max})$ for some absolute positive constant $\gamma_{\max}$. Suppose there exist three absolute constants $\mu, C_4, C_5 > 0$

such that

$$\lambda \geq 2\big(C_4\gamma_{\max}\sqrt{dp/(T-p+1)} + C_5\sqrt{dp/(T-p)}\big),$$

and

$$\mu \leq \lambda_{\min}(\mathbf{\Omega}) - C_4\sqrt{dp/(T-p+1)}.$$

Further assume there exists an absolute constant $\kappa_2' > 0$ such that $T \geq \kappa_2'(dp)^{2/\gamma_0 - 1}$ where $\gamma_0 = \gamma_1/(\gamma_1+1)$ and $T \geq 4$. We then have

$$\begin{aligned}\mathbb{P}\Big(\|\widehat{\mathbf{B}}_\lambda - \mathbf{B}\|_{\mathsf{F}} \geq \frac{1}{2\mu}\big(\lambda\sqrt{r} + 2\sqrt{2}C_4\gamma_{\max}\sqrt{\frac{drp}{T-p+1}} + 2\sqrt{2}C_5\sqrt{\frac{drp}{T-p}}\big)\Big)\\ = O\Big(\exp(-C_6\cdot dp) + (dp)^2\exp\Big(-2\sqrt{\frac{T-p}{dp}}\Big)\Big),\end{aligned}$$

where $C_6 > 0$ is a constant only depending on $C_4, C_5, \kappa_1, \kappa_2', \gamma_1, \lambda_{\max}(\Omega)$, and $\lambda_{\min}(\Omega)$.

Theorem 5.1 confirms, when p is fixed and λ is appropriately chosen, our approach could attain the same $\sqrt{dr/T}$ rate of convergence as in the VAR(1) model. This also confirms the minimax rate-optimality in studying the general VAR(p) models.

6 Experimental results

This section provides the empirical simulation and real stock market data analysis. It is split into two parts. In the first part, we provide a computationally efficient algorithm to calculate $\widehat{\mathbf{A}}_\lambda$. The second part includes the numerical results.

6.1 Algorithm

We first provide a computationally efficient algorithm to implement (2.2). For this, we exploit the contractive Peaceman-Rachford splitting method (PRSM) (Peaceman and Rachford, 1955; He et al., 2014). In detail, letting $\widetilde{\mathbf{A}}_\lambda^G := [\widehat{\mathbf{A}}_\lambda^G]^{\mathsf{T}}$, it is obvious

$$\widetilde{\mathbf{A}}_\lambda^G = \operatorname*{argmin}_{\mathbf{Q}\in\mathbb{R}^{n\times m}} \langle -2\widehat{\mathbf{M}}_1 + \widehat{\mathbf{M}}\mathbf{Q}, \mathbf{Q}\rangle + \lambda\|\mathbf{Q}\|_*.$$

The above optimization problem falls into the application regime of the contractive PRSM algorithm. Similar to Fan et al. (2014), we can show $\widetilde{\mathbf{A}}_\lambda^G$ is the convergence of the sequence $\{\mathbf{Y}^{(k)}\}_{k\geq 1}$ in the following iterative scheme:

$$\begin{cases}\mathbf{X}^{(k+1)} = (2\widehat{\mathbf{M}} + \beta\mathbf{I}_n)^{-1}(2\widehat{\mathbf{M}}_1 + \beta\mathbf{Y}^{(k)} + \mathbf{P}^{(k)}),\\ \mathbf{P}^{(k+1/2)} = \mathbf{P}^{(k)} - \alpha\beta(\mathbf{X}^{(k+1)} - \mathbf{Y}^{(k)}),\\ \mathbf{Y}^{(k+1)} = \mathcal{S}_{\lambda/\beta}(\mathbf{X}^{(k+1)} - \mathbf{P}^{(k+1/2)}/\beta),\\ \mathbf{P}^{(k+1)} = \mathbf{P}^{(k+1/2)} - \alpha\beta(\mathbf{X}^{(k+1)} - \mathbf{Y}^{(k+1)}),\end{cases}$$

where $\mathcal{S}_\tau(\cdot)$ is the conventional matrix soft thresholding operator defined in Fan et al. (2014), α is a relaxation parameter, and β is a penalty parameter. We choose $\alpha = 0.9$ and $\beta = 1$ following the discussions in Fan et al. (2014).

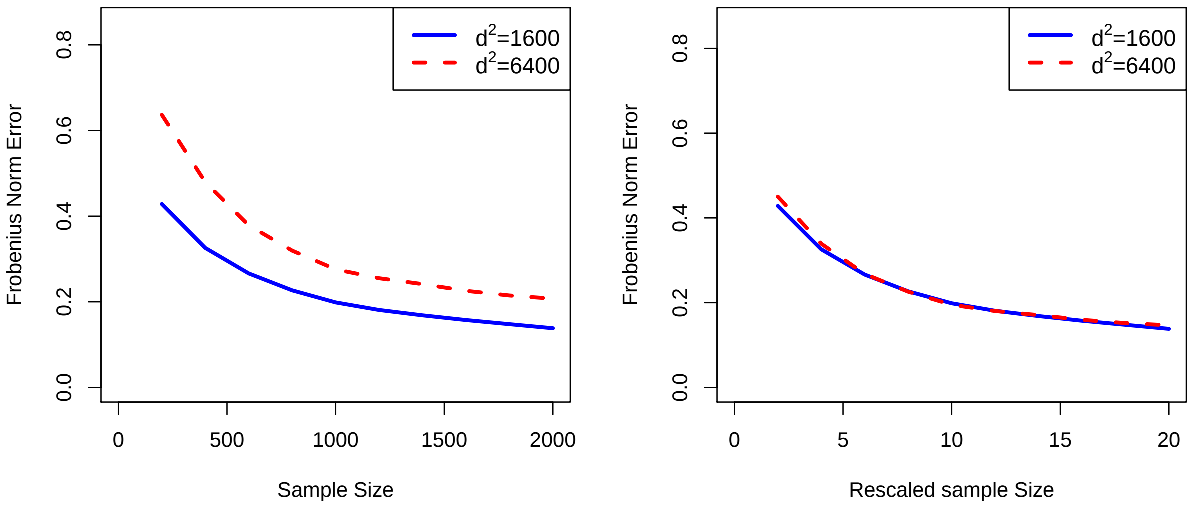

Figure 1: The estimation errors plotted against the sample sizes (T) and the rescaled sample sizes ($T/(rd)$). The errors between the estimators and the true transition matrix are measured under the Frobenius norm. We consider Scheme (**M1**). We fix $r = 3$ and select $d = 40$ or 80. In calculating the errors, we assume the rank r is known. The results are obtained over 1,000 replications.

We further discuss how to determine the tuning parameter λ. This involves a vast literature. Here we follow a procedure provided in Lam and Yao (2012), exploiting the ratio estimator. In detail, the determination of λ is in two steps.

- In the first step, we choose a small enough λ_0 such that
$$\text{rank}(\widetilde{\mathbf{A}}^G_{\lambda_0}) = d/2.$$
- In the second step, we select $\widehat{r}$ to be
$$\widehat{r} = \text{argmin}\left\{\frac{\sigma_{j+1}(\widetilde{\mathbf{A}}^G_{\lambda_0})}{\sigma_j(\widetilde{\mathbf{A}}^G_{\lambda_0})} : 1 \le j \le \frac{d}{2} - 1\right\}.$$
The selected tuning parameter λ is then chosen to be the minimum value such that the rank of $\widetilde{\mathbf{A}}^G_{\lambda}$ is $\widehat{r}$.

6.2 Numerical results

This section gives the results on studying synthetic and real stock market data.

6.2.1 Synthetic data analysis

This section provides the numerical results on the synthetic data. We consider the following two methods:

- NW: the linear regression procedure in Negahban and Wainwright (2011);
- HXL: the proposed robust-sign-transformation procedure.

For comparison fairness, we use the contractive PRSM algorithm for calculating the estimates of both NW and HXL. We focus on the following two time series schemes belonging to Model (1.1).

- **(M1).** $\{\boldsymbol{Z}_t \in \mathbb{R}^d\}_{t\in\mathbb{Z}}$ is a Gaussian sequence, and $\boldsymbol{X}_t$ is equal to $\boldsymbol{Z}_t$.
- **(M2).** $\{\boldsymbol{Z}_t \in \mathbb{R}^d\}_{t\in\mathbb{Z}}$ is multivariate t-distributed of degree of freedom 3, and $\boldsymbol{X}_t$ is a transformed version of $\boldsymbol{Z}_t$ using transformation functions $f_1(x) = \cdots = f_d(x) = x^{1/3}$.

We adopt a similar setting as in Negahban and Wainwright (2011) for generating $\mathbf{A}, \boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}_E := \mathrm{Cov}(\mathbf{E}_1)$. Specifically, let the transition matrix $\mathbf{A}$ be generated through $\mathbf{A} = \mathbf{X}\mathbf{Y}^T/(3d)$, where $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d\times r}$ have entries all independently distributed to $N_1(0,1)$. We let $\boldsymbol{\Sigma}_E = \mathbf{I}_d$ and hence further determine $\boldsymbol{\Sigma}$ due to the stationary condition. In the end, we rescale the system to make $\mathrm{diag}(\boldsymbol{\Sigma}) = \mathbf{I}_d$.

We first study the scaling of the estimation error $\|\widehat{\mathbf{A}}_\lambda - \mathbf{A}\|_{\mathsf{F}}$ compared to (T, d). To this end, we focus on Scheme **(M1)**, fix $r = 3$, increase d from 40 to 80, and assume r is known in tuning the parameter λ. Figure 1 illustrates the scaling of averaged errors, $\|\widehat{\mathbf{A}}_\lambda - \mathbf{A}\|_{\mathsf{F}}$, compared to T as well as $T/(rd)$ over 1,000 replications. It gives a similar "stacking" phenomenon as in Negahban and Wainwright (2011), and backs up Theorem 3.4.

Secondly, we compare HXL to NW on Schemes **(M1)** and **(M2)**. Fixing $r = 3, d = 40, T = 400$, and assuming the true rank r is known, Table 1 reports the estimators' averaged errors and their standard deviations when T increases from 200 to 2,000. Table 1 confirms HXL beats NW when the data are nonGaussian, while attaining comparable performance under the Gaussian model. In addition, HXL performs closely under the Gaussian and nonGaussian settings, though the variances slightly inflate under the nonGaussian setting.

6.2.2 Stock market data analysis

We analyze the log returns of daily closing prices from the Standard & Poor 100 (S&P 100) index (`finance.yahoo.com`). We focus on two periods: January 1, 2003 to December 31, 2006 and January 1, 2009 to December 31, 2012, the pre- and post-financial crisis time regions. This gives us 1,005 and 1,006 daily returns. We study 90 companies that are constantly in the S&P 100 index. The 90 stocks come from 10 Global Industry Classification Standard (GICS) sectors, including Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Materials, Telecommunications Services, and Utilities.

We apply HXL to the log returns in these two periods. This gives us two estimates, $\widehat{\mathbf{A}}_1$ of rank 2, and $\widehat{\mathbf{A}}_2$ of rank 1, via the rank selection procedure described in Lam and Yao (2012). Table 2 illustrates the ranks of $\widehat{\mathbf{A}}_1$ and $\widehat{\mathbf{A}}_2$. Table 2 further shows the top 10 stocks of largest magnitudes in right-singular vectors of $\widehat{\mathbf{A}}_1$ and $\widehat{\mathbf{A}}_2$[4]. Here the right-singular vectors correspond to the latent

[4] We present the abbreviations of the stock names, while the full names can be found on `finance.yahoo.com`.

Table 1: Comparison of averaged estimation errors (by taking the median) under the Frobenius norm for two competing methods over 1,000 replications. The standard deviations are presented in the parentheses. In each simulation, we set $d = 40$, $r = 3$, and T from 200 to 2,000. Schemes **(M1)** and **(M2)** are considered.

T	NW	HXL	NW	HXL
	(M1)		**(M2)**	
200	**0.39** (0.03)	0.44 (0.04)	1.65 (0.92)	**0.45** (0.08)
400	**0.32** (0.03)	**0.32** (0.04)	0.86 (0.80)	**0.32** (0.06)
600	**0.25** (0.03)	0.28 (0.03)	0.65 (0.73)	**0.28** (0.06)
800	**0.24** (0.03)	**0.24** (0.02)	0.55 (0.26)	**0.25** (0.03)
1000	**0.18** (0.02)	0.20 (0.02)	0.67 (0.18)	**0.22** (0.03)
1200	**0.16** (0.03)	0.18 (0.03)	0.65 (0.29)	**0.19** (0.07)
1400	**0.15** (0.02)	0.17 (0.03)	0.54 (0.13)	**0.18** (0.05)
1600	**0.16** (0.01)	**0.16** (0.02)	0.59 (0.27)	**0.18** (0.06)
1800	**0.13** (0.02)	0.15 (0.01)	0.71 (0.37)	**0.15** (0.04)
2000	**0.12** (0.02)	0.14 (0.02)	0.63 (0.35)	**0.15** (0.03)

driving factors in the model of Basu (2014), and hence the corresponding stocks can be interpreted as the obtained driving factors in each period.

There are some notable discoveries. First, the driving factors for pre- and post- 2007-2008 financial crisis are utterly different. This is consistent to the relevant economics and finance principles. Secondly, we find, for the period 2003-2006, the stocks from Energy and Industrials sectors dominate the driving factor. The domination of stocks in Energy and Industrials sectors (e.g., Occidental Petroleum, Caterpillar Inc., Raytheon Company) are relevant to the fact that in 2003 there was a huge increase in the international oil price and it was convoluted with the 2003 invasion of Iraq. In comparison, for the period 2009-2012, we find the stocks in Energy and Financials sectors dominate the driving factor. On one hand, the domination of stocks in the Energy sector in both pre- and post-crisis period is as expected, because the energy forms the foundation for economic development and has strong impact on many other sectors (Kilian and Park, 2009). On the other hand, the role of stocks in the Industrials sector on driving the stock market is replaced by those in the Financials sector. This is an interesting discovery and we will investigate this pattern shift in more details in the future.

Table 2: The list of ranks and driving factors calculated from the transition matrix estimators $\widehat{\mathbf{A}}_1$ and $\widehat{\mathbf{A}}_2$ using HXL. Here the m-th value in the column "percentage" represents the ratio of ℓ_2 norm of the top m entries' values to the ℓ_2 norm of the whole singular vector, which is 1.

$\widehat{\mathbf{A}}_1$				$\widehat{\mathbf{A}}_2$			
rank	driving factors	stock sector	percentage	rank	driving factors	stock sector	percentage
2	OXY	E	0.31	1	BK	F	0.25
	CVS	CS	0.38		AXP	F	0.34
	CAT	I	0.44		FDX	I	0.42
	AMZN	CD	0.49		T	TS	0.48
	AIG	F	0.53		NOV	E	0.53
	WMT	CS	0.56		UTX	CD	0.57
	RTN	I	0.59		HAL	E	0.60
	CVZ	TS	0.62		TWX	CD	0.64
	NSC	I	0.64		GS	F	0.66
	UPS	I	0.66		AAPL	IT	0.69
	JNJ	HC	0.22				
	PG	CS	0.26				
	WFC	F	0.29				
	ORCL	IT	0.31				
	WMT	CS	0.32				
	CMCSA	CD	0.33				
	SO	U	0.34				
	MSFT	IT	0.34				
	PEP	CS	0.34				
	GE	I	0.34				

Abbreviations for stock sectors: Consumer Discretionary(CD), Consumer Staples(CS), Energy(E), Financials(F), Health Care(HC), Industrials(I), Information Technology(IT), Materials(M), Telecommunications Services(TS), Utilities(U).

7 Discussions

7.1 Extension to study sparse transition matrix

The techniques developed for establishing upper and lower bounds in the time series analysis are very general, and could be extended to study many other problems.

For example, an immediate problem related to low-rank transition matrix estimation is sparse transition matrix estimation, where Han et al. (2015) and Basu and Michailidis (2015) have provided a set of results under the Gaussian assumption. Here we note the proof of Theorem 3.4 can easily devise the following proposition.

Proposition 7.1. Suppose the conditions in Theorem 3.4 hold. We then have

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{2,s} = O_P(\sqrt{s\log(d/s)/T}) \ \text{ and } \ \|\widehat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}_1\|_{2,s} = O_P(\sqrt{s\log(d/s)/T}), \tag{7.1}$$

where the restricted spectral norm $\|\cdot\|_{2,s}$ is defined as

$$\|\mathbf{M}\|_{2,s} := \sup_{\|\boldsymbol{v}\|_0 \le s} \frac{(\boldsymbol{v}^T\mathbf{M}\mathbf{M}^{\mathsf{T}}\boldsymbol{v})^{1/2}}{\|\boldsymbol{v}\|_2}.$$

Proposition 7.1 is in parallel to Proposition 2.4 in Basu and Michailidis (2015), which is the key in their analysis. Hence, by exploiting Proposition 7.1, we can similarly plug $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Sigma}}_1$ into the algorithm in Basu and Michailidis (2015), and build a similar upper bound for our robust-sign-transformation sparsity-induced estimator.

7.2 Estimation of transformation functions

In this manuscript we did not discuss estimation of transformation functions $f_1, \ldots, f_d$ in Model (1.1) since they are beyond the main interest. However, in practice, in addition to estimating the temporal correlation pattern, approximating transformation functions could also be of interest for rebuilding the latent factors, and hence deserves a discussion.

It might be attempting to conjecture that estimation of such nonparametric functions is statistically very challenging, especially when d is relatively large. However, as has been shown in Theorem 4.6 in Liu et al. (2012), when the observed sequence is i.i.d. Gaussian copula distributed with d moderately growing to infinity with T, $f_1, \ldots, f_d$ can be uniformly approached by a truncated nonparametric likelihood estimator. Exactly the same set of estimators can be employed under Model (1.1), and are able to be shown to enjoy the same uniform convergence property as their analogues in Liu et al. (2012) under the weak dependence assumption **(A2)**.

7.3 Addition comments

First, the theories developed in this paper heavily rely on the stationary assumption. The techniques developed in this paper cannot be trivially applied to study non-stationary VAR models. We will leave this for future studies. Secondly, we evaluate the temporal dependence of the VAR model via characterizing the corresponding mixing coefficients. This is in parallel to the settings of theoretical studies in low dimensions (Chen and Fan, 2006; Beare, 2010). In comparison, Negahban and Wainwright (2011), Loh and Wainwright (2012), Han et al. (2015), and Basu and Michailidis (2015) gave alternative upper bounds with $\|\mathbf{A}\|_2$ and spectral densities involved. For this, they require a stringent Gaussian assumption on the time series. Extending such results to study the non-linear nonGaussian time series is interesting as well as very challenging, and is left for future studies.

References

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. and Liao, Y. (2016). Efficient estimation of approximate factor models. *Journal of Econometrics*, 191(1):1–18.

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92.

Bandyopadhyay, S. (2006). A note on strong mixing. Technical report, Iowa State University.

Barber, R. F. and Kolar, M. (2015). ROCKET: Robust confidence intervals via Kendall's tau for transelliptical graphical models. *arXiv preprint arXiv:1502.07641*.

Basu, S. (2014). *Modeling and Estimation of High-dimensional Vector Autoregressions*. PhD thesis, The University of Michigan.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567.

Beare, B. K. (2010). Copulas and temporal dependence. *Econometrica*, 78(1):395–410.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2(2):107–144.

Bunea, F., She, Y., and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309.

Bunea, F., She, Y., and Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, 40(5):2359–2388.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11(1)–11(37).

Candes, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772.

Chang, J., Guo, B., and Yao, Q. (2015). High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. *Journal of Econometrics*, 189(2):297–312.

Chen, X. and Fan, Y. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, 130(2):307–335.

Chen, X., Koenker, R., and Xiao, Z. (2009). Copula-based nonlinear quantile autoregression. *The Econometrics Journal*, 12(s1):S50–S67.

Chen, X., Xu, M., and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics*, 41(6):2994–3021.

Christiano, L. J., Eichenbaum, M., and Evans, C. L. (1999). Monetary policy shocks: What have we learned and to what end? *Handbook of Macroeconomics*, 1(A):65–148.

Davis, R. A., Zang, P., and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096.

Den Haan, W. J. and Levin, A. (1998). Vector autoregressive covariance matrix estimation. Technical report, University of California, San Diego.

Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications*, 84(2):313–342.

Fan, J., Han, F., and Liu, H. (2014). PAGE: Robust pattern guided estimation of large covariance matrix. Technical report, Princeton University.

Fan, J., Liao, Y., and Mincheva, M. (2011a). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320–3356.

Fan, J., Lv, J., and Qi, L. (2011b). Sparse high dimensional models in economics. *Annual Review of Economics*, 3:291–317.

Fan, J., Zhang, J., and Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606.

Fang, H.-B., Fang, K.-T., and Kotz, S. (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82(1):1–16.

Fang, K.-T., Kotz, S., and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall.

Han, F. and Li, Y. (2017). Bernstein-type inequality for strongly mixing random matrices. Technical report, University of Washington.

Han, F. and Liu, H. (2013). Principal component analysis on nonGaussian dependent data. In *Proceedings of the 30th International Conference on Machine Learning*, pages 240–248.

Han, F. and Liu, H. (2017). Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli*, 23(1):23–57.

Han, F., Lu, H., and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 16:3115–3150.

He, B., Liu, H., Wang, Z., and Yuan, X. (2014). A strictly contractive Peaceman–Rachford splitting method for convex programming. *SIAM Journal on Optimization*, 24(3):1011–1040.

Hsing, T. and Wu, W. B. (2004). On weighted U-statistics for stationary processes. *Annals of Probability*, 32(2):1600–1631.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264.

Kilian, L. and Park, C. (2009). The impact of oil price shocks on the U.S. stock market. *International Economic Review*, 50(4):1267–1287.

Kolmogorov, A. N. and Rozanov, Y. A. (1960). On strong mixing conditions for stationary Gaussian processes. *Theory of Probability and Its Applications*, 5(2):204–208.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329.

Kruskal, W. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918.

Liebscher, E. (2005). Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *Journal of Time Series Analysis*, 26(5):669–689.

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–464.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., et al. (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Longla, M. and Peligrad, M. (2012). Some aspects of modeling dependence in copula-based Markov chains. *Journal of Multivariate Analysis*, 111:234–240.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer.

Merlevède, F., Peligrad, M., and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474.

Michailidis, G. (2012). Statistical challenges in biological networks. *Journal of Computational and Graphical Statistics*, 21(4):840–855.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097.

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, 95(2):365–379.

Patton, A. (2012a). Copula methods for forecasting multivariate time series. *Handbook of Economic Forecasting*, 2:899–960.

Patton, A. J. (2012b). A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18.

Peaceman, D. W. and Rachford, Jr, H. H. (1955). The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for Industrial and Applied Mathematics*, 3(1):28–41.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-Rank Regression*. Springer.

Rémillard, B., Papageorgiou, N., and Soustra, F. (2012). Copula-based semiparametric models for multivariate time series. *Journal of Multivariate Analysis*, 110:30–42.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345.

Sancetta, A. (2008). Sample covariance shrinkage for high dimensional dependent data. *Journal of Multivariate Analysis*, 99(5):949–967.

Shojaie, A. and Michailidis, G. (2010). Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523.

Smith, S. M. (2012). The future of FMRI connectivity. *Neuroimage*, 62(2):1257–1266.

Song, S. and Bickel, P. J. (2011). Large vector auto regressions. *arXiv preprint arXiv:1106.3915*.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.

Wang, T. and Xia, Y. (2015). Whittle likelihood estimation of nonlinear autoregressive models with moving average residuals. *Journal of the American Statistical Association*, 110(511):1083–1099.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14150–14154.

Xiao, H. and Wu, W. B. (2012). Covariance matrix estimation for stationary time series. *The Annals of Statistics*, 40(1):466–493.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346.

Supplementary material to “Rate-optimal estimation of a high-dimensional semiparametric time series model”

Fang Han[*], Sheng Xu[†] and Han Liu[‡]

This supplementary material provides all proofs of results in the main text as well as some auxiliary lemmas.

A1 Proofs of main results

This section provides the proofs of Theorem 3.4 and Theorem 4.1. In the following, for any two matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{m\times n}$, we define the Hadamard product between $\mathbf{M}$ and $\mathbf{N}$ to be $\mathbf{M} \circ \mathbf{N} := [\mathbf{M}_{jk}\mathbf{N}_{jk}]$.

A1.1 Proof of Theorem 3.4

Proof. Remind the robust-sign-transformation version of the observed data point $\boldsymbol{X}_t$, for $t = 1, \ldots, T$, is defined as

$$\boldsymbol{S}_t := \text{sign}(\boldsymbol{X}_t) = \text{sign}(\boldsymbol{Z}_t).$$

We further define

$$\mathbf{T} := \text{Cov}(\boldsymbol{S}_t) = \mathbb{E}(\boldsymbol{S}_t\boldsymbol{S}_t^{\mathsf{T}}),$$
$$\mathbf{T}_1 := \text{Cov}(\boldsymbol{S}_t, \boldsymbol{S}_{t+1}) = \mathbb{E}(\boldsymbol{S}_t\boldsymbol{S}_{t+1}^{\mathsf{T}}).$$

Based on Lemma 2.2, it suffices to determine $\delta_1, \delta_2 > 0$ such that

$$\mathbb{P}\Big(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 \le \delta_1\Big) \ge 1 - \epsilon_1, \tag{A1.1}$$

$$\mathbb{P}\Big(\|\widehat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}_1\|_2 \le \delta_2\Big) \ge 1 - \epsilon_2, \tag{A1.2}$$

where δ_1, δ_2 are functions of (T, d) and ϵ_1, ϵ_2 go to zero when (T, d) goes to infinity. The proof is then split into two parts.

Step I. First, we study (A1.1). For this, we relate $\mathbf{\Sigma}$ to $\mathbf{T}$ via the following formula

$$\mathbf{\Sigma} = \sin(\frac{\pi}{2}\mathbf{T}).$$

[*] Department of Statistics, University of Washington, Seattle, WA 98195, USA; e-mail: `fanghan@uw.edu`

[†] Department of Statistics, Yale University, New Haven, CT 06511, USA; e-mail: `sheng.xu@yale.edu`

[‡] Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: `hanliu@princeton.edu`

Secondly, by Taylor's theorem, we have

$$\begin{aligned}\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} &= \sin\Big(\frac{\pi}{2}\widehat{\mathbf{T}}\Big) - \sin\Big(\frac{\pi}{2}\mathbf{T}\Big)\\ &= \cos\Big(\frac{\pi}{2}\mathbf{T}\Big)\circ\frac{\pi}{2}(\widehat{\mathbf{T}}-\mathbf{T}) - \frac{1}{2}\sin\Big(\frac{\pi}{2}\bar{\mathbf{T}}\Big)\circ\frac{\pi}{2}(\widehat{\mathbf{T}}-\mathbf{T})\circ\frac{\pi}{2}(\widehat{\mathbf{T}}-\mathbf{T}), \qquad \text{(A1.3)}\end{aligned}$$

for some matrix $\bar{\mathbf{T}}$ with each entry $\bar{\mathbf{T}}_{jk}$ being a number on the closed interval between $\mathbf{T}_{jk}$ and $\widehat{\mathbf{T}}_{jk}$. Applying the operator norm on both sides of (A1.3) and then using the triangle inequality on the right-hand side yields

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 \le \frac{\pi}{2}\underbrace{\|\cos\Big(\frac{\pi}{2}\mathbf{T}\Big)\circ(\widehat{\mathbf{T}}-\mathbf{T})\|_2}_{E_1} + \frac{\pi^2}{8}\underbrace{\|\sin\Big(\frac{\pi}{2}\bar{\mathbf{T}}\Big)\circ(\widehat{\mathbf{T}}-\mathbf{T})\circ(\widehat{\mathbf{T}}-\mathbf{T})\|_2}_{E_2}. \qquad \text{(A1.4)}$$

Hence, in order to establish a bound for $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2$, it suffices to establish appropriate bounds separately for the first order term E_1, and the second order term E_2.

Step I.1 We first consider the first order term E_1. For any $\boldsymbol{v} \in \mathbb{S}^{d-1}$ (the unit sphere in the d-dimensional Euclidean space), we define, for $t = 1, \ldots, T$,

$$Y_t := \boldsymbol{v}^\top \boldsymbol{S}_t = \boldsymbol{v}^\top \mathrm{sign}(\boldsymbol{X}_t).$$

There are two observations on $\{Y_t\}_{t=1}^T$. First, since $\{\boldsymbol{Z}_t\}_{t=1}^T$ satisfies the strong mixing condition and the link functions between $\{\boldsymbol{Z}_t\}_{t=1}^T$ and $\{Y_t^2\}_{t=1}^T$ are measurable, $\{Y_t^2\}_{t=1}^T$ also satisfies the strong mixing condition with

$$\alpha(n; Y_t^2) \le \exp(-C_{11}n^{\gamma_1}), \qquad \text{(A1.5)}$$

for the absolute constant $C_{11} = \kappa_1$ and any $n \ge 2$.

Secondly, because $\{\boldsymbol{Z}_t\}_{t=1}^T$ is elliptically distributed, $\{Y_t\}_{t=1}^T$ satisfies the subgaussian condition in (3.5). In particular, we have $\{Y_t^2\}_{t=1}^T$ satisfies that there exists a positive constant C_{12}, only depending on $\lambda_{\max}(\mathbf{\Sigma})$ and $\lambda_{\min}(\mathbf{\Sigma})$, such that, for all $\xi > 0$,

$$\mathbb{P}(|Y_t^2 - \mathbb{E}Y_t^2| > \xi) \le \exp\Big(1 - \frac{\xi}{C_{12}}\Big). \qquad \text{(A1.6)}$$

Notice, for arbitrary fixed $\boldsymbol{v} \in \mathbb{S}^{d-1}$,

$$\boldsymbol{v}^\top \widehat{\mathbf{T}} \boldsymbol{v} = \frac{1}{T}\sum_{t=1}^T (\boldsymbol{v}^\top \boldsymbol{S}_t)^2 = \frac{1}{T}\sum_{t=1}^T Y_t^2$$

and

$$\boldsymbol{v}^\top \mathbf{T} \boldsymbol{v} = \frac{1}{T}\sum_{t=1}^T \mathbb{E}(\boldsymbol{v}^\top \boldsymbol{S}_t)^2 = \frac{1}{T}\sum_{t=1}^T \mathbb{E}Y_t^2.$$

We then exploit the strong mixing condition (A1.5), the tail condition (A1.6), and employ Lemma A2.7 to deduce that there exist positive constants $C_{13} - C_{17}$, only depending on C_{11}, C_{12}, γ_1, such that for all $T \geq 4$ and $\eta_1 > 0$, we have

$$\mathbb{P}\big(|\boldsymbol{v}^\top \widehat{\mathbf{T}} \boldsymbol{v} - \boldsymbol{v}^\top \mathbf{T} \boldsymbol{v}| \geq \eta_1\big) \leq T \exp\Big(-\frac{(T\eta_1)^{\gamma_0}}{C_{13}}\Big) + \exp\Big(-\frac{(T\eta_1)^2}{C_{14}(1+C_{15}T)}\Big) + \exp\Big(-\frac{(T\eta_1)^2}{C_{16}T}\exp\Big(\frac{(T\eta_1)^{\gamma_0(1-\gamma_0)}}{C_{17}(\log T\eta_1)^{\gamma_0}}\Big)\Big),$$

where $\gamma_0 = \gamma_1/(\gamma_1+1) < 1$. Setting $\eta_1 = C_{10}\sqrt{d/T}$, we further have

$$\mathbb{P}\big(|\boldsymbol{v}^\top \widehat{\mathbf{T}} \boldsymbol{v} - \boldsymbol{v}^\top \mathbf{T} \boldsymbol{v}| \geq C_{10}\sqrt{\frac{d}{T}}\big) \leq \underbrace{T \exp\Big(-\frac{(C_{10}^2 Td)^{\gamma_0/2}}{C_{13}}\Big)}_{F_1} + \underbrace{\exp\Big(-\frac{C_{10}^2 Td}{C_{14}(1+C_{15}T)}\Big)}_{F_2} + \underbrace{\exp\Big(-\frac{C_{10}^2 d}{C_{16}}\exp\Big(\frac{(C_{10}^2 Td)^{\gamma_0(1-\gamma_0)/2}}{C_{17}(\log(C_{10}\sqrt{Td}))^{\gamma_0}}\Big)\Big)}_{F_3}.$$

Now we focus on F_1, F_2, F_3 and bound them successively. For F_1, since $T \geq \kappa_2 d^{2/\gamma_0 - 1}$, we have

$$(Td)^{\frac{\gamma_0}{2}} \geq \kappa_2^{\frac{\gamma_0}{2}} d,$$

which implies

$$T \leq \exp\Big(\frac{C_{10}^{\gamma_0}(Td)^{\frac{\gamma_0}{2}}}{C_{13}} - \frac{C_{10}^{\gamma_0}\kappa_2^{\frac{\gamma_0}{2}}}{2C_{13}} d\Big),$$

for large T. Hence, we deduce that

$$F_1 \leq \exp\Big(-\frac{(C_{10}^2\kappa_2)^{\frac{\gamma_0}{2}}}{2C_{13}} d\Big). \tag{A1.7}$$

For F_2, since $T \geq 1$, we have

$$F_2 \leq \exp\Big(-\frac{C_{10}^2 Td}{C_{14}(1+C_{15})T}\Big) = \exp\Big(-\frac{C_{10}^2 d}{C_{14}(1+C_{15})}\Big). \tag{A1.8}$$

For F_3, since $Td > 1$, we have $\log\sqrt{Td} > 0$, which implies

$$F_3 \leq \exp\Big(-\frac{C_{10}^2 d}{C_{16}}\Big). \tag{A1.9}$$

Combining (A1.7), (A1.8), and (A1.9), we have

$$\mathbb{P}\big(|\boldsymbol{v}^\top \widehat{\mathbf{T}} \boldsymbol{v} - \boldsymbol{v}^\top \mathbf{T} \boldsymbol{v}| \geq C_{10}\sqrt{d/T}\big) \leq \exp\Big(-\frac{(C_{10}^2\kappa_2)^{\frac{\gamma_0}{2}}}{2C_{13}} d\Big) + \exp\Big(-\frac{C_{10}^2}{C_{14}(1+C_{15})} d\Big) + \exp\Big(-\frac{C_{10}^2}{C_{16}} d\Big). \tag{A1.10}$$

Define $C_{18} := \min\{C_{10}^{\gamma_0}\kappa_2^{\gamma_0/2}C_{13}^{-1}/2, C_{10}^2(C_{14}(1+C_{15}))^{-1}, C_{10}^2 C_{16}^{-1}\}$. It follows from (A1.10) that

$$\mathbb{P}\big(|\boldsymbol{v}^\top \widehat{\mathbf{T}} \boldsymbol{v} - \boldsymbol{v}^\top \mathbf{T} \boldsymbol{v}| \geq C_{10}\sqrt{d/T}\big) \leq 3\exp(-C_{18} d). \tag{A1.11}$$

We then aim to bound $\|\widehat{\mathbf{T}} - \mathbf{T}\|_2$. To this end, we first define some additional notation. For any metric space (Ω, ρ), a subset $S_\epsilon(\Omega)$ is called an ϵ-net of Ω if every point $\omega \in \Omega$ can be approximated to within ϵ by some point $\xi \in S_\epsilon(\Omega)$. The minimal cardinality of the ϵ-net, if finite, is denoted by $\mathcal{N}(\Omega, \epsilon)$.

We define the (1/4)-net of $\mathbb{S}^{d-1}$, equipped with the Euclidean distance, as $\mathcal{S}_{1/4}$. According to Lemma A2.6, we have

$$\|\widehat{\mathbf{T}} - \mathbf{T}\|_2 = \sup_{\boldsymbol{v} \in \mathbb{S}^{d-1}} \big|\langle(\widehat{\mathbf{T}} - \mathbf{T})\boldsymbol{v}, \boldsymbol{v}\rangle\big| \leq 2 \sup_{\boldsymbol{v} \in \mathcal{S}_{1/4}} \big|\langle(\widehat{\mathbf{T}} - \mathbf{T})\boldsymbol{v}, \boldsymbol{v}\rangle\big|. \tag{A1.12}$$

Using Lemma A2.5, we also have

$$\mathcal{N}(\mathbb{S}^{d-1}, 1/4) \leq 9^d. \tag{A1.13}$$

Using (A1.12), we have

$$\mathbb{P}\big(\|\widehat{\mathbf{T}} - \mathbf{T}\|_2 > 2C_{10}\sqrt{\frac{d}{T}}\big) \leq \mathbb{P}\big(\sup_{\boldsymbol{v} \in \mathcal{S}_{1/4}} \big|\boldsymbol{v}^\mathsf{T}(\widehat{\mathbf{T}} - \mathbf{T})\boldsymbol{v}\big| > C_{10}\sqrt{\frac{d}{T}}\big). \tag{A1.14}$$

Combining (A1.13) and (A1.14), we have, for arbitrary $\boldsymbol{v} \in \mathcal{S}_{1/4}$,

$$\mathbb{P}\big(\|\widehat{\mathbf{T}} - \mathbf{T}\|_2 > 2C_{10}\sqrt{\frac{d}{T}}\big) \leq 9^d \cdot \mathbb{P}\big(\big|\boldsymbol{v}^\mathsf{T}(\widehat{\mathbf{T}} - \mathbf{T})\boldsymbol{v}\big| > C_{10}\sqrt{\frac{d}{T}}\big).$$

Using (A1.11), we have

$$\mathbb{P}(\|\widehat{\mathbf{T}} - \mathbf{T}\|_2 > 2C_{10}\sqrt{\frac{d}{T}}) \leq 9^d \cdot 3\exp(-C_{18}d).$$

By Lemma A2.8, it follows that

$$\mathbb{P}\Big(\big\|\cos\big(\frac{\pi}{2}\mathbf{T}\big) \circ (\widehat{\mathbf{T}} - \mathbf{T})\big\|_2 \leq 4C_{10}\sqrt{\frac{d}{T}}\Big) > 1 - 9^d \cdot 3\exp(-C_{18}d), \tag{A1.15}$$

where we remind C_{18} is a constant only depending on $C_{10}, \kappa_1, \kappa_2, \gamma_1, \lambda_{\max}(\boldsymbol{\Sigma})$, and $\lambda_{\min}(\boldsymbol{\Sigma})$.

Step I.2. In this step we upper bound the second order term E_2. Due to Lemma A2.9 and the fact that, for any $1 \leq i, j \leq d$,

$$\Big|\big[\sin\big(\frac{\pi}{2}\bar{\mathbf{T}}\big) \circ (\widehat{\mathbf{T}} - \mathbf{T}) \circ (\widehat{\mathbf{T}} - \mathbf{T})\big]_{ij}\Big| \leq \big[(\widehat{\mathbf{T}} - \mathbf{T}) \circ (\widehat{\mathbf{T}} - \mathbf{T})\big]_{ij},$$

we have

$$\Big\|\sin\big(\frac{\pi}{2}\bar{\mathbf{T}}\big) \circ (\widehat{\mathbf{T}} - \mathbf{T}) \circ (\widehat{\mathbf{T}} - \mathbf{T})\Big\|_2 \leq \|(\widehat{\mathbf{T}} - \mathbf{T}) \circ (\widehat{\mathbf{T}} - \mathbf{T})\|_2. \tag{A1.16}$$

Similarly, using Lemma A2.7, we have there exists a positive constant C_{19}, only depending on $\kappa_1, \kappa_2, \gamma_1$, $\lambda_{\max}(\boldsymbol{\Sigma})$, and $\lambda_{\min}(\boldsymbol{\Sigma})$, such that, for $1 \leq i, j \leq d$ and $\eta_2 > 0$, we have

$$\mathbb{P}(|\widehat{\mathbf{T}}_{ij} - \mathbf{T}_{ij}| \geq \eta_2) \leq 3\exp(-C_{19}T\eta_2^2),$$

and, by the union bound, we further obtain

$$\mathbb{P}(\|\widehat{\mathbf{T}} - \mathbf{T}\|_{\max} \geq \eta_2) \leq \frac{3}{2} \cdot d^2 \exp(-C_{19}T\eta_2^2). \tag{A1.17}$$

Reminding the inequality $\|\mathbf{M}\|_2 \leq d \cdot \|\mathbf{M}\|_{\max}$ for any $\mathbf{M} \in \mathbb{R}^{d \times d}$, we deduce

$$\|(\widehat{\mathbf{T}} - \mathbf{T}) \circ (\widehat{\mathbf{T}} - \mathbf{T})\|_2 \leq d \cdot \|(\widehat{\mathbf{T}} - \mathbf{T}) \circ (\widehat{\mathbf{T}} - \mathbf{T})\|_{\max} = d \cdot \|\widehat{\mathbf{T}} - \mathbf{T}\|_{\max}^2. \tag{A1.18}$$

Combining (A1.16), (A1.17), and (A1.18), we conclude that

$$\mathbb{P}\Big(\big\|\sin\big(\frac{\pi}{2}\bar{\mathbf{T}}\big)\circ(\widehat{\mathbf{T}}-\mathbf{T})\circ(\widehat{\mathbf{T}}-\mathbf{T})\big\|_2\le d\eta_2^2\Big)>1-\frac{3}{2}\cdot d^2\exp(-C_{19}T\eta_2^2).$$

Setting $\eta_2^2=2\log(2d/\beta_1)/C_{19}T$ where $\beta_1>0$ is any absolute positive constant, we deduce

$$\mathbb{P}\Big(\big\|\sin\big(\frac{\pi}{2}\bar{\mathbf{T}}\big)\circ(\widehat{\mathbf{T}}-\mathbf{T})\circ(\widehat{\mathbf{T}}-\mathbf{T})\big\|_2\le\frac{2d\log(2d/\beta_1)}{C_{19}T}\Big)>1-\frac{3\beta_1^2}{8}. \tag{A1.19}$$

Combining (A1.4), (A1.15), and (A1.19), we have

$$\mathbb{P}\big(\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_2\le 2\pi C_{10}\sqrt{\frac{d}{T}}+\frac{\pi^2 d\log(2d/\beta_1)}{4C_{19}T}\big)>1-9^d\cdot 3\exp(-C_{18}d)-\frac{3\beta_1^2}{8}.$$

Setting $\beta_1=2d/\exp(\sqrt{T/d})$, we conclude

$$\begin{aligned}&\mathbb{P}\big(\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_2\le(2\pi C_{10}+\frac{\pi^2}{4C_{19}})\cdot\sqrt{\frac{d}{T}}\big)\\ >&1-9^d\cdot 3\exp(-C_{18}d)-\frac{3}{2}\cdot d^2\exp\big(-2\sqrt{\frac{T}{d}}\big),\end{aligned}$$

And it follows that there exist an absolute constant $C_1>0$ and a constant C_1', only depending on $C_1,\kappa_1,\gamma_1,\lambda_{\max}(\mathbf{\Sigma}),\lambda_{\min}(\mathbf{\Sigma})$, such that

$$\mathbb{P}\big(\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_2\le C_1\cdot\sqrt{\frac{d}{T}}\big)>1-\epsilon_1, \tag{A1.20}$$

where $\epsilon_1=9^d\cdot 3\exp(-C_1'd)+3d^2\exp(-2\sqrt{T/d})/2$, and ϵ_1 goes to zero when (T,d) goes to infinity. Therefore (A1.20) verifies (A1.1).

Step II. Now we turn to study (A1.2). Similar to the deviation in the first part, we have

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}_1-\mathbf{\Sigma}_1\|_2\le&\frac{\pi}{2}\cdot\underbrace{\|\cos\big(\frac{\pi}{2}\mathbf{T}_1\big)\circ(\widehat{\mathbf{T}}_1-\mathbf{T}_1)\|_2}_{E_3}+\\ &\frac{\pi^2}{8}\cdot\underbrace{\|\sin\big(\frac{\pi}{2}\bar{\mathbf{T}}_1\big)\circ(\widehat{\mathbf{T}}_1-\mathbf{T}_1)\circ(\widehat{\mathbf{T}}_1-\mathbf{T}_1)\|_2}_{E_4}.\end{aligned} \tag{A1.21}$$

Step II.1. For the first order term, since

$$\widehat{\mathbf{T}}_1=\frac{1}{T-1}\sum_{t=1}^{T-1}\boldsymbol{S}_t\boldsymbol{S}_{t+1}^{\mathsf{T}},$$

we have, for any $\boldsymbol{v}\in\mathbb{S}^{d-1}$,

$$\begin{aligned}\boldsymbol{v}^{\mathsf{T}}(\widehat{\mathbf{T}}_1-\mathbf{T}_1)\boldsymbol{v}&=\frac{1}{T-1}\sum_{t=1}^{T-1}\big(\boldsymbol{v}^{\mathsf{T}}\boldsymbol{S}_t\boldsymbol{S}_{t+1}^{\mathsf{T}}\boldsymbol{v}-\mathbb{E}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{S}_t\boldsymbol{S}_{t+1}^{\mathsf{T}}\boldsymbol{v}\big)\\&=\frac{1}{T-1}\sum_{t=1}^{T-1}\big((\boldsymbol{v}^{\mathsf{T}}\boldsymbol{S}_t)(\boldsymbol{v}^{\mathsf{T}}\boldsymbol{S}_{t+1})-\mathbb{E}(\boldsymbol{v}^{\mathsf{T}}\boldsymbol{S}_t)(\boldsymbol{v}^{\mathsf{T}}\boldsymbol{S}_{t+1})\big).\end{aligned} \tag{A1.22}$$

For $t=1,\ldots,T-1$, we define

$$A_t:=\boldsymbol{v}^{\mathsf{T}}\boldsymbol{S}_t\quad\text{and}\quad B_t:=\boldsymbol{v}^{\mathsf{T}}\boldsymbol{S}_{t+1}.$$

It follows from (A1.22) that

$$\boldsymbol{v}^{\mathsf{T}}(\widehat{\mathbf{T}}_1-\mathbf{T}_1)\boldsymbol{v}=\frac{1}{T-1}\sum_{t=1}^{T-1}(A_tB_t-\mathbb{E}A_tB_t).$$

Due to the fact that

$$A_tB_t=\frac{1}{4}((A_t+B_t)^2-(A_t-B_t)^2),$$

we have

$$\begin{aligned}\boldsymbol{v}^{\mathsf{T}}(\widehat{\mathbf{T}}_1-\mathbf{T}_1)\boldsymbol{v}=&\frac{1}{T-1}\sum_{t=1}^{T-1}\frac{(A_t+B_t)^2-\mathbb{E}(A_t+B_t)^2}{4}+\\&\frac{1}{T-1}\sum_{t=1}^{T-1}\frac{(A_t-B_t)^2-\mathbb{E}(A_t-B_t)^2}{4}.\end{aligned}$$

Thus, we also have

$$\begin{aligned}|\boldsymbol{v}^{\mathsf{T}}(\widehat{\mathbf{T}}_1-\mathbf{T}_1)\boldsymbol{v}|\leq&\Big|\frac{1}{T-1}\sum_{t=1}^{T-1}\frac{(A_t+B_t)^2-\mathbb{E}(A_t+B_t)^2}{4}\Big|+\\&\Big|\frac{1}{T-1}\sum_{t=1}^{T-1}\frac{(A_t-B_t)^2-\mathbb{E}(A_t-B_t)^2}{4}\Big|.\end{aligned}\tag{A1.23}$$

Since $\{\boldsymbol{Z}_t\}_{t=1}^{T}$ satisfies the strong mixing condition, we have $\{(A_t+B_t)^2\}_{t=1}^{T-1}$ and $\{(A_t-B_t)^2\}_{t=1}^{T-1}$ also satisfy the strong mixing condition. At the same time, $\{(A_t+B_t)\}_{t=1}^{T-1}$ and $\{(A_t-B_t)\}_{t=1}^{T-1}$ also satisfy the subguassian condition. According to Lemma A2.7, we have

$$\mathbb{P}\Big(\Big|\frac{1}{T-1}\sum_{t=1}^{T-1}\frac{(A_t+B_t)^2-\mathbb{E}(A_t+B_t)^2}{4}\Big|\geq C_{21}\sqrt{\frac{d}{T-1}}\Big)\leq 3\exp(-C_{22}d),\tag{A1.24}$$

$$\mathbb{P}\Big(\Big|\frac{1}{T-1}\sum_{t=1}^{T-1}\frac{(A_t-B_t)^2-\mathbb{E}(A_t-B_t)^2}{4}\Big|\geq C_{23}\sqrt{\frac{d}{T-1}}\Big)\leq 3\exp(-C_{24}d).\tag{A1.25}$$

where C_{21},C_{23} are absolute constants, C_{22} is a constant only depending on $C_{21},\kappa_1,\kappa_2,\gamma_1,\lambda_{\max}(\boldsymbol{\Sigma})$, $\lambda_{\min}(\boldsymbol{\Sigma})$, and C_{24} is a constant only depending on $C_{23},\kappa_1,\kappa_2,\gamma_1,\lambda_{\max}(\boldsymbol{\Sigma}),\lambda_{\min}(\boldsymbol{\Sigma})$. Combining (A1.23), (A1.24), and (A1.25), we have

$$\mathbb{P}\Big(|\boldsymbol{v}^{\mathsf{T}}(\widehat{\mathbf{T}}_1-\mathbf{T}_1)\boldsymbol{v}|\geq(C_{21}+C_{23})\sqrt{\frac{d}{T-1}}\Big)\leq 3\exp(-C_{22}d)+3\exp(-C_{24}d).$$

According to Lemma A2.5 and Lemma A2.6, it follows that

$$\mathbb{P}(\|\widehat{\mathbf{T}}_1-\mathbf{T}_1\|_2\geq 2(C_{21}+C_{23})\sqrt{\frac{d}{T-1}})\leq 9^d\cdot 3(\exp(-C_{22}d)+\exp(-C_{24}d)).$$

By Lemma A2.8, it follows that

$$\begin{aligned}&\mathbb{P}\Big(\Big\|\cos\big(\frac{\pi}{2}\mathbf{T}_1\big)\circ(\widehat{\mathbf{T}}_1-\mathbf{T}_1)\Big\|_2\leq 4(C_{21}+C_{23})\sqrt{\frac{d}{T-1}}\Big)\\>&1-9^d\cdot 3(\exp(-C_{22}d)+\exp(-C_{24}d)).\end{aligned}\tag{A1.26}$$

Step II.2. For the second order term E_4, due to Lemma A2.9, we have

$$\Big\| \sin\big(\frac{\pi}{2}\bar{\mathbf{T}}_1\big) \circ (\widehat{\mathbf{T}}_1 - \mathbf{T}_1) \circ (\widehat{\mathbf{T}}_1 - \mathbf{T}_1)\Big\|_2 \le \|(\widehat{\mathbf{T}}_1 - \mathbf{T}_1) \circ (\widehat{\mathbf{T}}_1 - \mathbf{T}_1)\|_2.$$

By Lemma A2.7 and the elementary inequality that $\|\mathbf{M}\|_2 \le d \cdot \|\mathbf{M}\|_{\max}$ for any $\mathbf{M} \in \mathbb{R}^{d\times d}$, we have, for any absolute positive constant $\beta_2 > 0$,

$$\begin{aligned}\mathbb{P}\Big(\Big\| \sin\big(\frac{\pi}{2}\bar{\mathbf{T}}_1\big) \circ (\widehat{\mathbf{T}}_1 - \mathbf{T}_1) \circ (\widehat{\mathbf{T}}_1 - \mathbf{T}_1)\Big\|_2 \le \frac{2d\log(2d/\beta_2)}{C_{25}(T-1)}\Big) \\ > 1 - \frac{3\beta_2^2}{8},\end{aligned} \tag{A1.27}$$

where C_{25} is a constant only depending on $\kappa_1, \kappa_2, \gamma_1, \lambda_{\max}(\mathbf{\Sigma})$, and $\lambda_{\min}(\mathbf{\Sigma})$. Combining (A1.21), (A1.26), and (A1.27), we have

$$\begin{aligned}\mathbb{P}\big(\|\widehat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}_1\|_2 \le 2\pi(C_{21} + C_{23}) \cdot \sqrt{\frac{d}{T-1}} + \frac{\pi^2 d\log(2d/\beta_2)}{4C_{25}(T-1)}\big) \\ > 1 - 9^d \cdot 3(\exp(-C_{22}d) + \exp(-C_{24}d)) - \frac{3\beta_2^2}{8}.\end{aligned}$$

Setting $\beta_2 = 2d/\sqrt{(T-1)/d}$, we conclude that

$$\begin{aligned}&\mathbb{P}\Big(\|\widehat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}_1\|_2 \le \big(2\pi(C_{21} + C_{23}) + \frac{\pi^2}{4C_{25}}\big) \cdot \sqrt{\frac{d}{T-1}}\Big) \\ >&1 - 9^d \cdot 3(\exp(-C_{22}d) + \exp(-C_{24}d)) - \frac{3}{2} \cdot d^2 \exp\Big(-2\sqrt{\frac{T-1}{d}}\Big),\end{aligned}$$

And it follows that there exist an absolute positive constant C_2 and a constant C_2', only depending on $C_2, \kappa_1, \kappa_2, \gamma_1, \lambda_{\max}(\mathbf{\Sigma}), \lambda_{\min}(\mathbf{\Sigma})$, such that

$$\mathbb{P}\big(\|\widehat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}_1\|_2 \le C_2 \cdot \sqrt{\frac{d}{T-1}}\big) > 1 - \epsilon_2, \tag{A1.28}$$

where $\epsilon_2 = 9^d \cdot 6\exp(-C_2'd) + 3d^2\exp(-2\sqrt{(T-1)/d})/2$, and ϵ_2 goes to zero when (T, d) goes to infinity. And (A1.28) verifies (A1.2).

In the end, based on the assumptions (3.3), (3.4), and the results (A1.20), (A1.28), we adopt Lemma 2.2 to conclude

$$\mathbb{P}\Big(\|\widehat{\mathbf{A}}_\lambda - \mathbf{A}\|_{\mathsf{F}} \le \frac{1}{2\mu}\big(\lambda\sqrt{r} + 2\sqrt{2}C_1\gamma_{\max}\sqrt{\frac{dr}{T}} + 2\sqrt{2}C_2\sqrt{\frac{dr}{T-1}}\big)\Big) \ge 1 - \epsilon,$$

where

$$\begin{aligned}\epsilon = 9^d \cdot 3\exp(-C_1'd) + 9^d \cdot 6\exp(-C_2'd) + \frac{3}{2} \cdot d^2\exp\Big(-2\sqrt{\frac{T}{d}}\Big) \\ + \frac{3}{2} \cdot d^2\exp\Big(-2\sqrt{\frac{T-1}{d}}\Big),\end{aligned}$$

and ϵ goes to zero when (T, d) goes to infinity. This completes the proof. □

A1.2 The remainder of the proof of Theorem 4.1

Proof. We aim to show

$$D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}} \| \mathbb{P}_{\mathbf{0}}) \leq \gamma^2 rd.$$

For this, we study the term $\log\det \mathbf{V}_{\mathbf{A}}$. We first focus on determining a lower triangular auxiliary matrix $\mathbf{U} \in \mathbb{R}^{Td\times Td}$, such that the product $\mathbf{V}_{\mathbf{A}}\mathbf{U}$ is upper triangular:

$$\mathbf{U} := \begin{bmatrix} \mathbf{I}_d & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{U}_{21} & \mathbf{I}_d & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{U}_{31} & \mathbf{U}_{32} & \mathbf{I}_d & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{U}_{T1} & \mathbf{U}_{T2} & \mathbf{U}_{T3} & \dots & \mathbf{I}_d \end{bmatrix}.$$

If such a $\mathbf{U}$ exists, it is straightforward that $\det(\mathbf{U}) = 1$ and accordingly

$$\det(\mathbf{V}_{\mathbf{A}}) = \det(\mathbf{V}_{\mathbf{A}}\mathbf{U}). \tag{A1.29}$$

We now prove that such a $\mathbf{U}$ indeed exists. For the sake of presentation clearness, with a little abuse of notation, we rewrite $\mathbf{V}_{\mathbf{A}}$ by using $\mathbf{V}_{ij}$, $1 \leq i, j \leq T$. Specifically, let $\mathbf{V}_{ij} = \mathbf{A}^{j-i}$ if $j > i$, $\mathbf{V}_{ij} = (\mathbf{A}^{i-j})^{\mathsf{T}}$ if $i < j$, and $\mathbf{V}_{ij} = \mathbf{I}_d$ if $j = i$. Accordingly, we have

$$\mathbf{V}_{\mathbf{A}} := \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} & \dots & \mathbf{V}_{1T} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \mathbf{V}_{23} & \dots & \mathbf{V}_{2T} \\ \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{V}_{33} & \dots & \mathbf{V}_{3T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{V}_{T1} & \mathbf{V}_{T2} & \mathbf{V}_{T3} & \dots & \mathbf{V}_{TT} \end{bmatrix}$$

and $\mathbf{V}_{\mathbf{A}}\mathbf{U}$ is equal to

$$\begin{bmatrix} \mathbf{V}_{11}+\mathbf{V}_{12}\mathbf{U}_{21}+\ldots+\mathbf{V}_{1T}\mathbf{U}_{T1} & \mathbf{V}_{12}+\mathbf{V}_{13}\mathbf{U}_{32}+\ldots+\mathbf{V}_{1T}\mathbf{U}_{T2} & \dots & \mathbf{V}_{1T} \\ \mathbf{V}_{21}+\mathbf{V}_{22}\mathbf{U}_{21}+\ldots+\mathbf{V}_{2T}\mathbf{U}_{T1} & \mathbf{V}_{22}+\mathbf{V}_{23}\mathbf{U}_{32}+\ldots+\mathbf{V}_{2T}\mathbf{U}_{T2} & \dots & \mathbf{V}_{2T} \\ \mathbf{V}_{31}+\mathbf{V}_{32}\mathbf{U}_{21}+\ldots+\mathbf{V}_{3T}\mathbf{U}_{T1} & \mathbf{V}_{32}+\mathbf{V}_{33}\mathbf{U}_{32}+\ldots+\mathbf{V}_{3T}\mathbf{U}_{T2} & \dots & \mathbf{V}_{3T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{V}_{T1}+\mathbf{V}_{T2}\mathbf{U}_{21}+\ldots+\mathbf{V}_{TT}\mathbf{U}_{T1} & \mathbf{V}_{T2}+\mathbf{V}_{T3}\mathbf{U}_{32}+\ldots+\mathbf{V}_{TT}\mathbf{U}_{T2} & \dots & \mathbf{V}_{TT} \end{bmatrix}. \tag{A1.30}$$

Now let us determine the entries of $\mathbf{U}$ such that the product $\mathbf{V}_{\mathbf{A}}\mathbf{U}$ is indeed upper triangular. In this case, we obtain $(T^2 - T)/2$ constraint equations. And the $T - k$ constraint equations arising from the k-th column of $\mathbf{V}_{\mathbf{A}}\mathbf{U}$ are

$$\begin{aligned} \mathbf{V}_{k+1,k} + \mathbf{V}_{k+1,k+1}\mathbf{U}_{k+1,k} + \ldots + \mathbf{V}_{k+1,T}\mathbf{U}_{Tk} &= \mathbf{0}, \\ \mathbf{V}_{k+2,k} + \mathbf{V}_{k+2,k+1}\mathbf{U}_{k+1,k} + \ldots + \mathbf{V}_{k+2,T}\mathbf{U}_{Tk} &= \mathbf{0}, \\ &\vdots \\ \mathbf{V}_{T,k} + \mathbf{V}_{T,k+1}\mathbf{U}_{k+1,k} + \ldots + \mathbf{V}_{T,T}\mathbf{U}_{Tk} &= \mathbf{0}. \end{aligned} \tag{A1.31}$$

For further calculation, we now define

$$\mathbf{V}^c_{ij} := [\mathbf{V}^\mathsf{T}_{ij},\ \mathbf{V}^\mathsf{T}_{i+1,j},\ \ \ldots,\ \ \mathbf{V}^\mathsf{T}_{Tj}]^\mathsf{T}, \quad \text{for } 1 \le i,j \le T,$$
$$\mathbf{U}^c_{ij} := [\mathbf{U}^\mathsf{T}_{ij},\ \mathbf{U}^\mathsf{T}_{i+1,j},\ \ \ldots,\ \ \mathbf{U}^\mathsf{T}_{Tj}]^\mathsf{T}, \quad \text{for } 1 \le j < i \le T.$$

And we similarly define

$$\mathbf{V}^r_{ij} := [\mathbf{V}_{ij},\ \mathbf{V}_{i,j+1},\ \ \ldots,\ \ \mathbf{V}_{iT}], \quad \text{for } 1 \le i,j \le T.$$

We then let $\widetilde{\mathbf{V}}_k$ represent the k by k block matrix formed from the lower right corner of $\mathbf{V}_\mathbf{A}$ for $k = 1, \ldots, T$:

$$\widetilde{\mathbf{V}}_k := \begin{bmatrix} \mathbf{V}_{T-k+1,T-k+1} & \mathbf{V}_{T-k+1,T-k+2} & \cdots & \mathbf{V}_{T-k+1,T} \\ \mathbf{V}_{T-k+2,T-k+1} & \mathbf{V}_{T-k+2,T-k+2} & \cdots & \mathbf{V}_{T-k+2,T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{V}_{T,T-k+1} & \mathbf{V}_{T,T-k+2} & \cdots & \mathbf{V}_{T,T} \end{bmatrix}.$$

With these definitions, we can rewrite (A1.31) as

$$\mathbf{V}^c_{k+1,k} + \widetilde{\mathbf{V}}_{T-k}\mathbf{U}^c_{k+1,k} = \mathbf{0}, \quad \text{for } k = 1, \ldots, T-1,$$

and we have

$$\mathbf{U}^c_{k+1,k} = -\widetilde{\mathbf{V}}^{-1}_{T-k}\mathbf{V}^c_{k+1,k}, \tag{A1.32}$$

which determines the lower triangular auxiliary matrix $\mathbf{U}$ such that $\mathbf{V}_\mathbf{A}\mathbf{U}$ is upper triangular. Of note, $\widetilde{\mathbf{V}}_{T-k}$ is invertible since $\mathbf{V}_\mathbf{A}$ is positive definite.

Secondly, we study $\mathbf{V}_\mathbf{A}\mathbf{U}$ in more details. From (A1.30), we can express the k-th diagonal element of $\mathbf{V}_\mathbf{A}\mathbf{U}$, for $k = 1, \ldots, T$, in the form:

$$(\mathbf{V}_\mathbf{A}\mathbf{U})_{kk} = \mathbf{V}_{kk} + \mathbf{V}_{k,k+1}\mathbf{U}_{k+1,k} + \ldots + \mathbf{V}_{kT}\mathbf{U}_{Tk} = \mathbf{V}_{kk} + \mathbf{V}^r_{k,k+1}\mathbf{U}^c_{k+1,k}. \tag{A1.33}$$

Combining (A1.32) and (A1.33) we have

$$(\mathbf{V}_\mathbf{A}\mathbf{U})_{kk} = \mathbf{V}_{kk} - \mathbf{V}^r_{k,k+1}\widetilde{\mathbf{V}}^{-1}_{T-k}\mathbf{V}^c_{k+1,k}.$$

For $k = 0, 1, \ldots, T$ and $1 \le i,j \le T$, define $\mathbf{V}^{(k)}_{ij}$ as follows:

$$\begin{aligned} \mathbf{V}^{(0)}_{ij} &= \mathbf{V}_{ij}, \\ \mathbf{V}^{(k)}_{ij} &= \mathbf{V}_{ij} - \mathbf{V}^r_{i,T-k+1}\widetilde{\mathbf{V}}^{-1}_k\mathbf{V}^c_{T-k+1,j}, \quad \text{for } k = 1, \ldots, T. \end{aligned} \tag{A1.34}$$

Then we have

$$(\mathbf{V}_\mathbf{A}\mathbf{U})_{kk} = \mathbf{V}^{(T-k)}_{kk}.$$

Noting that $\mathbf{V}_\mathbf{A}\mathbf{U}$ is upper triangular, we obtain

$$\det(\mathbf{V}_\mathbf{A}\mathbf{U}) = \prod_{k=1}^{T} \det\left(\mathbf{V}^{(T-k)}_{kk}\right). \tag{A1.35}$$

Combining (A1.29) and (A1.35), we have

$$\det(\mathbf{V}_\mathbf{A}) = \prod_{k=1}^{T} \det\left(\mathbf{V}^{(T-k)}_{kk}\right). \tag{A1.36}$$

We then aim to solve the expression of $\mathbf{V}_{kk}^{(T-k)}$ for $k = 1, \ldots, T$. Now we focus on $\mathbf{V}_{ij}^{(k)}$ for $k = 0, 1, \ldots, T$ and $1 \leq i, j \leq T$. It is sufficient to establish the recursive relationship between matrices for difference choices of k. Let us start with

$$\begin{aligned}
\mathbf{V}_{ij}^{(k+1)} &= \mathbf{V}_{ij} - \mathbf{V}_{i,T-k}^{r} \widetilde{\mathbf{V}}_{k+1}^{-1} \mathbf{V}_{T-k,j}^{c} \\
&= \mathbf{V}_{ij} - \left[\begin{array}{cc} \mathbf{V}_{i,T-k} & \mathbf{V}_{i,T-k+1}^{r} \end{array} \right] \left[\begin{array}{cc} \mathbf{V}_{T-k,T-k} & \mathbf{V}_{T-k,T-k+1}^{r} \\ \mathbf{V}_{T-k+1,T-k}^{c} & \widetilde{\mathbf{V}}_{k} \end{array} \right]^{-1} \left[\begin{array}{c} \mathbf{V}_{T-k,j} \\ \mathbf{V}_{T-k+1,j}^{c} \end{array} \right].
\end{aligned}$$

Noting that

$$\mathbf{V}_{T-k,T-k} - \mathbf{V}_{T-k,T-k+1}^{r} \widetilde{\mathbf{V}}_{k}^{-1} \mathbf{V}_{T-k+1,T-k}^{c} = \mathbf{V}_{T-k,T-k}^{(k)},$$

and using the Banachiewic identity (Brualdi and Schneider, 1983),

$$\begin{aligned}
&\left[\begin{array}{cc} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{array} \right]^{-1} \\
&= \left[\begin{array}{cc} (\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21})^{-1} & -(\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21})^{-1}\mathbf{M}_{12}\mathbf{M}_{22}^{-1} \\ -\mathbf{M}_{22}^{-1}\mathbf{M}_{21}(\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21})^{-1} & \mathbf{M}_{22}^{-1}[\mathbf{I} + \mathbf{M}_{21}(\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21})^{-1}\mathbf{M}_{12}\mathbf{M}_{22}^{-1}] \end{array} \right],
\end{aligned}$$

we deduce that

$$\begin{aligned}
&\mathbf{V}_{ij}^{(k+1)} = \mathbf{V}_{ij} - \left[\begin{array}{cc} \mathbf{V}_{i,T-k} & \mathbf{V}_{i,T-k+1}^{r} \end{array} \right] \cdot \\
&\left[\begin{array}{cc} (\mathbf{V}_{T-k,T-k}^{(k)})^{-1} & -(\mathbf{V}_{T-k,T-k}^{(k)})^{-1}\mathbf{V}_{T-k,T-k+1}^{r}\widetilde{\mathbf{V}}_{k}^{-1} \\ -\widetilde{\mathbf{V}}_{k}^{-1}\mathbf{V}_{T-k+1,T-k}^{c}(\mathbf{V}_{T-k,T-k}^{(k)})^{-1} & \widetilde{\mathbf{V}}_{k}^{-1}[\mathbf{I} + \mathbf{V}_{T-k+1,T-k}^{c}(\mathbf{V}_{T-k,T-k}^{(k)})^{-1}\mathbf{V}_{T-k,T-k+1}^{r}\widetilde{\mathbf{V}}_{k}^{-1}] \end{array} \right] \\
&\cdot \left[\begin{array}{c} \mathbf{V}_{T-k,j} \\ \mathbf{V}_{T-k+1,j}^{c} \end{array} \right].
\end{aligned}$$

Multiplying the last two matrices implies

$$\begin{aligned}
&\mathbf{V}_{ij}^{(k+1)} = \mathbf{V}_{ij} - \left[\begin{array}{cc} \mathbf{V}_{i,T-k} & \mathbf{V}_{i,T-k+1}^{r} \end{array} \right] \\
&\cdot \left[\begin{array}{c} (\mathbf{V}_{T-k,T-k}^{(k)})^{-1}(\mathbf{V}_{T-k,j} - \mathbf{V}_{T-k,T-k+1}^{r}\widetilde{\mathbf{V}}_{k}^{-1}\mathbf{V}_{T-k+1,j}^{c}) \\ \widetilde{\mathbf{V}}_{k}^{-1}\mathbf{V}_{T-k+1,j}^{c} - \widetilde{\mathbf{V}}_{k}^{-1}\mathbf{V}_{T-k+1,T-k}^{c}(\mathbf{V}_{T-k,T-k}^{(k)})^{-1}(\mathbf{V}_{T-k,j} - \mathbf{V}_{T-k,T-k+1}^{r}\widetilde{\mathbf{V}}_{k}^{-1}\mathbf{V}_{T-k+1,j}^{c}) \end{array} \right].
\end{aligned}$$

By the fact that

$$\mathbf{V}_{T-k,j} - \mathbf{V}_{T-k,T-k+1}^{r}\widetilde{\mathbf{V}}_{k}^{-1}\mathbf{V}_{T-k+1,j}^{c} = \mathbf{V}_{T-k,j}^{(k)},$$

we have

$$\begin{aligned}
\mathbf{V}_{ij}^{(k+1)} =& \mathbf{V}_{ij} - \left[\begin{array}{cc} \mathbf{V}_{i,T-k} & \mathbf{V}_{i,T-k+1}^{r} \end{array} \right] \\
&\cdot \left[\begin{array}{c} (\mathbf{V}_{T-k,T-k}^{(k)})^{-1}\mathbf{V}_{T-k,j}^{(k)} \\ \widetilde{\mathbf{V}}_{k}^{-1}\mathbf{V}_{T-k+1,j}^{c} - \widetilde{\mathbf{V}}_{k}^{-1}\mathbf{V}_{T-k+1,T-k}^{c}(\mathbf{V}_{T-k,T-k}^{(k)})^{-1}\mathbf{V}_{T-k,j}^{(k)} \end{array} \right].
\end{aligned}$$

Multiplying the last two matrices further implies

$$\begin{aligned}
\mathbf{V}_{ij}^{(k+1)} =& \mathbf{V}_{ij} - \mathbf{V}_{i,T-k}(\mathbf{V}_{T-k,T-k}^{(k)})^{-1}\mathbf{V}_{T-k,j}^{(k)} - \mathbf{V}_{i,T-k+1}^{r}\widetilde{\mathbf{V}}_{k}^{-1}\mathbf{V}_{T-k+1,j}^{c} \\
&+ \mathbf{V}_{i,T-k+1}^{r}\widetilde{\mathbf{V}}_{k}^{-1}\mathbf{V}_{T-k+1,T-k}^{c}(\mathbf{V}_{T-k,T-k}^{(k)})^{-1}\mathbf{V}_{T-k,j}^{(k)}.
\end{aligned}$$

Reorganizing the above equation implies

$$\begin{aligned}\mathbf{V}_{ij}^{(k+1)} =&(\mathbf{V}_{ij} - \mathbf{V}^r_{i,T-k+1}\widetilde{\mathbf{V}}_k^{-1}\mathbf{V}^c_{T-k+1,j})\\ &- (\mathbf{V}_{i,T-k} - \mathbf{V}^r_{i,T-k+1}\widetilde{\mathbf{V}}_k^{-1}\mathbf{V}^c_{T-k+1,T-k})(\mathbf{V}^{(k)}_{T-k,T-k})^{-1}\mathbf{V}^{(k)}_{T-k,j}.\end{aligned}$$

Noting that

$$\mathbf{V}_{ij} - \mathbf{V}^r_{i,T-k+1}\widetilde{\mathbf{V}}_k^{-1}\mathbf{V}^c_{T-k+1,j} = \mathbf{V}^{(k)}_{ij}$$

and

$$\mathbf{V}_{i,T-k} - \mathbf{V}^r_{i,T-k+1}\widetilde{\mathbf{V}}_k^{-1}\mathbf{V}^c_{T-k+1,T-k} = \mathbf{V}^{(k)}_{i,T-k},$$

we further have

$$\mathbf{V}_{ij}^{(k+1)} = \mathbf{V}_{ij}^{(k)} - \mathbf{V}^{(k)}_{i,T-k}(\mathbf{V}^{(k)}_{T-k,T-k})^{-1}\mathbf{V}^{(k)}_{T-k,j}, \tag{A1.37}$$

which gives us a recursive relationship between matrices of consecutive values of k. Now we calculate the exact value of $\mathbf{V}_{ij}^{(k)}$ for $1 \leq i,j \leq T-k$ and $k = 0, 1, \ldots, T-1$ by (4.4) and the recursive relationship in (A1.37). Here we employ the induction strategy to solve the problem. By the definition in (A1.34), we have, for $k = 0$ and $1 \leq i, j \leq T$,

$$\mathbf{V}_{ij}^{(0)} = \begin{cases} \mathbf{I}_d, & \text{if } i = j,\\ \mathbf{A}^{j-i}, & \text{if } i < j,\\ (\mathbf{A}^\top)^{i-j}, & \text{if } i > j,\end{cases}$$

and for $k = 1,\ 1 \leq i, j \leq T-1$,

$$\mathbf{V}_{ij}^{(1)} = \begin{cases} \mathbf{I}_d - \mathbf{A}^{T-i}(\mathbf{A}^\top)^{T-j}, & \text{if } i = j,\\ \mathbf{A}^{j-i}\left[\mathbf{I}_d - \mathbf{A}^{T-j}(\mathbf{A}^\top)^{T-j}\right], & \text{if } i < j,\\ \left[\mathbf{I}_d - \mathbf{A}^{T-i}(\mathbf{A}^\top)^{T-i}\right](\mathbf{A}^\top)^{i-j}, & \text{if } i > j.\end{cases}$$

And then we aim to prove, for $k = 1, \ldots, T-1,\ 1 \leq i, j \leq T-k$,

$$\mathbf{V}_{ij}^{(k)} = \begin{cases} \mathbf{I}_d - \mathbf{A}^{T+1-k-i}(\mathbf{A}^\top)^{T+1-k-j}, & \text{if } i = j,\\ \mathbf{A}^{j-i}\left[\mathbf{I}_d - \mathbf{A}^{T+1-k-j}(\mathbf{A}^\top)^{T+1-k-j}\right], & \text{if } i < j,\\ \left[\mathbf{I}_d - \mathbf{A}^{T+1-k-i}(\mathbf{A}^\top)^{T+1-k-i}\right](\mathbf{A}^\top)^{i-j}, & \text{if } i > j.\end{cases} \tag{A1.38}$$

Let us assume (A1.38) holds for $k = 1, \ldots, n$ and consider $k = n+1$. Then,

- if $i = j$, we have

$$\begin{aligned}&\mathbf{V}_{ij}^{(n+1)}\\ =&\mathbf{V}_{ij}^{(n)} - \mathbf{V}^{(n)}_{i,T-n}(\mathbf{V}^{(n)}_{T-n,T-n})^{-1}\mathbf{V}^{(n)}_{T-n,j}\\ =&\mathbf{I}_d - \mathbf{A}^{T+1-n-i}(\mathbf{A}^\top)^{T+1-n-j} - \mathbf{A}^{T-n-i}(\mathbf{I}_d - \mathbf{A}\mathbf{A}^\top)(\mathbf{A}^\top)^{T-n-j}\\ =&\mathbf{I}_d - \mathbf{A}^{T-n-i}(\mathbf{A}^\top)^{T-n-j};\end{aligned}$$

- if $i < j$, we have

$$\begin{aligned}
&\mathbf{V}_{ij}^{(n+1)} \\
=&\mathbf{A}^{j-i}\left[\mathbf{I}_d - \mathbf{A}^{T+1-n-j}(\mathbf{A}^\mathsf{T})^{T+1-n-j}\right] - \mathbf{A}^{T-n-i}(\mathbf{I}_d - \mathbf{A}\mathbf{A}^\mathsf{T})(\mathbf{A}^\mathsf{T})^{T-n-j} \\
=&\mathbf{A}^{j-i}\left[\mathbf{I}_d - \mathbf{A}^{T-n-j}(\mathbf{A}^\mathsf{T})^{T-n-j}\right];
\end{aligned}$$

- if $i > j$, we have

$$\begin{aligned}
&\mathbf{V}_{ij}^{(n+1)} \\
=&\left[\mathbf{I}_d - \mathbf{A}^{T+1-n-i}(\mathbf{A}^\mathsf{T})^{T+1-n-i}\right](\mathbf{A}^\mathsf{T})^{i-j} - \mathbf{A}^{T-n-i}(\mathbf{I}_d - \mathbf{A}\mathbf{A}^\mathsf{T})(\mathbf{A}^\mathsf{T})^{T-n-j} \\
=&\left[\mathbf{I}_d - \mathbf{A}^{T-n-i}(\mathbf{A}^\mathsf{T})^{T-n-i}\right](\mathbf{A}^\mathsf{T})^{i-j}.
\end{aligned}$$

Therefore, we prove (A1.38) holds for $k = 1, \ldots, T-1$ and $1 \le i, j \le T-k$. This implies, for $k = 1, \ldots, T-1$,

$$\mathbf{V}_{kk}^{(T-k)} = \mathbf{I}_d - \mathbf{A}\mathbf{A}^\mathsf{T}. \tag{A1.39}$$

Thus, combining (A1.36), (A1.39), and the fact that $\mathbf{V}_{TT}^{(0)} = \mathbf{I}_d$, we have

$$-\frac{1}{2}\log\det\mathbf{V}_\mathbf{A} = -\frac{T-1}{2}\log\det(\mathbf{I}_d - \mathbf{A}\mathbf{A}^\mathsf{T}). \tag{A1.40}$$

We note that $\mathbf{A} \in \mathcal{A}^0$ and accordingly $\mathbf{A}$ can be written as $\mathbf{A} = (\bar{\mathbf{A}} \mid \ldots \mid \bar{\mathbf{A}} \mid \mathbf{0})$ where $\bar{\mathbf{A}} \in \mathbb{R}^{d\times r}$ is a d by r matrix in the form of (4.2). We then have

$$\det(\mathbf{I}_d - \mathbf{A}\mathbf{A}^\mathsf{T}) = \det\left(\mathbf{I}_d - \left\lfloor\frac{d}{r}\right\rfloor \cdot \bar{\mathbf{A}}\bar{\mathbf{A}}^\mathsf{T}\right).$$

Let the singular values of $\bar{\mathbf{A}}$ be $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_d$ with $\sigma_{r+1} = \cdots = \sigma_d = 0$. By definition, the eigenvalues of $\bar{\mathbf{A}}\bar{\mathbf{A}}^\mathsf{T}$ are $\sigma_1^2, \ldots, \sigma_d^2$, and we have

$$\det\left(\mathbf{I}_d - \left\lfloor\frac{d}{r}\right\rfloor \cdot \bar{\mathbf{A}}\bar{\mathbf{A}}^\mathsf{T}\right) = \prod_{j=1}^{d}\left(1 - \left\lfloor\frac{d}{r}\right\rfloor\sigma_j^2\right) \ge \prod_{j=1}^{d}\left(1 - \frac{d\sigma_j^2}{r}\right).$$

Using the fact that

$$\sum_{j=1}^{d}\sigma_i^2 = \|\bar{\mathbf{A}}\|_\mathsf{F}^2 = \mathrm{Tr}(\bar{\mathbf{A}}\bar{\mathbf{A}}^\mathsf{T}) \le \gamma^2\frac{r}{dT}dr = \gamma^2\frac{r^2}{T},$$

and, for $2 \le k \le d$,

$$\left(1 - \frac{d}{r}(\sigma_1^2 + \ldots + \sigma_{k-1}^2)\right)\left(1 - \frac{d}{r}\sigma_k^2\right) \ge 1 - \frac{d}{r}(\sigma_1^2 + \ldots + \sigma_k^2),$$

we have

$$\det\left(\mathbf{I}_d - \left\lfloor\frac{d}{r}\right\rfloor \cdot \bar{\mathbf{A}}\bar{\mathbf{A}}^\mathsf{T}\right) \ge 1 - \frac{d}{r}\sum_{i=1}^{d}\sigma_i^2 \ge 1 - \gamma^2\frac{rd}{T}.$$

Combining (4.5) and (A1.40), we conclude

$$D_\mathsf{KL}(\mathbb{P}_\mathbf{A}\|\mathbb{P}_\mathbf{0}) \le -\frac{T-1}{2}\log(1 - \gamma^2\frac{rd}{T}).$$

Under the condition that $rd/T = o(1)$, for T sufficiently large, we have

$$-\log(1-\gamma^2\frac{rd}{T}) \le 2\gamma^2\frac{rd}{T},$$

yielding

$$D_{\mathsf{KL}}(\mathbb{P}_{\mathbf{A}}\|\mathbb{P}_{\mathbf{0}}) \le \frac{T-1}{T}\gamma^2 rd \le \gamma^2 rd.$$

This completes the proof. □

A2 Proofs of the rest results

A2.1 Proof of Lemma 2.2

Proof. The proof is very straightforward given the literature (see, for example, Fan et al. (2014)). However, the general version we presented does not exist, and is arguably worth a separate proof. We begin by introducing some additional notation. First, for an arbitrary convex function $f(\mathbf{M}) : \mathbb{R}^{m\times n} \mapsto \mathbb{R}$, we define $\partial f(\mathbf{M})$ to be its subdifferential:

$$\partial f(\mathbf{M}) := \Big\{\mathbf{G}\in\mathbb{R}^{m\times n} : f(\mathbf{N}) \ge f(\mathbf{M}) + \langle \mathbf{N}-\mathbf{M}, \mathbf{G}\rangle, \forall\ \mathbf{N}\in\mathbb{R}^{m\times n}\Big\}.$$

Secondly, for arbitrary matrix $\mathbf{Q}\in\mathbb{R}^{m\times n}$ of rank $\widetilde{r}$ and with spectral representation

$$\mathbf{Q} = \sum_{j=1}^{\widetilde{r}} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^{\mathsf{T}},$$

we define the support of $\mathbf{Q}$ as the pair of linear vector subspaces $(\mathcal{S}_1, \mathcal{S}_2)$. Here $\mathcal{S}_1$ is the linear span of $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{\widetilde{r}}$, and $\mathcal{S}_2$ is the linear span of $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{\widetilde{r}}$. We denote by $\mathcal{S}_j^{\perp}$ the orthogonal complement of $\mathcal{S}_j$ for $j = 1, 2$. We denote by $\mathbf{P}_{\mathcal{S}}$ the projection matrix on the linear vector subspace $\mathcal{S}$ of $\mathbb{R}^d$.

Due to the fact that $L_\lambda(\mathbf{Q}; \widehat{\mathbf{M}}, \widehat{\mathbf{M}}_1)$, defined in (2.2), is a convex function of $\mathbf{Q}$, we can write its subdifferential as:

$$\partial L_\lambda(\mathbf{Q}; \widehat{\mathbf{M}}, \widehat{\mathbf{M}}_1) = \Big\{ -2\widehat{\mathbf{M}}_1^{\mathsf{T}} + 2\mathbf{Q}\widehat{\mathbf{M}} + \lambda\mathbf{V} : \mathbf{V}\in\partial\|\mathbf{Q}\|_*\Big\}, \tag{A2.1}$$

and accordingly

$$\partial L_\lambda(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}; \widehat{\mathbf{M}}, \widehat{\mathbf{M}}_1) = \Big\{ -2\widehat{\mathbf{M}}_1^{\mathsf{T}} + 2\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\widehat{\mathbf{M}} + \lambda\widehat{\mathbf{V}} : \widehat{\mathbf{V}}\in\partial\|\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\|_*\Big\}. \tag{A2.2}$$

Since $\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}$ is the minimizer of $L_\lambda(\mathbf{Q}; \widehat{\mathbf{M}}, \widehat{\mathbf{M}}_1)$, we have

$$\mathbf{0}\in\partial L_\lambda(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}; \widehat{\mathbf{M}}, \widehat{\mathbf{M}}_1).$$

In particular, there exists $\widehat{\mathbf{V}}\in\partial\|\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\|_*$ such that, for all $\mathbf{Q}\in\mathbb{R}^{m\times n}$, we have

$$\langle -2\widehat{\mathbf{M}}_1^{\mathsf{T}} + 2\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\widehat{\mathbf{M}} + \lambda\widehat{\mathbf{V}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle = 0.$$

Using the bilinear property of inner product yields

$$\langle -2(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}), \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle + \langle 2(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle + \lambda\langle\widehat{\mathbf{V}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle = 0.$$

By subtracting $\lambda\langle \mathbf{V}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle$ on each side of the above inequality, we have

$$\langle -2(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}), \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle + \langle 2(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle + \lambda\langle \widehat{\mathbf{V}} - \mathbf{V}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle \qquad \text{(A2.3)}$$
$$= -\lambda\langle \mathbf{V}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle.$$

Noticing that $\widehat{\mathbf{V}} \in \partial\|\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\|_*$ and $\mathbf{V} \in \partial\|\mathbf{Q}\|_*$, we have

$$\|\mathbf{Q}\|_* \geq \|\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\|_* + \langle \mathbf{Q} - \widehat{\mathbf{A}}_\lambda^{\mathrm{G}}, \widehat{\mathbf{V}}\rangle \quad \text{and} \quad \|\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\|_* \geq \|\mathbf{Q}\|_* + \langle \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}, \mathbf{V}\rangle.$$

Then adding the above two inequalities implies

$$\langle \widehat{\mathbf{V}} - \mathbf{V}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle \geq 0. \qquad \text{(A2.4)}$$

It follows from (A2.3) and (A2.4) that

$$-2\langle \widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle + 2\langle (\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle \leq -\lambda\langle \mathbf{V}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle.$$

Moving the first part of the left-hand side to the right-hand side yields

$$2\langle (\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle \leq -\lambda\langle \mathbf{V}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle + 2\langle \widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle.$$

Using the representation introduced in Watson (1992), any element in $\partial\|\mathbf{Q}\|_*$ can be written as:

$$\mathbf{V} = \sum_{j=1}^{r} \boldsymbol{u}_j \boldsymbol{v}_j^{\mathsf{T}} + \mathbf{P}_{\mathcal{S}_1^\perp}\mathbf{W}\mathbf{P}_{\mathcal{S}_2^\perp} \text{ and } \|\mathbf{W}\|_2 \leq 1.$$

We further have

$$2\langle (\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle + \lambda\langle \mathbf{P}_{\mathcal{S}_1^\perp}\mathbf{W}\mathbf{P}_{\mathcal{S}_2^\perp}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle \qquad \text{(A2.5)}$$
$$\leq -\lambda\langle \sum_{j=1}^{r} \boldsymbol{u}_j \boldsymbol{v}_j^{\mathsf{T}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle + 2\langle \widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle.$$

This gives the inequality we will focus on in the rest of the proof.

We then aim to introduce $\|\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A}\|_{\mathsf{F}}$ in bounding the left and right terms in (A2.5). For this, we begin with the second term on the left. Noting that $\mathbf{Q}$ has the support $(\mathcal{S}_1, \mathcal{S}_2)$, we have

$$\langle \mathbf{P}_{\mathcal{S}_1^\perp}\mathbf{W}\mathbf{P}_{\mathcal{S}_2^\perp}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle = \langle \mathbf{P}_{\mathcal{S}_1^\perp}\mathbf{W}\mathbf{P}_{\mathcal{S}_2^\perp}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\rangle.$$

Since both $\mathbf{P}_{\mathcal{S}_1^\perp}$ and $\mathbf{P}_{\mathcal{S}_2^\perp}$ are projection matrices, we further have

$$\langle \mathbf{P}_{\mathcal{S}_1^\perp}\mathbf{W}\mathbf{P}_{\mathcal{S}_2^\perp}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\rangle = \langle \mathbf{W}, \mathbf{P}_{\mathcal{S}_1^\perp}\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\mathbf{P}_{\mathcal{S}_2^\perp}\rangle.$$

By Lemma A2.3, there exists $\mathbf{W}_0$ with $\|\mathbf{W}_0\|_2 \leq 1$ such that

$$\langle \mathbf{W}_0, \mathbf{P}_{\mathcal{S}_1^\perp}\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\mathbf{P}_{\mathcal{S}_2^\perp}\rangle = \|\mathbf{P}_{\mathcal{S}_1^\perp}\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\mathbf{P}_{\mathcal{S}_2^\perp}\|_*.$$

Thus, for this particular choice $\mathbf{W}_0$, we deduce from (A2.5) that

$$\langle 2(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle + \lambda\|\mathbf{P}_{\mathcal{S}_1^\perp}\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\mathbf{P}_{\mathcal{S}_2^\perp}\|_* \qquad \text{(A2.6)}$$
$$\leq \underbrace{-\lambda\Big\langle \sum_{j=1}^{r} \boldsymbol{u}_j \boldsymbol{v}_j^{\mathsf{T}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\Big\rangle}_{A_1} + \underbrace{2\langle \widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle}_{A_2}.$$

Then, we turn to bound the right-hand side of (A2.6). By the fact that

$$\|\sum_{j=1}^{r}\boldsymbol{u}_j\boldsymbol{v}_j^{\mathsf{T}}\|_2 = 1 \text{ and } \langle\sum_{j=1}^{r}\boldsymbol{u}_j\boldsymbol{v}_j^{\mathsf{T}}, \widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q}\rangle = \langle\sum_{j=1}^{r}\boldsymbol{u}_j\boldsymbol{v}_j^{\mathsf{T}}, \mathbf{P}_{\mathcal{S}_1}(\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q})\mathbf{P}_{\mathcal{S}_2}\rangle,$$

using Lemma A2.3, we deduce

$$A_1 \le \lambda\|\sum_{j=1}^{r}\boldsymbol{u}_j\boldsymbol{v}_j^{\mathsf{T}}\|_2 \cdot \|\mathbf{P}_{\mathcal{S}_1}(\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q})\mathbf{P}_{\mathcal{S}_2}\|_* = \lambda\|\mathbf{P}_{\mathcal{S}_1}(\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q})\mathbf{P}_{\mathcal{S}_2}\|_*. \tag{A2.7}$$

By Cauchy-Schwarz inequality, we have

$$\|\mathbf{P}_{\mathcal{S}_1}(\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q})\mathbf{P}_{\mathcal{S}_2}\|_* \le \sqrt{\operatorname{rank}(\mathbf{Q})} \cdot \|\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q}\|_{\mathsf{F}}. \tag{A2.8}$$

Combining (A2.7) and (A2.8) yields

$$A_1 \le \lambda\sqrt{\operatorname{rank}(\mathbf{Q})} \cdot \|\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q}\|_{\mathsf{F}}. \tag{A2.9}$$

To provide an upper bound for A_2, we define a projection operator $\mathcal{P}_{\mathbf{Q}}(\mathbf{M}) := \mathbf{M} - \mathbf{P}_{\mathcal{S}_1^{\perp}}\mathbf{M}\mathbf{P}_{\mathcal{S}_2^{\perp}}$. We then have

$$A_2 = \underbrace{2\langle\mathcal{P}_{\mathbf{Q}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}), \widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q}\rangle}_{A_{21}} + \underbrace{2\langle\mathbf{P}_{\mathcal{S}_1^{\perp}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}})\mathbf{P}_{\mathcal{S}_2^{\perp}}, \widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q}\rangle}_{A_{22}}. \tag{A2.10}$$

For A_{21}, define $\Lambda := 2\|\mathcal{P}_{\mathbf{Q}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}})\|_{\mathsf{F}}$. By Cauchy-Schwarz inequality, we then have

$$A_{21} = 2\langle\mathcal{P}_{\mathbf{Q}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}), \widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q}\rangle \le \Lambda\|\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q}\|_{\mathsf{F}}. \tag{A2.11}$$

For A_{22}, define $\Gamma := 2\|\mathbf{P}_{\mathcal{S}_1^{\perp}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}})\mathbf{P}_{\mathcal{S}_2^{\perp}}\|_2$. By Lemma A2.3, we have

$$A_{22} = 2\langle\mathbf{P}_{\mathcal{S}_1^{\perp}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}})\mathbf{P}_{\mathcal{S}_2^{\perp}}, \widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda}\rangle \le \Gamma\|\mathbf{P}_{\mathcal{S}_1^{\perp}}\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda}\mathbf{P}_{\mathcal{S}_2^{\perp}}\|_*. \tag{A2.12}$$

Combining (A2.10), (A2.11), and (A2.12), we have

$$A_2 \le \Lambda\|\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q}\|_{\mathsf{F}} + \Gamma\|\mathbf{P}_{\mathcal{S}_1^{\perp}}\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda}\mathbf{P}_{\mathcal{S}_2^{\perp}}\|_*. \tag{A2.13}$$

Combining (A2.6), (A2.9), and (A2.13), we have

$$\begin{aligned}&\langle 2(\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q}\rangle + \lambda\|\mathbf{P}_{\mathcal{S}_1^{\perp}}\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda}\mathbf{P}_{\mathcal{S}_2^{\perp}}\|_* \\ \le&(\lambda\sqrt{\operatorname{rank}(\mathbf{Q})} + \Lambda) \cdot \|\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda} - \mathbf{Q}\|_{\mathsf{F}} + \Gamma\|\mathbf{P}_{\mathcal{S}_1^{\perp}}\widehat{\mathbf{A}}^{\mathrm{G}}_{\lambda}\mathbf{P}_{\mathcal{S}_2^{\perp}}\|_*.\end{aligned} \tag{A2.14}$$

And now we turn to analyze Λ and Γ. Due to the inequality between the ℓ_2 norm and the Frobenius norm of matrices, we have

$$\Lambda = 2\|\mathcal{P}_{\mathbf{Q}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}})\|_{\mathsf{F}} \le 2\sqrt{\operatorname{rank}(\mathcal{P}_{\mathbf{Q}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}))} \cdot \|(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}})\|_2. \tag{A2.15}$$

By the definition of $\mathcal{P}_{\mathbf{Q}}$, we have

$$\begin{aligned}\mathcal{P}_{\mathbf{Q}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}) &= (\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}) - \mathbf{P}_{\mathcal{S}_1^{\perp}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}})\mathbf{P}_{\mathcal{S}_2^{\perp}} \\ &= \mathbf{P}_{\mathcal{S}_1^{\perp}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}})\mathbf{P}_{\mathcal{S}_2} + \mathbf{P}_{\mathcal{S}_1}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}).\end{aligned}$$

Combining the above equality with the fact that

$$\operatorname{rank}(\mathbf{P}_{\mathcal{S}_j}) \le \operatorname{rank}(\mathbf{Q}), \text{ for } j = 1, 2,$$

we have

$$\operatorname{rank}(\mathcal{P}_{\mathbf{Q}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}})) \leq 2 \cdot \operatorname{rank}(\mathbf{Q}). \tag{A2.16}$$

Combining (A2.15) and (A2.16) and setting $\Delta := \|\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}\|_2$, we have

$$\Lambda \leq 2\sqrt{2\operatorname{rank}(\mathbf{Q})} \cdot \|\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}\|_2 \leq 2\Delta\sqrt{2\operatorname{rank}(\mathbf{Q})}. \tag{A2.17}$$

We then turn to bound Γ. By the submultiplicity of ℓ_2 norm, we have

$$\Gamma = 2\|\mathbf{P}_{\mathcal{S}_1^{\perp}}(\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}})\mathbf{P}_{\mathcal{S}_2^{\perp}}\|_2 \leq 2\|\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}\|_2 = 2\Delta. \tag{A2.18}$$

Combining (A2.14), (A2.17), and (A2.18), we have

$$\begin{aligned}&\langle 2(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle + \lambda\|\mathbf{P}_{\mathcal{S}_1^{\perp}}\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\mathbf{P}_{\mathcal{S}_2^{\perp}}\|_* \\ \leq &(\lambda\sqrt{\operatorname{rank}(\mathbf{Q})} + 2\Delta\sqrt{2\operatorname{rank}(\mathbf{Q})}) \cdot \|\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\|_{\mathsf{F}} + 2\Delta\|\mathbf{P}_{\mathcal{S}_1^{\perp}}\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\mathbf{P}_{\mathcal{S}_2^{\perp}}\|_*.\end{aligned}$$

And it follows that

$$\begin{aligned}\langle 2(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle \leq &(\lambda\sqrt{\operatorname{rank}(\mathbf{Q})} + 2\Delta\sqrt{2\operatorname{rank}(\mathbf{Q})}) \cdot \|\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\|_{\mathsf{F}} \\ &+ (2\Delta - \lambda)\|\mathbf{P}_{\mathcal{S}_1^{\perp}}\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\mathbf{P}_{\mathcal{S}_2^{\perp}}\|_*.\end{aligned} \tag{A2.19}$$

Now we decompose the left-hand side of (A2.19) in the form

$$\begin{aligned}\langle 2(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle = \langle(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A}\rangle + \\ \langle(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle - \langle(\mathbf{Q} - \mathbf{A})\widehat{\mathbf{M}}, \mathbf{Q} - \mathbf{A}\rangle.\end{aligned}$$

Thus, it follows from (A2.19) that

$$\begin{aligned}&\langle(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A}\rangle + \langle(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\rangle \\ \leq &\langle(\mathbf{Q} - \mathbf{A})\widehat{\mathbf{M}}, \mathbf{Q} - \mathbf{A}\rangle + (\lambda\sqrt{\operatorname{rank}(\mathbf{Q})} + 2\Delta\sqrt{2\operatorname{rank}(\mathbf{Q})}) \cdot \|\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{Q}\|_{\mathsf{F}} \\ &+ (2\Delta - \lambda)\|\mathbf{P}_{\mathcal{S}_1^{\perp}}\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\mathbf{P}_{\mathcal{S}_2^{\perp}}\|_*.\end{aligned}$$

Because $\mathbf{Q}$ is an arbitrary matrix in $\mathbb{R}^{m\times n}$, setting $\mathbf{Q} = \mathbf{A}$, we have

$$\begin{aligned}2\langle(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A}\rangle \leq (\lambda\sqrt{\operatorname{rank}(\mathbf{A})} + 2\Delta\sqrt{2\operatorname{rank}(\mathbf{A})}) \cdot \|\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A}\|_{\mathsf{F}} \\ +(2\Delta - \lambda)\|\mathbf{P}_{\mathcal{S}_1^{\perp}}\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\mathbf{P}_{\mathcal{S}_2^{\perp}}\|_*.\end{aligned}$$

Replacing $\operatorname{rank}(\mathbf{A})$ by r implies

$$\begin{aligned}2\langle(\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A}\rangle \leq (\lambda\sqrt{r} + 2\Delta\sqrt{2r}) \cdot \|\widehat{\mathbf{A}}_\lambda^{\mathrm{G}} - \mathbf{A}\|_{\mathsf{F}} \\ +(2\Delta - \lambda)\|\mathbf{P}_{\mathcal{S}_1^{\perp}}\widehat{\mathbf{A}}_\lambda^{\mathrm{G}}\mathbf{P}_{\mathcal{S}_2^{\perp}}\|_*.\end{aligned} \tag{A2.20}$$

Finally, we bound Δ. To this end, we have

$$\Delta = \|\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{A}\widehat{\mathbf{M}}\|_2 \leq \|\widehat{\mathbf{M}}_1^{\mathsf{T}} - \mathbf{M}_1^{\mathsf{T}}\|_2 + \|\mathbf{A}(\widehat{\mathbf{M}} - \mathbf{M})\|_2 + \|\mathbf{M}_1^{\mathsf{T}} - \mathbf{A}\mathbf{M}\|_2.$$

Noting that $\mathbf{A}$ satisfies the following condition

$$\|\mathbf{A}\|_2 \leq \gamma_{\max} \quad \text{and} \quad \mathbf{M}_1 = \mathbf{M}\mathbf{A}^{\mathsf{T}},$$

it follows that

$$\Delta \leq \|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2 + \|\mathbf{A}\|_2\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 \leq \|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2 + \gamma_{\max} \cdot \|\widehat{\mathbf{M}} - \mathbf{M}\|_2.$$

By the condition (2.3), we have

$$\mathbb{P}\Big(\|\widehat{\mathbf{M}}-\mathbf{M}\|_2 \le \delta_1 \text{ and } \|\widehat{\mathbf{M}}_1-\mathbf{M}_1\|_2 \le \delta_2\Big) \ge 1-\epsilon_1-\epsilon_2. \tag{A2.21}$$

Therefore, we have, with probability no smaller than $1-\epsilon_1-\epsilon_2$,

$$\Delta \le \gamma_{\max}\delta_1+\delta_2.$$

By the assumption $\lambda \ge 2(\gamma_{\max}\delta_1+\delta_2)$, we deduce from (A2.20) that, under the event of (A2.21),

$$2\langle(\widehat{\mathbf{A}}^{\mathrm{G}}_\lambda-\mathbf{A})\widehat{\mathbf{M}}, \widehat{\mathbf{A}}^{\mathrm{G}}_\lambda-\mathbf{A}\rangle \le (\lambda\sqrt{r}+2\Delta\sqrt{2r})\cdot\|\widehat{\mathbf{A}}^{\mathrm{G}}_\lambda-\mathbf{A}\|_{\mathsf{F}}. \tag{A2.22}$$

Meanwhile, due to Lemma A2.4, we have

$$|\lambda_{\min}(\widehat{\mathbf{M}})-\lambda_{\min}(\mathbf{\Sigma})| \le \|\widehat{\mathbf{M}}-\mathbf{\Sigma}\|_2. \tag{A2.23}$$

Combining the assumption $\mu \le \lambda_{\min}(\mathbf{\Sigma})-\delta_1$, (A2.21), and (A2.23), we have, with probability no smaller than $1-\epsilon_1-\epsilon_2$,

$$\mu \le \lambda_{\min}(\widehat{\mathbf{M}}). \tag{A2.24}$$

Hence we deduce from (A2.22) that

$$2\mu\|\widehat{\mathbf{A}}^{\mathrm{G}}_\lambda-\mathbf{A}\|_{\mathsf{F}}^2 \le (\lambda\sqrt{r}+2\Delta\sqrt{2r})\cdot\|\widehat{\mathbf{A}}^{\mathrm{G}}_\lambda-\mathbf{A}\|_{\mathsf{F}}.$$

It then follows that

$$\|\widehat{\mathbf{A}}^{\mathrm{G}}_\lambda-\mathbf{A}\|_{\mathsf{F}} \le \frac{\lambda+2\sqrt{2}\Delta}{2\mu}\sqrt{r}.$$

So we conclude

$$\mathbb{P}\left(\|\widehat{\mathbf{A}}^{\mathrm{G}}_\lambda-\mathbf{A}\|_{\mathsf{F}} \le \frac{\lambda+2\sqrt{2}(\gamma_{\max}\delta_1+\delta_2)}{2\mu}\sqrt{r}\right) \ge 1-\epsilon_1-\epsilon_2,$$

where ϵ_1 and ϵ_2 go to zero when (T,d) goes to infinity.

□

A2.2 Proof of Theorem 5.1

Proof. Let's define

$$\mathbf{K} := \mathbb{E}\big[(\boldsymbol{S}_t^{\mathsf{T}},\boldsymbol{S}_{t-1}^{\mathsf{T}},\ldots,\boldsymbol{S}_{t-p+1}^{\mathsf{T}})^{\mathsf{T}}\cdot(\boldsymbol{S}_t^{\mathsf{T}},\boldsymbol{S}_{t-1}^{\mathsf{T}},\ldots,\boldsymbol{S}_{t-p+1}^{\mathsf{T}})\big],$$
$$\mathbf{K}_1 := \mathbb{E}\big[(\boldsymbol{S}_t^{\mathsf{T}},\boldsymbol{S}_{t-1}^{\mathsf{T}},\ldots,\boldsymbol{S}_{t-p+1}^{\mathsf{T}})^{\mathsf{T}}\cdot\boldsymbol{S}_{t+1}^{\mathsf{T}}\big].$$

Based on Lemma 2.2, it suffices to determine $\delta_1,\delta_2>0$ such that

$$\mathbb{P}\Big(\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}\|_2 \le \delta_1\Big) \ge 1-\epsilon_1, \tag{A2.25}$$

$$\mathbb{P}\Big(\|\widehat{\mathbf{\Omega}}_1-\mathbf{\Omega}_1\|_2 \le \delta_2\Big) \ge 1-\epsilon_2, \tag{A2.26}$$

where both δ_1,δ_2 are functions of (T,d) and both ϵ_1,ϵ_2 go to zero when (T,d) goes to infinity. The following proof is split into two parts according to the above statement.

Step I. First, to prove (A2.25), since $\{\boldsymbol{Z}_t\}_{t\in\mathbb{Z}}$ is globally elliptically distributed, we have

$$\mathbf{\Omega} = \sin\Big(\frac{\pi}{2}\mathbf{K}\Big).$$

Secondly, similar to Theorem 3.4, by Taylor's theorem, we have

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2 \leq \frac{\pi}{2} \underbrace{\| \cos\big(\frac{\pi}{2}\mathbf{K}\big) \circ (\widehat{\mathbf{K}} - \mathbf{K})\|_2}_{E_5} + \qquad \text{(A2.27)}$$

$$\frac{\pi^2}{8} \underbrace{\| \sin\big(\frac{\pi}{2}\bar{\mathbf{K}}\big) \circ (\widehat{\mathbf{K}} - \mathbf{K}) \circ (\widehat{\mathbf{K}} - \mathbf{K})\|_2}_{E_6},$$

for some matrix $\bar{\mathbf{K}}$ with each entry $\bar{\mathbf{K}}_{jk}$ as a value between $\mathbf{K}_{jk}$ and $\widehat{\mathbf{K}}_{jk}$.

Step I.1 For E_5, we define, for any $\boldsymbol{v} \in \mathbb{S}^{dp-1}$ (the unit sphere in the dp-dimensional Euclidean space) and $t = p, \ldots, T$,

$$W_t := \boldsymbol{v}^\top(\boldsymbol{S}_t^\top, \boldsymbol{S}_{t-1}^\top, \ldots, \boldsymbol{S}_{t-p+1}^\top)^\top = \boldsymbol{v}^\top \text{sign}\big((\boldsymbol{X}_t^\top, \boldsymbol{X}_{t-1}^\top, \ldots, \boldsymbol{X}_{t-p+1}^\top)^\top\big). \qquad \text{(A2.28)}$$

Meanwhile, similar to Theorem 3.4, we have $\{W_t^2\}_{t=p}^T$ also satisfies the strong mixing condition and the tail condition. According to the strong mixing condition, the tail condition, Lemma A2.5, Lemma A2.6, Lemma A2.7, and Lemma A2.8, there exists an absolute positive constant C_{41}, such that the following inequality holds,

$$\mathbb{P}\Big(\Big\| \cos\big(\frac{\pi}{2}\mathbf{K}\big) \circ (\widehat{\mathbf{K}} - \mathbf{K})\Big\|_2 \leq 4C_{41}\sqrt{\frac{dp}{T-p+1}}\Big) > 1 - 9^{dp} \cdot 3\exp(-C_{42}dp), \qquad \text{(A2.29)}$$

where C_{42} is a constant only depending on $C_{41}, \kappa_1, \kappa_2', \gamma_1, \lambda_{\max}(\mathbf{\Omega})$, and $\lambda_{\min}(\mathbf{\Omega})$.

Step I.2. In this step we upper bound the second order term E_6. Similar to Theorem 3.4, we have, for any positive constant β_3,

$$\mathbb{P}\Big(\Big\| \sin\big(\frac{\pi}{2}\bar{\mathbf{K}}\big) \circ (\widehat{\mathbf{K}} - \mathbf{K}) \circ (\widehat{\mathbf{K}} - \mathbf{K})\Big\|_2 \leq \frac{2dp\log(2dp/\beta_3)}{C_{43}(T-p+1)}\Big) > 1 - \frac{3\beta_3^2}{8}, \qquad \text{(A2.30)}$$

where C_{43} is a constant only depending on $\kappa_1, \kappa_2', \gamma_1, \lambda_{\max}(\Omega)$, and $\lambda_{\min}(\Omega)$.

Combining (A2.27), (A2.29), and (A2.30), there exist an absolute constant $C_4 > 0$ and another constant C_4', only depending on $C_4, \kappa_1, \kappa_2', \gamma_1, \lambda_{\max}(\mathbf{\Omega})$, and $\lambda_{\min}(\mathbf{\Omega})$, such that

$$\mathbb{P}\big(\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2 \leq C_4 \cdot \sqrt{\frac{dp}{T-p+1}}\big) > 1 - \epsilon_1, \qquad \text{(A2.31)}$$

where

$$\epsilon_1 = 9^{dp} \cdot 3\exp(-C_4'dp) + \frac{3}{2} \cdot (dp)^2 \cdot \exp\big(-2\sqrt{\frac{T-p+1}{dp}}\big),$$

and ϵ_1 goes to zero when (T, d) goes to infinity. (A2.31) then verifies (A2.25).

Step II. Now we study (A2.26). First, we have

$$\mathbf{\Omega}_1 = \sin(\frac{\pi}{2}\mathbf{K}_1).$$

And we also have

$$\begin{aligned} &\|\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}_1\|_2 \\ \leq \sqrt{p} \cdot \max_{1\leq j\leq p} \Big\| \sin\big(\frac{\pi}{2} \cdot \frac{1}{T-p}\sum_{t=p}^{T-1} &\boldsymbol{S}_{t+1-j}\boldsymbol{S}_{t+1}^{\mathsf{T}}\big) - \sin\big(\frac{\pi}{2}\cdot \mathbb{E}\boldsymbol{S}_{t+1-j}\boldsymbol{S}_{t+1}^{\mathsf{T}}\big)\Big\|_2. \end{aligned} \tag{A2.32}$$

By Taylor's theorem, similar to Theorem 3.4, for $j = 1, \ldots, p$, we only need to consider the first and second order terms for the formula in the maximum. And we conclude that there exists an absolute positive constant C_5 such that

$$\mathbb{P}\Big(\|\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}_1\|_2 \leq C_5 \cdot \sqrt{\frac{dp}{T-p}}\Big) > 1 - \epsilon_4, \tag{A2.33}$$

and ϵ_4 goes to zero when (T, d) goes to infinity.

Combining the assumption of λ, μ and (A2.31), (A2.33), we complete the proof. □

A2.3 Auxiliary lemmas

The following two lemmas are from Tsybakov (2009) and used in proving Theorem 4.1.

Lemma A2.1. Let $m \geq 8$. Then there exists a subset $\{\omega^{(0)}, \ldots, \omega^{(M)}\}$ of $\Omega = \{\omega = (\omega_1, \ldots, \omega_m) : \omega_i \in \{0, 1\}\}$ such that $\omega^{(0)} = (0, \ldots, 0)$,

$$\rho_{\mathsf{H}}(\omega^{(j)}, \omega^{(k)}) \geq \frac{m}{8}, \quad \text{for } 0 \leq j < k \leq M, \quad \text{and} \quad M \geq 2^{\frac{m}{8}}.$$

Here the metric $\rho(\cdot)_{\mathsf{H}}$ represents the Hamming distance.

Lemma A2.2. Assume $M \geq 2$ and suppose Θ, equipped with a distance $d(\cdot, \cdot)$, contains elements $\theta_0, \theta_1, \ldots, \theta_M$ such that $d(\theta_j, \theta_k) \geq 2s > 0$ for any $0 \leq j < k \leq M$. Also assume that $\mathbb{P}_j \ll \mathbb{P}_0$ for $j = 1, \ldots, M$ and

$$\frac{1}{M}\sum_{j=1}^{M} D_{\mathsf{KL}}(\mathbb{P}_j, \mathbb{P}_0) \leq \alpha \log M,$$

with $0 < \alpha < 1/8$ and $\mathbb{P}_j = \mathbb{P}_{\theta_j}, j = 0, 1, \ldots, M$. Then

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{P}_\theta(d(\widehat{\theta}, \theta) \geq s) \geq \frac{\sqrt{M}}{1+\sqrt{M}}\Big(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\Big) > 0.$$

The following lemma is the duality property of matrix Schatten norms, and is used in proving Theorem 4.1.

Lemma A2.3. For any two matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{m\times n}$, we have

$$|\langle \mathbf{M}, \mathbf{N}\rangle| \leq \|\mathbf{M}\|_* \cdot \|\mathbf{N}\|_2. \tag{A2.34}$$

And for given $\mathbf{M} \in \mathbb{R}^{m\times n}$, there exists $\mathbf{N}_0 \in \mathbb{R}^{m\times n}$ with $\|\mathbf{N}_0\|_2 = 1$ such that

$$|\langle \mathbf{M}, \mathbf{N}\rangle| = \|\mathbf{M}\|_*.$$

The following lemma comes from Theorem 3.3.16 in Horn and Johnson (1994) and is used in proving Lemma 2.2, Theorem 3.4, and Theorem 5.1.

Lemma A2.4. For any two matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{m\times n}$ and $r := \min\{m, n\}$, the following inequality holds,

$$|\sigma_j(\mathbf{M}+\mathbf{N}) - \sigma_j(\mathbf{M})| \le \sigma_1(\mathbf{N}), \quad \text{for } j = 1, \dots, r.$$

The following two lemmas from Ledoux and Talagrand (1991) are about ϵ-net arguments and used in the proofs of Theorem 3.4 and Theorem 5.1.

Lemma A2.5. Let (Ω, ρ) be a metric space and let $\epsilon > 0$. A subset $S_\epsilon(\Omega)$ of Ω is called an ϵ-net of Ω if every point $\omega \in \Omega$ can be approximated to within ϵ by some point $\xi \in S_\epsilon(\Omega)$, i.e., such that $\rho(\omega, \xi) \le \epsilon$. The minimal cardinality of an ϵ-net Ω, if finite, is denoted by $\mathcal{N}(\Omega, \epsilon)$ and is called the covering number of Ω at scale ϵ. With these, the unit Euclidean sphere $\mathbb{S}^{n-1}$ equipped with the Euclidean metric satisfies, for every $\epsilon > 0$,

$$\mathcal{N}(\mathbb{S}^{n-1}, \epsilon) \le \big(1 + 2/\epsilon\big)^n.$$

Lemma A2.6. Let $\mathbf{M}$ be a symmetric n by n matrix, and let S_ϵ be an ϵ-net of $\mathbb{S}^{n-1}$ for some $\epsilon \in [0, 1/2)$. Then

$$\sup_{\boldsymbol{v}\in\mathbb{S}^{n-1}} |\langle \mathbf{M}\boldsymbol{v}, \boldsymbol{v}\rangle| \le (1-2\epsilon)^{-1} \sup_{\boldsymbol{v}\in S_\epsilon} |\langle \mathbf{M}\boldsymbol{v}, \boldsymbol{v}\rangle|.$$

The following lemma, from Merlevède et al. (2011), depicts a Bernstein type bound on the tail probability of the partial sums of a sequence of dependent random variables satisfying a certain tail condition. It is used in the proofs of Theorem 3.4 and Theorem 5.1.

Lemma A2.7. Let $\{X_t\}_{t\in\mathbb{Z}}$ be a sequence of centered real-valued random variables. Suppose that the sequence satisfies the strong mixing condition,

$$\alpha(n) \le \exp(-2D_1 n^{\gamma_1}),$$

for some absolute constants $D_1 > 0$ and $\gamma_1 > 0$. And it also satisfies the tail condition that

$$\mathbb{P}(|X_t| > \xi) \le \exp(1 - (\xi/D_2)^{\gamma_2}),$$

for absolute constants $D_2 > 0$, $\gamma_2 > 0$ and all $\xi > 0$, $t \ge 1$. Define $\gamma_0 := \gamma_1\gamma_2/(\gamma_1+\gamma_2)$. Suppose $\gamma_0 < 1$. Then there exist positive constants D_3, D_4, D_5, D_6, D_7, depending on D_1, D_2, γ_1, and γ_2, such that, for all $T \ge 4$,

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{T}\Big|\sum_{t=1}^{T} X_t\Big| \ge \eta\Big) \le & T\exp\Big(-\frac{(T\eta)^{\gamma_0}}{D_3}\Big) + \exp\Big(-\frac{(T\eta)^2}{D_4(1+D_5T)}\Big) \\ & + \exp\Big(-\frac{(T\eta)^2}{D_6T}\exp\Big(\frac{(T\eta)^{\gamma_0(1-\gamma_0)}}{D_7(\log T\eta)^{\gamma_0}}\Big)\Big).\end{aligned}$$

The following two lemmas, coming from Lemmas 4.3 and 4.4 in Wegkamp and Zhao (2016), are used in the proof of Theorem 3.4.

Lemma A2.8. Let $\mathbf{M}$ and $\mathbf{N}$ be any m by m square matrices. If $\sin(\frac{\pi}{2}\mathbf{M})$ is positive semidefinite with its diagonal elements all equal to one, then we have

$$\|\cos\Big(\frac{\pi}{2}\mathbf{M}\Big) \circ (\mathbf{N}-\mathbf{M})\|_2 \le 2\|\mathbf{N}-\mathbf{M}\|_2. \tag{A2.35}$$

Lemma A2.9. For any two matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{m\times n}$, if $|\mathbf{M}_{jk}| \le \mathbf{N}_{jk}$ for all $1 \le j \le m$ and $1 \le k \le n$, then $\|\mathbf{M}\|_2 \le \|\mathbf{N}\|_2$.

References

Brualdi, R. A. and Schneider, H. (1983). Determinantal identities: Gauss, Schur, Cauchy, Sylvester, Kronecker, Jacobi, Binet, Laplace, Muir, and Cayley. *Linear Algebra and its Applications*, 52–53:769–791.

Fan, J., Han, F., and Liu, H. (2014). PAGE: Robust pattern guided estimation of large covariance matrix. Technical report, Princeton University.

Horn, R. A. and Johnson, C. R. (1994). *Topics in Matrix Analysis.* Cambridge University Press.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes.* Springer.

Merlevède, F., Peligrad, M., and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer.

Watson, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45.

Wegkamp, M. and Zhao, Y. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli*, 22(2):1184–1226.