MATRIX AUTOREGRESSIVE MODEL WITH VECTOR TIME SERIES COVARIATES FOR SPATIO-TEMPORAL DATA

Hu Sun[1], Zuofeng Shang[2], Yang Chen[1]

[1]*University of Michigan and* [2]*New Jersey Institute of Technology*

Supplementary Material

This supplemental material is organized as follows. Section S1 presents the algorithmic details of the penalized MLE using alternating minimization outlined in Section 3.1. In Section S2, we prove Proposition 1 on the equivalence of the estimation problem of MARAC to a kernel ridge regression problem. In Section S3, we prove Theorem 1 on the joint stationarity condition of the matrix and auxiliary vector time series. Then in Section S4, we provide proofs of the theoretical results under fixed spatial dimensionality, including Proposition 2, Theorem 2 and Corollary 1. In Section S5, we present proofs of the theoretical results under high spatial dimensionality, namely Theorem 3. All essential lemmas used throughout the proofs are presented and proved in Section S6. Finally, we include additional details of the simulation in Section S7 as well as an approximated estimating algorithm for obtaining the penalized MLE via kernel truncation.

In this supplemental material, we use $\bar{\rho}(\cdot)$, $\rho_i(\cdot)$, $\underline{\rho}(\cdot)$, and $\|\cdot\|_s$ to denote the maximum, i^{th} largest, minimum eigenvalue, and spectral norm of a matrix. We use $a \vee b, a \wedge b$ to denote the maximum and minimum of a and b, respectively. For two sequences of random variables,

say X_n, Y_n, we use $X_n \lesssim Y_n$ to denote the case where $X_n/Y_n = O_P(1)$, and $X_n \gtrsim Y_n$ to denote the case where $Y_n/X_n = O_P(1)$. We then use $X_n \asymp Y_n$ to denote the case where both $X_n \lesssim Y_n$ and $X_n \gtrsim Y_n$ hold.

S1 Alternating Minimization Algorithm for PMLE

To solve the optimization problem in (3.7) for $\mathbf{A}_p$ at the $(l+1)^{\text{th}}$ iteration, it suffices to solve the following least-square problem:

$$\min_{\mathbf{A}_p} \left\{ \sum_{t\in[T]} \operatorname{tr}\left(\widetilde{\mathbf{X}}_t(\mathbf{A}_p)^\top \left(\mathbf{\Sigma}_r^{(l)}\right)^{-1} \widetilde{\mathbf{X}}_t(\mathbf{A}_p) \left(\mathbf{\Sigma}_c^{(l)}\right)^{-1} \right) \right\}, \tag{S1.1}$$

where $\widetilde{\mathbf{X}}_t(\mathbf{A}_p)$ is the residual matrix when predicting $\mathbf{X}_t$:

$$\begin{aligned}\widetilde{\mathbf{X}}_t(\mathbf{A}_p) = \mathbf{X}_t &- \sum_{p'<p} \mathbf{A}_{p'}^{(l+1)} \mathbf{X}_{t-p'} \left(\mathbf{B}_{p'}^{(l+1)}\right)^\top - \sum_{p'>p} \mathbf{A}_{p'}^{(l)} \mathbf{X}_{t-p'} \left(\mathbf{B}_{p'}^{(l)}\right)^\top \\ &- \sum_{q\in[Q]} \boldsymbol{\mathcal{G}}_q^{(l)} \bar{\times} \mathbf{z}_{t-q} - \mathbf{A}_p \mathbf{X}_{t-p} \left(\mathbf{B}_p^{(l)}\right)^\top = \widetilde{\mathbf{X}}_{t,-p} - \mathbf{A}_p \mathbf{X}_{t-p} \left(\mathbf{B}_p^{(l)}\right)^\top\end{aligned}$$

and we use $\widetilde{\mathbf{X}}_{t,-p}$ to denote the partial residual excluding the term involving $\mathbf{X}_{t-p}$ and use $\boldsymbol{\mathcal{G}}_q^{(l)}$ to denote the tensor coefficient satisfying $[\boldsymbol{\mathcal{G}}_q^{(l)}]_{ijd} = \langle [\mathbf{K}]_{u,:}^\top, [\mathbf{\Gamma}_q^{(l)}]_{:d} \rangle$, with $u = i + (j-1)M$. The superscript l represents the value at the l^{th} iteration. To simplify the notation, we define $\mathbf{\Phi}(\mathbf{A}_t, \mathbf{B}_t, \mathbf{\Sigma}) = \sum_t \mathbf{A}_t^\top \mathbf{\Sigma}^{-1} \mathbf{B}_t$, where $\mathbf{\Sigma}, \mathbf{A}_t, \mathbf{B}_t$ are arbitrary matrices/vectors with conformal matrix sizes and we simply write $\mathbf{\Phi}(\mathbf{A}_t, \mathbf{\Sigma})$ if $\mathbf{A}_t = \mathbf{B}_t$. Solving (S1.1) yields the following updating formula for $\mathbf{A}_p^{(l+1)}$:

$$\mathbf{A}_p^{(l+1)} \leftarrow \mathbf{\Phi}\left(\widetilde{\mathbf{X}}_{t,-p}^\top, \mathbf{B}_p^{(l)} \mathbf{X}_{t-p}^\top, \mathbf{\Sigma}_c^{(l)}\right) \mathbf{\Phi}\left(\mathbf{B}_p^{(l)} \mathbf{X}_{t-p}^\top, \mathbf{\Sigma}_c^{(l)}\right)^{-1} \tag{S1.2}$$

Similarly, we have the following updating formula for $\mathbf{B}_p^{(l+1)}$:

$$\mathbf{B}_p^{(l+1)} \leftarrow \mathbf{\Phi}\left(\widetilde{\mathbf{X}}_{t,-p}, \mathbf{A}_p^{(l+1)} \mathbf{X}_{t-p}, \mathbf{\Sigma}_r^{(l)}\right) \mathbf{\Phi}\left(\mathbf{A}_p^{(l+1)} \mathbf{X}_{t-p}, \mathbf{\Sigma}_r^{(l)}\right)^{-1} \tag{S1.3}$$

For updating $\boldsymbol{\Gamma}_q$, or its vectorized version $\boldsymbol{\gamma}_q = \mathbf{vec}\left(\boldsymbol{\Gamma}_q\right)$, it is required to solve the following kernel ridge regression problem:

$$\min_{\boldsymbol{\gamma}_q}\left\{\frac{1}{2T}\boldsymbol{\Phi}\left(\widetilde{\mathbf{x}}_{t,-q}-\left(\mathbf{z}_{t-q}^{\top}\otimes\mathbf{K}\right)\boldsymbol{\gamma}_q,\boldsymbol{\Sigma}^{(l)}\right)+\frac{\lambda}{2}\boldsymbol{\gamma}_q^{\top}\left(\mathbf{I}_D\otimes\mathbf{K}\right)\boldsymbol{\gamma}_q\right\},$$

where $\boldsymbol{\Sigma}^{(l)}=\boldsymbol{\Sigma}_c^{(l)}\otimes\boldsymbol{\Sigma}_r^{(l)}$ and $\widetilde{\mathbf{x}}_{t,-q}$ is the vectorized partial residual of $\mathbf{X}_t$ by leaving out the lag-q auxiliary predictor, defined in a similar way as $\widetilde{\mathbf{X}}_{t,-p}$. Solving the kernel ridge regression leads to the following updating formula for $\boldsymbol{\gamma}_q^{(l+1)}$:

$$\boldsymbol{\gamma}_q^{(l+1)}\leftarrow\left[\left(\sum_{t\in[T]}\mathbf{z}_{t-q}\mathbf{z}_{t-q}^{\top}\right)\otimes\mathbf{K}+\lambda T\left(\mathbf{I}_D\otimes\boldsymbol{\Sigma}^{(l)}\right)\right]^{-1}\left[\sum_{t\in[T]}\left(\mathbf{z}_{t-q}\otimes\widetilde{\mathbf{x}}_{t,-q}\right)\right]. \tag{S1.4}$$

The step in (S1.4) can be slow since one needs to invert a square matrix of size $MND\times MND$. In the supplemental material, we propose an approximation to (S1.4) to avoid inverting large matrices.

The updating rule of $\boldsymbol{\Sigma}_r^{(l+1)}$ and $\boldsymbol{\Sigma}_c^{(l+1)}$ can be easily derived by taking their derivative in (3.7) and setting it to zero. Specifically, we have:

$$\boldsymbol{\Sigma}_r^{(l+1)}\leftarrow\frac{1}{NT}\boldsymbol{\Phi}\left(\widetilde{\mathbf{X}}_t^{\top},\boldsymbol{\Sigma}_c^{(l)}\right) \tag{S1.5}$$

$$\boldsymbol{\Sigma}_c^{(l+1)}\leftarrow\frac{1}{MT}\boldsymbol{\Phi}\left(\widetilde{\mathbf{X}}_t,\boldsymbol{\Sigma}_r^{(l+1)}\right). \tag{S1.6}$$

where $\widetilde{\mathbf{X}}_t$ is the full residual when predicting $\mathbf{X}_t$.

The algorithm cycles through (S1.2), (S1.3), (S1.4), (S1.5) and (S1.6) and terminates when $\mathbf{B}_p^{(l)}\otimes\mathbf{A}_p^{(l)}$, $\boldsymbol{\mathcal{G}}_q^{(l)}$, $\boldsymbol{\Sigma}_c^{(l)}\otimes\boldsymbol{\Sigma}_r^{(l)}$ have their relative changes between iterations fall under a pre-specified threshold. We summarize the algorithm in pseudo-code in Algorithm 1.

Remark S1.1. (Convergence of Kronecker Product) When dealing with high-dimensional matrices, it is cumbersome to compute the change between $\mathbf{B}_p^{(l)}\otimes\mathbf{A}_p^{(l)}$ and $\mathbf{B}_p^{(l+1)}\otimes\mathbf{A}_p^{(l+1)}$ under the Frobenius norm. An upper bound of $\|\mathbf{B}_p^{(l+1)}\otimes\mathbf{A}_p^{(l+1)}-\mathbf{B}_p^{(l)}\otimes\mathbf{A}_p^{(l)}\|_{\mathrm{F}}$ can be used

Algorithm 1 Alternating Minimization Algorithm for PMLE

Randomly initialize parameters $\boldsymbol{\Theta}^{(0)} = \{\mathbf{A}_1^{(0)}, \mathbf{B}_1^{(0)}, \ldots, \mathbf{A}_P^{(0)}, \mathbf{B}_P^{(0)}, \boldsymbol{\Gamma}_1^{(0)}, \ldots, \boldsymbol{\Gamma}_Q^{(0)}, \boldsymbol{\Sigma}_r^{(0)}, \boldsymbol{\Sigma}_c^{(0)}\}$.

$k \leftarrow 0$.

while not converge **do**

for $\boldsymbol{\eta}^{(k)}$ in $[\mathbf{A}_1^{(k)}, \mathbf{B}_1^{(k)}, \ldots, \mathbf{A}_P^{(k)}, \mathbf{B}_P^{(k)}, \boldsymbol{\Gamma}_1^{(k)}, \ldots, \boldsymbol{\Gamma}_Q^{(k)}, \boldsymbol{\Sigma}_r^{(k)}, \boldsymbol{\Sigma}_c^{(k)}]$ **do**

$\boldsymbol{\eta}^{(k+1)} \leftarrow \arg\min_{\boldsymbol{\eta}} \mathfrak{L}_\lambda \left(\boldsymbol{\eta}; \boldsymbol{\Theta}^{(k)} \setminus \{\boldsymbol{\eta}^{(k)}\}\right)$. ▷ Details in (S1.2), (S1.3), (S1.4), (S1.5) and (S1.6).

Replace $\boldsymbol{\eta}^{(k)}$ with $\boldsymbol{\eta}^{(k+1)}$ in $\boldsymbol{\Theta}^{(k)}$.

end for

$\boldsymbol{\Theta}^{(k+1)} \leftarrow \boldsymbol{\Theta}^{(k)}$.

$k \leftarrow k + 1$.

end while

for $p = 1, 2, \ldots, P$ **do**

$c \leftarrow \text{sign}(\text{tr}(\mathbf{A}_p^{(k)})) \cdot \|\mathbf{A}_p^{(k)}\|_{\text{F}}$.

$\mathbf{A}_p^{(k)} \leftarrow c^{-1} \cdot \mathbf{A}_p^{(k)}, \mathbf{B}_p^{(k)} \leftarrow c \cdot \mathbf{B}_p^{(k)}$.

end for

return $\boldsymbol{\Theta}^{(k)}$.

instead:

$$\|\mathbf{B}_p^{(l+1)} - \mathbf{B}_p^{(l)}\|_{\mathrm{F}} \cdot \|\mathbf{A}_p^{(l+1)}\|_{\mathrm{F}} + \|\mathbf{B}_p^{(l)}\|_{\mathrm{F}} \cdot \|\mathbf{A}_p^{(l+1)} - \mathbf{A}_p^{(l)}\|_{\mathrm{F}}, \tag{S1.7}$$

and a similar bound can be used for the convergence check of $\mathbf{\Sigma}_c^{(l)} \otimes \mathbf{\Sigma}_r^{(l)}$.

S2 Proof of Proposition 1

Proof. For each function $g_{q,d}(\cdot) \in \mathbb{H}_k$, we can decompose it as follows:

$$g_{q,d}(\cdot) = \sum_{s \in \mathbb{S}} \gamma_{q,d,s} k(\cdot, s) + \sum_{j=1}^{J} \alpha_{q,d,j} \phi_j(\cdot) + h_{q,d}(\cdot),$$

where $h_{q,d}(\cdot)$ does not belong to the null space of $\mathbb{H}_k$ nor the span of $\{k(\cdot, s) | s \in \mathbb{S}\}$. Here we assume that the null space of $\mathbb{H}_k$ contains only the zero function, so $\phi_j(\cdot) = 0,$ for all j.

By the reproducing property of the kernel $k(\cdot, \cdot)$, we have $\langle g_{q,d}, k(\cdot, s') \rangle_{\mathbb{H}_k} = g_{q,d}(s') = \sum_{s \in \mathbb{S}} \gamma_{q,d,s} k(s, s')$, which is independent of $h_{q,d}(\cdot)$, and therefore $h_{q,d}(\cdot)$ is independent of the prediction for $\mathbf{x}_t$ in the MARAC model. In addition, for any $h_{q,d}(\cdot) \notin \mathrm{span}(\{k(\cdot, s) | s \in \mathbb{S}\})$, we have:

$$\|g_{q,d}\|_{\mathbb{H}_k}^2 = \gamma_{q,d}^\top \mathbf{K} \gamma_{q,d} + \|h_{q,d}\|_{\mathbb{H}_k}^2 \geq \|\sum_{s \in \mathbb{S}} \gamma_{q,d,s} k(\cdot, s)\|_{\mathbb{H}_k}^2,$$

and the equality holds only if $h_{q,d}(\cdot) = 0$. Consequently, the global minimizer for the constrained optimization problem (3.3) must have $h_{q,d}(\cdot) = 0$. It then follows that the squared RKHS functional norm penalty for $g_{q,d}$ can be written as $\gamma_{q,d}^\top \mathbf{K} \gamma_{q,d}$ and the tensor coefficient $\boldsymbol{\mathcal{G}}_q$ satisfies $\mathbf{vec}\left([\boldsymbol{\mathcal{G}}]_{::d}\right) = \mathbf{K} \gamma_{q,d}$. The remainder of the proof is straightforward by simple linear algebra, and thus we omit it here. □

S3 Proof of Theorem 1

Proof. Under Assumption 1 that the vector time series $\mathbf{z}_t$ follows a VAR($\widetilde{Q}$) process, we can derive that the vectorized matrix time series $\mathbf{X}_t$ and the vector time series $\mathbf{z}_t$ jointly follow a VAR($\max(P, Q, \widetilde{Q})$) process, namely,

$$\begin{bmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{bmatrix} = \sum_{l=1}^{\max(P,Q,\widetilde{Q})} \begin{bmatrix} (\mathbf{B}_l \otimes \mathbf{A}_l) \odot \mathbf{1}_{\{l \le P\}} & \mathbf{G}_l^\top \odot \mathbf{1}_{\{l \le Q\}} \\ \mathbf{O}_{D \times S} & \mathbf{C}_l \odot \mathbf{1}_{\{l \le \widetilde{Q}\}} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{t-l} \\ \mathbf{z}_{t-l} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_t \\ \boldsymbol{\nu}_t \end{bmatrix}. \tag{S3.8}$$

Let $L = \max(P, Q, \widetilde{Q})$ and $\mathbf{y}_t = [\mathbf{x}_t^\top, \mathbf{z}_t^\top]$. Denote the transition matrix in (S3.8) at lag-l as $\mathbf{J}_l \in \mathbb{R}^{(S+D)\times(S+D)}$ and the error term as $\mathbf{u}_t^\top = [\mathbf{e}_t^\top, \boldsymbol{\nu}_t^\top]$, then we can rewrite the VAR(L) process in (S3.8) as a VAR(1) process as:

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-L+1} \end{bmatrix} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{J}_2 & \cdots & \mathbf{J}_{L-1} & \mathbf{J}_L \\ \mathbf{I}_{S+D} & \mathbf{O}_{S+D} & \cdots & \cdots & \mathbf{O}_{S+D} \\ \mathbf{O}_{S+D} & \mathbf{I}_{S+D} & \mathbf{O}_{S+D} & \cdots & \mathbf{O}_{S+D} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{O}_{S+D} & \mathbf{O}_{S+D} & \cdots & \mathbf{I}_{S+D} & \mathbf{O}_{S+D} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-L} \end{bmatrix} + \begin{bmatrix} \mathbf{u}_t \\ \mathbf{0}_{S+D} \\ \vdots \\ \mathbf{0}_{S+D} \end{bmatrix}, \tag{S3.9}$$

where we use $\mathbf{O}_{S+D}$ to denote a zero matrix of size $(S+D) \times (S+D)$. For this VAR(1) process to be stationary, we require that $\det(\lambda \mathbf{I} - \mathbf{J}) \neq 0$ for all $|\lambda| \geq 1, \lambda \in \mathbb{C}$, where $\mathbf{J}$ is the transition matrix in (S3.9). The determinant $\det(\lambda \mathbf{I} - \mathbf{J})$ can be simplified by column operations as:

$$\begin{aligned} &\det(\lambda \mathbf{I} - \mathbf{J}) \\ &= \det \begin{bmatrix} \lambda^L \mathbf{I}_S - \sum_{l=1}^{L} \lambda^{L-l} (\mathbf{B}_l \otimes \mathbf{A}_l) \odot \mathbf{1}_{\{l \le P\}} & -\sum_{l=1}^{L} \lambda^{L-l} \mathbf{G}_l^\top \odot \mathbf{1}_{\{l \le Q\}} \\ \mathbf{O} & \lambda^L \mathbf{I}_D - \sum_{l=1}^{L} \lambda^{L-l} \mathbf{C}_l \odot \mathbf{1}_{\{l \le \widetilde{Q}\}} \end{bmatrix} \\ &= \lambda^{2L} \det[\boldsymbol{\Phi}_1(\lambda)] \det[\boldsymbol{\Phi}_2(\lambda)], \end{aligned}$$

where $\mathbf{\Phi}_1(\lambda) = \mathbf{I}_S - \sum_{p=1}^{P} \lambda^{-p} (\mathbf{B}_p \otimes \mathbf{A}_p)$ and $\mathbf{\Phi}_2(\lambda) = \mathbf{I}_D - \sum_{\widetilde{q}=1}^{\widetilde{Q}} \lambda^{-\widetilde{q}} \mathbf{C}_{\widetilde{q}}$, and setting $y = 1/\lambda$ completes the proof. □

S4 Theory under Fixed Spatial Dimension

S4.1 Proof of Proposition 2

Proof. For the brevity of the presentation, we fix P, Q as 1, but the proofs presented below can be easily extended to an arbitrary P, Q. For the vectorized MARAC$(1,1)$ model (2.4), we can equivalently write it as:

$$\mathbf{x}_t = \mathbf{y}_t \boldsymbol{\theta} + \mathbf{e}_t, \tag{S4.10}$$

where $\mathbf{y}_t = [\mathbf{x}_{t-1}^\top \otimes \mathbf{I}_S; \mathbf{z}_{t-1}^\top \otimes \mathbf{K}]$ and $\boldsymbol{\theta} = [\mathbf{vec}\,(\mathbf{B}_1 \otimes \mathbf{A}_1)^\top, \boldsymbol{\gamma}_1^\top]^\top$. Using $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ to denote the precision matrix for $\mathbf{e}_t$, we can rewrite the penalized likelihood in (3.7) for $(\boldsymbol{\theta}, \mathbf{\Omega})$ as:

$$h(\boldsymbol{\theta}, \mathbf{\Omega}) = -\frac{1}{2} \log |\mathbf{\Omega}| + \frac{1}{2} \mathrm{tr}\,(\mathbf{\Omega} \mathbf{S}(\boldsymbol{\theta})) + \frac{\lambda}{2} \boldsymbol{\theta}^\top \widetilde{\mathbf{K}} \boldsymbol{\theta}, \tag{S4.11}$$

where $\mathbf{S}(\boldsymbol{\theta}) = T^{-1} \sum_{t=1}^{T} (\mathbf{x}_t - \mathbf{y}_t \boldsymbol{\theta})(\mathbf{x}_t - \mathbf{y}_t \boldsymbol{\theta})^\top$, $\widetilde{\mathbf{K}}$ is defined as:

$$\widetilde{\mathbf{K}} = \begin{bmatrix} \mathbf{O}_{S \times S} \otimes \mathbf{K} & \mathbf{O}_{S \times D} \otimes \mathbf{K} \\ \mathbf{O}_{D \times S} \otimes \mathbf{K} & \mathbf{I}_D \otimes \mathbf{K} \end{bmatrix}.$$

We use $\boldsymbol{\theta}^*, \mathbf{\Omega}^*$ to denote the ground truth of $\boldsymbol{\theta}, \mathbf{\Omega}$, respectively. We define $\mathbb{F}_{\boldsymbol{\theta}}$ and $\mathbb{F}_{\mathbf{\Omega}}$ as:

$$\mathbb{F}_{\boldsymbol{\theta}} = \{[\mathbf{vec}\,(\mathbf{B}_1 \otimes \mathbf{A}_1)^\top, \boldsymbol{\gamma}_1^\top]^\top | \|\mathbf{A}_1\|_{\mathrm{F}} = 1, \mathrm{sign}(\mathrm{tr}\,(\mathbf{A}_1)) = 1\}$$

$$\mathbb{F}_{\mathbf{\Omega}} = \{\mathbf{\Sigma}_c^{-1} \otimes \mathbf{\Sigma}_r^{-1} | \mathbf{\Sigma}_r \in \mathbb{R}^{M \times M}, \mathbf{\Sigma}_c \in \mathbb{R}^{N \times N}, \underline{\rho}(\mathbf{\Sigma}_r), \underline{\rho}(\mathbf{\Sigma}_c) > 0\}.$$

The estimators of MARAC, denoted as $\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{\Omega}}$, is the minimizer of $h(\boldsymbol{\theta}, \mathbf{\Omega})$ with $\boldsymbol{\theta} \in \mathbb{F}_{\boldsymbol{\theta}}, \mathbf{\Omega} \in \mathbb{F}_{\mathbf{\Omega}}$.

In order to establish the consistency of $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Omega}}^{-1}$, it suffices to show that for any constant $c > 0$:

$$\mathrm{P}\left(\inf_{\|\bar{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{\mathrm{F}} \geq c} \inf_{\bar{\boldsymbol{\theta}}} h(\bar{\boldsymbol{\theta}}, \bar{\mathbf{\Omega}}) \leq h(\boldsymbol{\theta}^*, \mathbf{\Omega}^*) \right) \to 0, \text{ as } T \to \infty. \tag{S4.12}$$

This is because if (S4.12) is established, then as $T \to \infty$ we have:

$$\mathrm{P}\left(\inf_{\|\bar{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^*\|_{\mathrm{F}}\geq c}\inf_{\bar{\boldsymbol{\theta}}\in\mathbb{F}_{\boldsymbol{\theta}}} h(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\Omega}}) \geq \inf_{\|\bar{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^*\|_{\mathrm{F}}\geq c}\inf_{\bar{\boldsymbol{\theta}}} h(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\Omega}}) > h(\boldsymbol{\theta}^*,\boldsymbol{\Omega}^*) \geq h(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Omega}})\right)$$

approaching 1 and thus we must have $\|\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^*\|_{\mathrm{F}} < c$ with probability approaching 1 as $T \to \infty$, and the consistency is established since c is arbitrary.

To prove (S4.12), we first fix $\boldsymbol{\Omega} = \bar{\boldsymbol{\Omega}}$ and let $\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) = \arg\min_{\boldsymbol{\theta}} h(\boldsymbol{\theta},\bar{\boldsymbol{\Omega}})$, thus we have:

$$\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) = \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}}\mathbf{y}_t}{T} + \lambda\widetilde{\mathbf{K}}\right)^{-1}\left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}}\mathbf{x}_t}{T}\right), \tag{S4.13}$$

which is a consistent estimator of $\boldsymbol{\theta}^*$ for any $\bar{\boldsymbol{\Omega}}$ given that $\lambda \to 0$ and the matrix and vector time series are covariance-stationary. To see that $\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) \xrightarrow{p} \boldsymbol{\theta}^*$, notice that:

$$\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) = (\mathbf{I}-\lambda\widetilde{\mathbf{K}})\boldsymbol{\theta}^* + \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}}\mathbf{y}_t}{T} + \lambda\widetilde{\mathbf{K}}\right)^{-1}\left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}}\mathbf{e}_t}{T}\right), \tag{S4.14}$$

and the first term converges to $\boldsymbol{\theta}^*$ since $\lambda = o(1)$. In the second term of (S4.14), we have:

$$\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}}\mathbf{y}_t}{T} + \lambda\widetilde{\mathbf{K}} \xrightarrow{p} \begin{bmatrix} \boldsymbol{\Sigma}^*_{\mathbf{x},\mathbf{x}}\otimes\bar{\boldsymbol{\Omega}} & \boldsymbol{\Sigma}^*_{\mathbf{x},\mathbf{z}}\otimes\bar{\boldsymbol{\Omega}}\mathbf{K} \\ \boldsymbol{\Sigma}^*_{\mathbf{z},\mathbf{x}}\otimes\mathbf{K}\bar{\boldsymbol{\Omega}} & \boldsymbol{\Sigma}^*_{\mathbf{z},\mathbf{z}}\otimes\mathbf{K}\bar{\boldsymbol{\Omega}}\mathbf{K} \end{bmatrix}, \tag{S4.15}$$

where $\boldsymbol{\Sigma}^*_{\mathbf{x},\mathbf{x}} = \mathrm{Var}(\mathbf{x}_t)$, $\boldsymbol{\Sigma}^*_{\mathbf{x},\mathbf{z}} = \mathrm{Cov}(\mathbf{x}_t,\mathbf{z}_t)$ and $\boldsymbol{\Sigma}^*_{\mathbf{z},\mathbf{z}} = \mathrm{Var}(\mathbf{z}_t)$. The convergence in probability in (S4.15) holds due to the joint stationarity of $\mathbf{x}_t$ and $\mathbf{z}_t$ and the assumption that $\lambda = o(1)$. We further note that the sequence $\{\mathbf{y}_t^\top\bar{\boldsymbol{\Omega}}\mathbf{e}_t\}_{t=1}^T$ is a martingale difference sequence (MDS), and we have $\sum_{t=1}^T \mathbf{y}_t^\top\bar{\boldsymbol{\Omega}}\mathbf{e}_t/T = O_P(T^{-1/2})$ by the central limit theorem (CLT) of MDS (see proposition 7.9 of Hamilton (2020) for the central limit theorem of martingale difference sequence). Combining this result together with (S4.15), we conclude that the second term in (S4.14) is $o_P(1)$ and thus $\widetilde{\boldsymbol{\theta}}(\bar{\Omega})$ is consistent for $\boldsymbol{\theta}^*$.

Plugging $\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})$ into $h(\boldsymbol{\theta},\bar{\boldsymbol{\Omega}})$ yields the profile likelihood of $\bar{\boldsymbol{\Omega}}$:

$$\ell(\bar{\boldsymbol{\Omega}}) = -\frac{1}{2}\log|\bar{\boldsymbol{\Omega}}| + \frac{1}{2}\mathrm{tr}\left(\bar{\boldsymbol{\Omega}}\frac{\sum_t \mathbf{x}_t[\mathbf{x}_t - \mathbf{y}_t\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})]^\top}{T}\right).$$

To prove (S4.12), it suffices to show that:

$$\mathrm{P}\left(\inf_{\|\bar{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^*\|_{\mathrm{F}}\geq c}\ell(\bar{\boldsymbol{\Omega}})\leq\ell(\boldsymbol{\Omega}^*)\right)\to 0,\ \text{as } T\to\infty, \tag{S4.16}$$

since $\ell(\boldsymbol{\Omega}^*)\leq h(\boldsymbol{\theta}^*,\boldsymbol{\Omega}^*)$. Now, since $\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\xrightarrow{p}\boldsymbol{\theta}^*$, we can write $\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})=\boldsymbol{\theta}^*+\boldsymbol{\zeta}$, with $\|\boldsymbol{\zeta}\|_{\mathrm{F}}=o_P(1)$. Using this new notation, we can rewrite $\ell(\bar{\boldsymbol{\Omega}})$ as:

$$\begin{aligned}\ell(\bar{\boldsymbol{\Omega}}) &= -\frac{1}{2}\log|\bar{\boldsymbol{\Omega}}|+\frac{1}{2}\mathrm{tr}\left(\bar{\boldsymbol{\Omega}}\frac{\sum_t\mathbf{x}_t\mathbf{e}_t^\top}{T}\right)-\frac{1}{2}\mathrm{tr}\left(\left(\frac{\sum_t\mathbf{x}_t^\top\bar{\boldsymbol{\Omega}}\mathbf{y}_t}{T}\right)\boldsymbol{\zeta}\right) \\ &= \widetilde{\ell}(\bar{\boldsymbol{\Omega}})-\frac{1}{2}\mathrm{tr}\left(\left(\frac{\sum_t\mathbf{x}_t^\top\bar{\boldsymbol{\Omega}}\mathbf{y}_t}{T}\right)\boldsymbol{\zeta}\right),\end{aligned} \tag{S4.17}$$

where we define the first two terms in (S4.17) as $\widetilde{\ell}(\bar{\boldsymbol{\Omega}})$.

By the Cauchy-Schwartz inequality, we have:

$$\left|\frac{1}{2}\mathrm{tr}\left(\left(\frac{\sum_t\mathbf{x}_t^\top\bar{\boldsymbol{\Omega}}\mathbf{y}_t}{T}\right)\boldsymbol{\zeta}\right)\right|\leq\frac{1}{2}\left\|\frac{\sum_t\mathbf{x}_t^\top\bar{\boldsymbol{\Omega}}\mathbf{y}_t}{T}\right\|_{\mathrm{F}}\cdot\|\boldsymbol{\zeta}\|_{\mathrm{F}}. \tag{S4.18}$$

By the definition of $\mathbf{y}_t$, we have:

$$\frac{\sum_t\mathbf{x}_t^\top\bar{\boldsymbol{\Omega}}\mathbf{y}_t}{T}=\left[\left(\frac{\sum_t\mathbf{x}_{t-1}\otimes\mathbf{x}_t}{T}\right)^\top\left(\mathbf{I}_S\otimes\bar{\boldsymbol{\Omega}}\right);\left(\frac{\sum_t\mathbf{z}_{t-1}\otimes\mathbf{x}_t}{T}\right)^\top\left(\mathbf{I}_D\otimes\bar{\boldsymbol{\Omega}}\mathbf{K}\right)\right],$$

and notice that $\mathbf{x}_{t-1}\otimes\mathbf{x}_t$ and $\mathbf{z}_{t-1}\otimes\mathbf{x}_t$ are just rearranged versions of $\mathbf{x}_t\mathbf{x}_{t-1}^\top$ and $\mathbf{x}_t\mathbf{z}_{t-1}^\top$, respectively. Therefore, by the joint stationarity of $\mathbf{x}_t$ and $\mathbf{z}_t$, we have the time average of $\mathbf{x}_{t-1}\otimes\mathbf{x}_t$ and $\mathbf{z}_{t-1}\otimes\mathbf{x}_t$ converging to the rearranged version of some constant auto-covariance matrices and therefore we have the term on the right-hand side of (S4.18) being $o_P(1)$.

Given this argument, proving (S4.16) is now equivalent to proving:

$$\mathrm{P}\left(\inf_{\|\bar{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^*\|_{\mathrm{F}}\geq c}\widetilde{\ell}(\bar{\boldsymbol{\Omega}})\leq\widetilde{\ell}(\boldsymbol{\Omega}^*)\right)\to 0,\ \text{as } T\to\infty. \tag{S4.19}$$

Define $\widetilde{\boldsymbol{\Omega}}$ as the unconstrained minimizer of $\widetilde{\ell}(\boldsymbol{\Omega})$, then explicitly, we have:

$$\begin{aligned}\widetilde{\boldsymbol{\Omega}}=\arg\min_{\boldsymbol{\Omega}}\widetilde{\ell}(\boldsymbol{\Omega}) &= \left(\frac{\sum_t\mathbf{e}_t\mathbf{x}_t^\top}{T}\right)^{-1} \\ &= \left(\frac{\sum_t\mathbf{e}_t\mathbf{e}_t^\top}{T}+\frac{\sum_t\mathbf{e}_t\left(\mathbf{y}_t\boldsymbol{\theta}^*\right)^\top}{T}\right)^{-1}\xrightarrow{p}\boldsymbol{\Omega}^*,\end{aligned}$$

where the final argument on the convergence in probability to $\mathbf{\Omega}^*$ is based on the fact that $\sum_{t=1}^{T} \mathbf{e}_t \left(\mathbf{y}_t \boldsymbol{\theta}^*\right)^\top / T = O_P(T^{-1/2})$ by the CLT of MDS. By the second-order Taylor expansion of $\widetilde{\ell}(\bar{\mathbf{\Omega}})$ at $\widetilde{\mathbf{\Omega}}$, we have:

$$\widetilde{\ell}(\bar{\mathbf{\Omega}}) = \widetilde{\ell}(\widetilde{\mathbf{\Omega}}) + \frac{1}{4}\mathbf{vec}\left(\bar{\mathbf{\Omega}} - \widetilde{\mathbf{\Omega}}\right)^\top \left[\check{\mathbf{\Omega}}^{-1} \otimes \check{\mathbf{\Omega}}^{-1}\right] \mathbf{vec}\left(\bar{\mathbf{\Omega}} - \widetilde{\mathbf{\Omega}}\right), \tag{S4.20}$$

where $\check{\mathbf{\Omega}} = \widetilde{\mathbf{\Omega}} + \eta(\bar{\mathbf{\Omega}} - \widetilde{\mathbf{\Omega}})$, for some $\eta \in [0, 1]$. For any constant $c > 0$ such that $\|\bar{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{\mathrm{F}} = c$, let $c = \kappa \bar{\rho}(\mathbf{\Omega}^*)$, where $\kappa > 0$ is also a constant that relates to c only. Consequently, we have:

$$\left|\bar{\rho}(\bar{\mathbf{\Omega}}) - \bar{\rho}(\mathbf{\Omega}^*)\right| \leq \|\bar{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_s \leq \|\bar{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{\mathrm{F}} = \kappa \bar{\rho}(\mathbf{\Omega}^*),$$

and thus $\bar{\rho}(\bar{\mathbf{\Omega}}) \leq (1 + \kappa)\bar{\rho}(\mathbf{\Omega}^*)$. Conditioning on the event that $\|\bar{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{\mathrm{F}} = c$, we first have $\|\bar{\mathbf{\Omega}} - \widetilde{\mathbf{\Omega}}\|_{\mathrm{F}} \geq c/2$ to hold with probability approaching one, due to the consistency of $\widetilde{\mathbf{\Omega}}$. Furthermore, we also have:

$$\begin{aligned}
\underline{\rho}(\check{\mathbf{\Omega}}^{-1} \otimes \check{\mathbf{\Omega}}^{-1}) = \underline{\rho}(\check{\mathbf{\Omega}}^{-1})^2 &= \frac{1}{\bar{\rho}(\check{\mathbf{\Omega}})^2} \\
&\geq \left[\frac{1}{\bar{\rho}(\widetilde{\mathbf{\Omega}}) + \bar{\rho}(\bar{\mathbf{\Omega}})}\right]^2 \\
&\geq \left[\frac{1}{2\bar{\rho}(\mathbf{\Omega}^*) + (c + \bar{\rho}(\mathbf{\Omega}^*))}\right]^2 = \frac{1}{(3+\kappa)^2} \cdot \frac{1}{\bar{\rho}(\mathbf{\Omega}^*)^2},
\end{aligned}$$

where the last inequality holds with probability approaching one since $\mathrm{P}\left[\bar{\rho}(\widetilde{\mathbf{\Omega}}) \leq 2\bar{\rho}(\mathbf{\Omega}^*)\right] \to 1$. Utilizing these facts together with (S4.20), we end up having:

$$\mathrm{P}\left[\widetilde{\ell}(\bar{\mathbf{\Omega}}) \geq \widetilde{\ell}(\widetilde{\mathbf{\Omega}}) + \frac{1}{16} \cdot \left(\frac{\kappa}{3+\kappa}\right)^2\right] \to 1, \text{ as } T \to \infty, \tag{S4.21}$$

for any $\bar{\mathbf{\Omega}}$ such that $\|\bar{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{\mathrm{F}} = c = \kappa \bar{\rho}(\mathbf{\Omega}^*)$. Since κ is an arbitrary positive constant and $\widetilde{\ell}(\widetilde{\mathbf{\Omega}}) \xrightarrow{p} \widetilde{\ell}(\mathbf{\Omega}^*)$, we establish (S4.19) and thereby completes the proof. □

S4.2 Proof of Theorem 2

To prove Theorem 2, we first establish the consistency and the convergence rate of the estimators in Lemma S4.1 below.

Lemma S4.1. *Under the same assumption as Theorem 2, all model estimators for MARAC are $\sqrt{T}$-consistent, namely:*

$$\|\widehat{\mathbf{A}}_p - \mathbf{A}_p^*\|_{\mathrm{F}} = O_P\left(\frac{1}{\sqrt{T}}\right), \|\widehat{\mathbf{B}}_p - \mathbf{B}_p^*\|_{\mathrm{F}} = O_P\left(\frac{1}{\sqrt{T}}\right), \|\widehat{\boldsymbol{\gamma}}_q - \boldsymbol{\gamma}_q^*\|_{\mathrm{F}} = O_P\left(\frac{1}{\sqrt{T}}\right),$$

for $p \in [P], q \in [Q]$. As a direct result, we also have:

$$\|\widehat{\mathbf{B}}_p \otimes \widehat{\mathbf{A}}_p - \mathbf{B}_p^* \otimes \mathbf{A}_p^*\|_{\mathrm{F}} = O_P\left(\frac{1}{\sqrt{T}}\right), \text{ for } p \in [P].$$

We delay the proof of Lemma S4.1 to Section S6.2. With this lemma, we are now ready to present the proof of Theorem 2.

Proof. For the simplicity of notation and presentation, we fix P, Q as 1, but the proving technique can be generalized to arbitrary P, Q. To start with, we revisit the updating rule for $\mathbf{A}_p^{(l+1)}$ in (S1.2). By plugging in the data-generating model for $\mathbf{X}_t$ according to MARAC$(1,1)$ model, we can transform (S1.2) into:

$$\sum_{t\in[T]} \left[\Delta\mathbf{A}_1\mathbf{X}_{t-1}\widehat{\mathbf{B}}_1^\top + \mathbf{A}_1^*\mathbf{X}_{t-1}\Delta\mathbf{B}_1^\top + \Delta\boldsymbol{\mathcal{G}}_1\bar{\times}\mathbf{z}_{t-1} - \mathbf{E}_t\right]\widehat{\boldsymbol{\Sigma}}_c^{-1}\widehat{\mathbf{B}}_1\mathbf{X}_{t-1}^\top = \mathbf{O}_{M\times M},$$

where for any arbitrary matrix/tensor $\mathbf{M}$, we define $\Delta\mathbf{M}$ as $\Delta\mathbf{M} = \widehat{\mathbf{M}} - \mathbf{M}^*$. One can simplify the estimating equation above by left multiplying $\widehat{\boldsymbol{\Sigma}}_r^{-1}$ and then vectorize both

sides to obtain:

$$\begin{aligned}
&\sum_{t\in[T]}\left[(\mathbf{B}_1^*\mathbf{X}_{t-1}^\top)^\top(\mathbf{\Sigma}_c^*)^{-1}(\mathbf{B}_1^*\mathbf{X}_{t-1}^\top)\otimes(\mathbf{\Sigma}_r^*)^{-1}\right]\mathbf{vec}\left(\widehat{\mathbf{A}}_1-\mathbf{A}_1^*\right)\\
+&\sum_{t\in[T]}\left[(\mathbf{B}_1^*\mathbf{X}_{t-1}^\top)^\top(\mathbf{\Sigma}_c^*)^{-1}\otimes(\mathbf{\Sigma}_r^*)^{-1}\mathbf{A}_1^*\mathbf{X}_{t-1}\right]\mathbf{vec}\left(\widehat{\mathbf{B}}_1^\top-(\mathbf{B}_1^*)^\top\right)\\
+&\sum_{t\in[T]}\left\{\mathbf{z}_{t-1}^\top\otimes\left[(\mathbf{B}_1^*\mathbf{X}_{t-1}^\top)^\top(\mathbf{\Sigma}_c^*)^{-1}\otimes(\mathbf{\Sigma}_r^*)^{-1}\mathbf{K}\right]\right\}\mathbf{vec}\left(\widehat{\boldsymbol{\gamma}}_1-\boldsymbol{\gamma}_1^*\right)\\
=&\sum_{t\in[T]}\left[(\mathbf{B}_1^*\mathbf{X}_{t-1}^\top)^\top(\mathbf{\Sigma}_c^*)^{-1}\otimes(\mathbf{\Sigma}_r^*)^{-1}\right]\mathbf{vec}\left(\mathbf{E}_t\right)+o_P(\sqrt{T}).
\end{aligned}$$

On the left-hand side of the equation above, we replace $\widehat{\mathbf{B}}_1, \widehat{\mathbf{\Sigma}}_r, \widehat{\mathbf{\Sigma}}_c$ with their true values $\mathbf{B}_1^*, \mathbf{\Sigma}_r^*, \mathbf{\Sigma}_c^*$, since the discrepancies are of order $o_P(1)$ and can thus be incorporated into the $o_P(\sqrt{T})$ term given the $\sqrt{T}$-consistency of $\widehat{\mathbf{A}}_1, \widehat{\mathbf{B}}_1, \widehat{\boldsymbol{\gamma}}_1$. On the right-hand side, we have:

$$\begin{aligned}
&\sum_t \mathbf{vec}\left(\widehat{\mathbf{\Sigma}}_r^{-1}\mathbf{E}_t\widehat{\mathbf{\Sigma}}_c^{-1}\widehat{\mathbf{B}}_1\mathbf{X}_{t-1}^\top\right)\\
&=\sum_t\left[\mathbf{e}_t^\top\otimes(\mathbf{X}_{t-1}\otimes\mathbf{I}_M)\right]\mathbf{vec}\left[\left(\widehat{\mathbf{B}}_1^\top\otimes\mathbf{I}_M\right)\widehat{\mathbf{\Sigma}}^{-1}\right],
\end{aligned}$$

where the process $\{\mathbf{e}_t^\top\otimes(\mathbf{X}_{t-1}\otimes\mathbf{I}_M)\}_{t=1}^T$ is a martingale difference sequence and the martingale central limit theorem (Hall and Heyde, 2014) implies that $\sum_t\left[\mathbf{e}_t^\top\otimes(\mathbf{X}_{t-1}\otimes\mathbf{I}_M)\right]=O_P(\sqrt{T})$, and thus by the consistency of $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{B}}_1$, we can replace $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{B}}_1$ with their true values and incorporate the remainders into $o_P(\sqrt{T})$.

Similar transformations can be applied to (S1.3) and (S1.4), where the penalty term is incorporated into $o_P(\sqrt{T})$ due to the assumption that $\lambda=o(T^{-\frac{1}{2}})$. With the notation that $\mathbf{U}_t=\mathbf{I}_N\otimes\mathbf{A}_1^*\mathbf{X}_{t-1}$, $\mathbf{V}_t=\mathbf{B}_1^*\mathbf{X}_{t-1}^\top\otimes\mathbf{I}_M$, $\mathbf{Y}_t=\mathbf{z}_{t-1}^\top\otimes\mathbf{K}$ and $\mathbf{W}_t=[\mathbf{V}_t;\mathbf{U}_t;\mathbf{Y}_t]$, these transformed estimating equations can be converted altogether into:

$$\begin{aligned}
\left(\frac{1}{T}\sum_{t\in[T]}\mathbf{W}_t^\top(\mathbf{\Sigma}^*)^{-1}\mathbf{W}_t\right)\mathbf{vec}\left(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*\right)&=\frac{1}{T}\sum_{t\in[T]}\mathbf{W}_t^\top(\mathbf{\Sigma}^*)^{-1}\mathbf{vec}\left(\mathbf{E}_t\right)\\
&\quad+o_P(T^{-1/2}),
\end{aligned}\tag{S4.22}$$

where $\mathbf{vec}\left(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\right)=[\mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{A}}}-\boldsymbol{\mathcal{A}}^*\right)^\top,\mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{B}}}-\boldsymbol{\mathcal{B}}^*\right)^\top,\mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{R}}}-\boldsymbol{\mathcal{R}}^*\right)^\top]^\top$, and $\widehat{\boldsymbol{\mathcal{A}}},\widehat{\boldsymbol{\mathcal{B}}},\widehat{\boldsymbol{\mathcal{R}}}$ are defined as $[\widehat{\boldsymbol{\mathcal{A}}}]_{::p}=\widehat{\mathbf{A}}_p$, $[\widehat{\boldsymbol{\mathcal{B}}}]_{::p}=\widehat{\mathbf{B}}_p^\top$, $[\widehat{\boldsymbol{\mathcal{R}}}]_{:dq}=\widehat{\boldsymbol{\gamma}}_{q,d}$ and $\boldsymbol{\mathcal{A}}^*,\boldsymbol{\mathcal{B}}^*,\boldsymbol{\mathcal{R}}^*$ are the corresponding true coefficients.

In (S4.22), we first establish that:

$$(1/T)\sum_{t\in[T]}\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t\xrightarrow{p.}\mathrm{E}\left[\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t\right]. \tag{S4.23}$$

To prove S4.23, by the assumption that $\mathbf{X}_t$ and $\mathbf{z}_t$ are zero-meaned and jointly stationary, we have $T^{-1}\sum_{t\in[T]}\widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^\top\xrightarrow{p.}\mathrm{E}[\widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^\top]$ by Lemma S6.1 and Corollary S6.2, where $\widetilde{\mathbf{x}}_t=[\mathbf{x}_t^\top,\mathbf{z}_t^\top]^\top$. See details of Lemma S6.1 and Corollary S6.2 in Section S6.1. Then since each element of $\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t$ is a linear combination of terms in $\widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^\top$ (thus a continuous mapping), it is straightforward that (S4.23) holds elementwise.

Given (S4.23) and the fact that $\widehat{\boldsymbol{\Theta}}$ is $\sqrt{T}$-consistent, we can rewrite (S4.22) as:

$$\begin{aligned}\mathrm{E}\left[\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t\right]\mathbf{vec}\left(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\right)=&\frac{1}{T}\sum_{t\in[T]}\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{vec}\left(\mathbf{E}_t\right)\\&+o_P(T^{-1/2}),\end{aligned} \tag{S4.24}$$

For the term on the right-hand side of (S4.24), first notice that the sequence $\{\boldsymbol{\eta}_t\}_{t=1}^T$, where $\boldsymbol{\eta}_t=\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{vec}\left(\mathbf{E}_t\right)$, is a zero-meaned, stationary vector martingale difference sequence (MDS), thanks to the independence of $\mathbf{E}_t$ from the jointly stationary $\mathbf{X}_{t-1}$ and $\mathbf{z}_{t-1}$. By the martingale central limit theorem (Hall and Heyde, 2014), we have:

$$\frac{1}{\sqrt{T}}\sum_{t\in[T]}\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{vec}\left(\mathbf{E}_t\right)\xrightarrow{d.}\mathcal{N}(\mathbf{0},\mathrm{E}\left[\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t\right]). \tag{S4.25}$$

Combining (S4.24) and (S4.25), we end up having:

$$\mathrm{E}\left[\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t\right]\sqrt{T}\mathbf{vec}\left(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\right)\xrightarrow{d.}\mathcal{N}(\mathbf{0},\mathrm{E}\left[\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t\right]). \tag{S4.26}$$

The asymptotic distribution of $\sqrt{T}\mathbf{vec}\left(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*\right)$ can thus be derived by multiplying both sides of (S4.26) by the inverse of $\mathbf{L}=\mathrm{E}\left[\mathbf{W}_t^\top(\mathbf{\Sigma}^*)^{-1}\mathbf{W}_t\right]$. However, the matrix $\mathbf{L}$ is not a full-rank matrix, because $\mathbf{L}\boldsymbol{\mu}=\mathbf{0}$, where $\boldsymbol{\mu}=[\mathbf{vec}\left(\boldsymbol{\mathcal{A}}^*\right)^\top,-\mathbf{vec}\left(\boldsymbol{\mathcal{B}}^*\right)^\top,\mathbf{0}^\top]^\top$. As a remedy, let $\boldsymbol{\zeta}=[\mathbf{vec}\left(\mathbf{A}_1^*\right)^\top\mathbf{0}^\top]^\top\in\mathbb{R}^{M^2+N^2+DMN}$, then given the identifiability constraint that $\|\mathbf{A}_1^*\|_{\mathrm{F}}=\|\widehat{\mathbf{A}}_1\|_{\mathrm{F}}=1$ and the fact that $\widehat{\mathbf{A}}_1$ is $\sqrt{T}$-consistent, we have $\mathbf{vec}\left(\mathbf{A}_1^*\right)^\top\mathbf{vec}\left(\widehat{\mathbf{A}}_1-\mathbf{A}_1^*\right)=o_P(T^{-1/2})$. Therefore, we have:

$$\sqrt{T}\boldsymbol{\zeta}^\top\mathbf{vec}\left(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*\right)\xrightarrow{p.}0. \tag{S4.27}$$

Combining (S4.26) and (S4.27) and using the Slutsky's theorem, we have $\mathbf{H}\sqrt{T}\mathbf{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*)\xrightarrow{d.}\mathcal{N}(\mathbf{0},\mathbf{L})$, where $\mathbf{H}=\mathbf{L}+\boldsymbol{\zeta}\boldsymbol{\zeta}^\top$ and thus:

$$\sqrt{T}\mathbf{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*)\xrightarrow{d.}\mathcal{N}(\mathbf{0},\mathbf{H}^{-1}\mathbf{L}\mathbf{H}^{-1}). \tag{S4.28}$$

The final asymptotic distribution of $\mathrm{vec}(\widehat{\mathbf{B}}_1^\top)\otimes\mathrm{vec}(\widehat{\mathbf{A}}_1)$ and $\mathbf{K}\widehat{\boldsymbol{\gamma}}_{q,d}$ can be derived easily from (S4.28) with the multivariate delta method, and we omit the details here. □

S4.3 Proof of Corollary 1

Proof. Based on the asymptotic distribution of the MARAC model estimators in (4.3), it is straightforward that the marginal asymptotic distribution of $\widehat{\boldsymbol{\mathcal{G}}}_1,\ldots,\widehat{\boldsymbol{\mathcal{G}}}_Q$ follows:

$$\sqrt{T}\begin{bmatrix}\mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{G}}}_1-\boldsymbol{\mathcal{G}}_1^*\right)\\ \cdots\\ \mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{G}}}_Q-\boldsymbol{\mathcal{G}}_Q^*\right)\end{bmatrix}\xrightarrow{d.}\mathcal{N}\left(\mathbf{0},\left[\mathbf{O}:\mathbf{I}_{QD}\otimes\mathbf{K}\right]\mathbf{\Xi}\begin{bmatrix}\mathbf{O}\\ \\ \mathbf{I}_{QD}\otimes\mathbf{K}\end{bmatrix}\right). \tag{S4.29}$$

Unwrapping the matrix $\mathbf{\Xi}$, one can simplify the asymptotic variance in (S4.29) as:

$$\mathbf{\Psi}\coloneqq\left[\mathbf{O}:\mathbf{I}_{QD}\otimes\mathbf{K}\right]\mathbf{\Xi}\begin{bmatrix}\mathbf{O}\\ \\ \mathbf{I}_{QD}\otimes\mathbf{K}\end{bmatrix}=(\mathbf{I}\otimes\mathbf{K})\left[\mathbf{D}-\mathbf{C}\boldsymbol{\gamma}\left(\mathbf{C}\boldsymbol{\gamma}\right)^\top\right](\mathbf{I}\otimes\mathbf{K}),$$

where $\mathbf{D}$ is the lower-right $MNQD \times MNQD$ block of $\mathbf{H}^{-1}$, and $\mathbf{C}$ is the lower-left block under the same block partition. To estimate the rank of matrix $\mathbf{\Psi}$, it is sufficient to estimate the rank of $\mathbf{D}$, as $\mathbf{I} \otimes \mathbf{K}$ is full-rank, and $\mathbf{C}\boldsymbol{\gamma} \left(\mathbf{C}\boldsymbol{\gamma}\right)^\top$ is rank-1. Note that matrix $\mathbf{H}$ is full-rank, and the top-left block of $\mathbf{H}$, denoted as $\mathbf{H}_{11}$, is:

$$\mathrm{E}\left\{\begin{bmatrix} \ddots & \cdots & \iddots & \ddots & \cdots & \iddots \\ \cdot & \mathbf{X}_{t-i}\mathbf{B}_i^\top\mathbf{\Sigma}_c^{-1}\mathbf{B}_j\mathbf{X}_{t-j}^\top \otimes \mathbf{\Sigma}_r^{-1} & \cdots & \cdots & \mathbf{X}_{t-i}\mathbf{B}_i^\top\mathbf{\Sigma}_c^{-1} \otimes \mathbf{\Sigma}_r^{-1}\mathbf{A}_j\mathbf{X}_{t-j} & \cdots \\ \iddots & \cdots & \ddots & \iddots & \cdots & \ddots \\ \ddots & \cdots & \iddots & \ddots & \cdots & \iddots \\ \cdot & \mathbf{\Sigma}_c^{-1}\mathbf{B}_i\mathbf{X}_{t-i}^\top \otimes \mathbf{X}_{t-j}^\top\mathbf{A}_j^\top\mathbf{\Sigma}_r^{-1} & \cdots & \cdots & \mathbf{\Sigma}_c^{-1} \otimes \mathbf{X}_{t-i}^\top\mathbf{A}_i^\top\mathbf{\Sigma}_r^{-1}\mathbf{A}_j\mathbf{X}_{t-j} & \cdots \\ \iddots & \cdots & \ddots & \iddots & \cdots & \ddots \end{bmatrix}\right\} + \boldsymbol{\alpha}\boldsymbol{\alpha}^\top,$$

where $1 \leq i, j \leq P$, and $\boldsymbol{\alpha} = [\mathbf{vec}\left(\mathbf{A}_1\right)^\top, \ldots, \mathbf{vec}\left(\mathbf{A}_P\right)^\top, \mathbf{0}^\top]^\top$. Here, all model parameters are the ground truth values, and we omit the asterisk notation for simplicity. This matrix is the key component of the asymptotic variance of the MAR(P) model, see Theorem 3 of Chen et al. (2021), and is thus invertible. Consequently, the Schur complement of $\mathbf{H}$ is invertible and thus $\mathbf{D}$ is a full-rank matrix. Therefore, we have $\mathrm{rank}(\mathbf{\Psi}) \geq MNQD - 1$.

Finally, based on (S4.29), we have $T \cdot (\widehat{\mathbf{g}} - \mathbf{g}^*)^\top \mathbf{\Psi}^\dagger (\widehat{\mathbf{g}} - \mathbf{g}^*) \xrightarrow{d.} \chi^2_r$, where $r = \mathrm{rank}(\mathbf{\Psi}) \geq MNQD - 1$, and thus completes the proof. In practice, when we utilize this result to test the hypothesis of $\mathbf{g}^* = \mathbf{0}$, we will plug in the estimator of all parameters and compute the test statistics $T \cdot \widehat{\mathbf{g}}^\top \widehat{\mathbf{\Psi}}^\dagger \widehat{\mathbf{g}}$, and set the critical region based on χ^2_{MNQD-1}. □

S5 Theory under High Spatial Dimension

S5.1 Proof of Theorem 3

Proof. In this proof, we will fix P, Q as 1 again for the ease of presentation, but the technical details can be generalized to arbitrary P, Q. Since we fix the lags to be 1, we drop the subscript of the coefficients for convenience.

Under the specification of the MARAC$(1,1)$ model, we restate the model as:

$$\mathbf{x}_t = \left(\mathbf{x}_{t-1}^\top \otimes \mathbf{I}_S\right) \mathbf{vec}\left(\mathbf{B}^* \otimes \mathbf{A}^*\right) + \left(\mathbf{z}_{t-1}^\top \otimes \mathbf{K}\right) \boldsymbol{\gamma}^* + \mathbf{e}_t,$$

where $S = MN$ and we introduce the following additional notations:

$$\mathbf{Y}_T \coloneqq \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_T \end{bmatrix}, \quad \widetilde{\mathbf{X}}_T \coloneqq \begin{bmatrix} \mathbf{x}_0^\top \\ \vdots \\ \mathbf{x}_{T-1}^\top \end{bmatrix} \otimes \mathbf{I}_S, \quad \widetilde{\mathbf{z}}_T \coloneqq \begin{bmatrix} \mathbf{z}_0^\top \\ \vdots \\ \mathbf{z}_{T-1}^\top \end{bmatrix}, \quad \boldsymbol{\mathcal{E}}_T = \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_T \end{bmatrix}.$$

We will drop the subscript T for convenience. Let $\boldsymbol{\phi}^* = \mathbf{vec}\left(\mathbf{B}^* \otimes \mathbf{A}^*\right)$, and $g_1^*, \ldots, g_D^* \in \mathbb{H}_k$ be the true autoregressive and functional parameters. Correspondingly, let $\boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_D^*$ be the coefficients for the representers when evaluating $g_1^*, \ldots, g_D^*$ on a matrix grid, i.e., $\mathbf{K}\boldsymbol{\gamma}_d^*$ is a discrete evaluation of g_d^* on the matrix grid. Let $\mathbb{F}_{\boldsymbol{\phi}} = \{\mathbf{vec}\left(\mathbf{B} \otimes \mathbf{A}\right) | \|\mathbf{A}\|_{\mathrm{F}} = \mathrm{sign}(\mathrm{tr}\left(\mathbf{A}\right)) = 1, \mathbf{A} \in \mathbb{R}^{M \times M}, \mathbf{B} \in \mathbb{R}^{N \times N}\}$. Using these new notations, the MARAC estimator is obtained by solving the following penalized least squares problem:

$$\min_{\boldsymbol{\phi} \in \mathbb{F}_{\boldsymbol{\phi}}, \boldsymbol{\gamma} \in \mathbb{R}^{SD}} \mathfrak{L}_\lambda(\boldsymbol{\phi}, \boldsymbol{\gamma}) \coloneqq \left\{ \frac{1}{2T} \|\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\phi} - \left(\widetilde{\mathbf{z}} \otimes \mathbf{K}\right)\boldsymbol{\gamma}\|_{\mathrm{F}}^2 + \frac{\lambda}{2} \boldsymbol{\gamma}^\top \left(\mathbf{I}_D \otimes \mathbf{K}\right) \boldsymbol{\gamma} \right\}. \tag{S5.30}$$

By fixing $\boldsymbol{\phi}$, the estimator for $\boldsymbol{\gamma}$ is given by $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\phi}) = \arg\min_{\boldsymbol{\gamma}} \mathfrak{L}_\lambda(\boldsymbol{\phi}, \boldsymbol{\gamma})$, and can be explicitly written as:

$$\widehat{\boldsymbol{\gamma}}(\boldsymbol{\phi}) = T^{-1} \left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}} \otimes \mathbf{K} + \lambda \cdot \mathbf{I}_{SD}\right)^{-1} \left(\widetilde{\mathbf{z}}^\top \otimes \mathbf{I}_S\right) \left(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\phi}\right). \tag{S5.31}$$

Plugging (S5.31) into (S5.30) yields the profile likelihood for $\boldsymbol{\phi}$:

$$\ell_\lambda(\boldsymbol{\phi}) = \mathfrak{L}_\lambda(\boldsymbol{\phi}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\phi})) = \frac{1}{2T}\left(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\phi}\right)^\top \mathbf{W}\left(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\phi}\right), \tag{S5.32}$$

where $\mathbf{W}$ is defined as:

$$\mathbf{W} = \left\{\mathbf{I} - \frac{(\widetilde{\mathbf{z}} \otimes \mathbf{K})\left[\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}} \otimes \mathbf{K} + \lambda \cdot \mathbf{I}_{SD}\right]^{-1}(\widetilde{\mathbf{z}}^\top \otimes \mathbf{I}_S)}{T}\right\} = \left(\mathbf{I} + \frac{\widetilde{\mathbf{z}}\widetilde{\mathbf{z}}^\top}{\lambda T} \otimes \mathbf{K}\right)^{-1}, \tag{S5.33}$$

and the second equality in (S5.33) is by the Woodbury matrix identity. It can be seen that $\mathbf{W}$ is positive semi-definite and has all of its eigenvalues within $(0, 1)$. To improve the clarity and organization of the proof, we break down the proof into several major steps. In the first step, we establish the following result on $\widehat{\boldsymbol{\phi}}$:

Proposition S5.1. *Under the assumptions of Theorem 3, we have:*

$$\left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\right)^\top \left(\frac{\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}}{T}\right)\left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\right) \lesssim O_P(C_g\lambda) + O_P(c_{1,S} \cdot SD/T), \tag{S5.34}$$

where $C_g = \sum_{d=1}^D \|g_d^*\|_{\mathbb{H}_k}^2$.

In order to derive the convergence rate of $\widehat{\boldsymbol{\phi}}$, we still require one additional result:

Lemma S5.2. *Under the assumptions of Theorem 3 and the requirement that* $S \log S/T \to 0$*, it holds that:*

$$\underline{\rho}\left(\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}/T\right) \geq \frac{c_{0,S}}{2} > 0, \tag{S5.35}$$

with probability approaching 1 *as* $S, T \to \infty$*, where* $\underline{\rho}(\cdot)$ *is the minimum eigenvalue of a matrix and* $c_{0,S} = \underline{\rho}(\boldsymbol{\Sigma}^*_{\mathbf{x},\mathbf{x}} - \left(\boldsymbol{\Sigma}^*_{\mathbf{z},\mathbf{x}}\right)^\top \left(\boldsymbol{\Sigma}^*_{\mathbf{z},\mathbf{z}}\right)^{-1} \boldsymbol{\Sigma}^*_{\mathbf{z},\mathbf{x}})$.

The proof of Proposition S5.1 and Lemma S5.2 are relegated to Section S5.2 and S6.3, respectively. Combining Proposition S5.1 and Lemma S5.2, we can derive the error bound of $\widehat{\boldsymbol{\phi}}$ as:

$$\frac{1}{S}\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\|_{\mathrm{F}} \lesssim O_P(\sqrt{\frac{C_g\gamma_S}{c_{0,S}S}}) + O_P(\sqrt{\frac{c_{1,S}D}{c_{0,S}TS}}). \tag{S5.36}$$

Now with this error bound of the autoregressive parameter $\widehat{\boldsymbol{\phi}}$, we further derive the prediction error bound for the functional parameters. To start with, we have:

$$\begin{aligned}\frac{1}{\sqrt{TS}}\|(\widetilde{\mathbf{z}}\otimes\mathbf{K})(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\|_{\mathrm{F}} &= \frac{1}{\sqrt{TS}}\left\|(\mathbf{I}-\mathbf{W})(\mathbf{Y}-\widetilde{\mathbf{X}}\widehat{\boldsymbol{\phi}})-(\widetilde{\mathbf{z}}\otimes\mathbf{K})\boldsymbol{\gamma}^*\right\|_{\mathrm{F}} \\ &\leq \frac{1}{\sqrt{TS}}\Bigg[\underbrace{\|(\mathbf{I}-\mathbf{W})\boldsymbol{\mathcal{E}}\|_{\mathrm{F}}}_{J_1}+\underbrace{\|(\mathbf{I}-\mathbf{W})\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*)\|_{\mathrm{F}}}_{J_2} \\ &\quad+\underbrace{\|\mathbf{W}(\widetilde{\mathbf{z}}\otimes\mathbf{K})\boldsymbol{\gamma}^*\|_{\mathrm{F}}}_{J_3}\Bigg],\end{aligned}$$

and we will bound the terms J_1, J_2, J_3 separately.

To bound J_1, we first establish two lemmas.

Lemma S5.3. *Given the definition of* $\mathbf{W}$ *in* (S5.33) *and under the assumptions of Theorem 3, we have* $O_P(\gamma_S^{-1/2r_0}) \leq tr(\mathbf{I}-\mathbf{W}) \leq O_P(\sqrt{S}\gamma_S^{-1/2r_0})$*, where* $\gamma_S = \lambda/S$*. Furthermore, we have* $tr(\mathbf{W}) \leq SD$*.*

Lemma S5.4. *Given the definition of* $\mathbf{W}$ *in* (S5.33) *and under the assumptions of Theorem 3, we have that:*

$$\boldsymbol{\mathcal{E}}^\top\mathbf{W}\boldsymbol{\mathcal{E}}/tr(\mathbf{W}) = O_P(c_{1,S}),$$

where $c_{1,S} = \|\boldsymbol{\Sigma}\|_s$*. Furthermore, we have* $\boldsymbol{\mathcal{E}}^\top(\mathbf{I}-\mathbf{W})^2\boldsymbol{\mathcal{E}}/tr\left((\mathbf{I}-\mathbf{W})^2\right) = O_P(c_{1,S})$*.*

We leave the proof of Lemma S5.3 and Lemma S5.4 to Section S6.4 and S6.5. By Lemma S5.4, we have:

$$J_1^2 \asymp c_{1,S}\cdot\mathrm{tr}\left((\mathbf{I}-\mathbf{W})^2\right) \lesssim c_{1,S}\cdot\mathrm{tr}\left(\mathbf{I}-\mathbf{W}\right).$$

And by Lemma S5.3, we have $J_1 \leq O_P(c_{1,S}^{1/2}\cdot S^{1/4}\gamma_S^{-1/4r_0})$.

For J_2, we have the following bound:

$$J_2 = \|(\mathbf{I} - \mathbf{W})\mathbf{W}^{-1/2}\mathbf{W}^{1/2}\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*)\|_{\mathrm{F}} \tag{S5.37}$$

$$\leq \|(\mathbf{I} - \mathbf{W})\mathbf{W}^{-1/2}\|_s \cdot \|\mathbf{W}^{1/2}\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*)\|_{\mathrm{F}}$$

$$\leq \|\mathbf{W}^{-1/2}\|_s \cdot \|\mathbf{W}^{1/2}\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*)\|_{\mathrm{F}}. \tag{S5.38}$$

To bound $\|\mathbf{W}^{-1/2}\|_s$, we can take advantage of the simpler form of $\mathbf{W}$ using the Woodbury matrix identity in (S5.33) and obtain:

$$\|\mathbf{W}^{-1/2}\|_s = \bar{\rho}(\mathbf{W}^{-1})^{\frac{1}{2}} = \bar{\rho}\left(\mathbf{I} + (\lambda T)^{-1}\widetilde{\mathbf{z}}\widetilde{\mathbf{z}}^\top \otimes \mathbf{K}\right)^{\frac{1}{2}}$$

$$\leq \left[1 + \lambda^{-1}\bar{\rho}(\mathbf{K})\bar{\rho}(T^{-1}\widetilde{\mathbf{z}}\widetilde{\mathbf{z}}^\top)\right]^{\frac{1}{2}} \leq \left[1 + \lambda^{-1}\bar{\rho}(\mathbf{K})\mathrm{tr}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\right)\right]^{\frac{1}{2}}.$$

In Lemma S6.1, which we state later in Section S6.1, we have shown that for N-dimensional stationary vector autoregressive process, the covariance estimator is consistent in the spectral norm as long as $N \log N/T \to 0$. Therefore, since $\{\mathbf{z}_t\}_{t=1}^T$ follows a stationary VAR($\widetilde{Q}$) process and its dimensionality D is fixed, we have $\|\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}} - \boldsymbol{\Sigma}_{\mathbf{z}}^*\|_s \xrightarrow{p} 0$ and thus with probability approaching 1, we have $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}) \leq 2\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{z}}^*)$. Therefore, we have $\|\mathbf{W}^{-1/2}\|_s \leq O_P(\sqrt{1 + c_0/\lambda})$, where c_0 is a constant related to $\mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathbf{z}}^*\right)$ and $\bar{\rho}(\mathbf{K})$. Combining this with the result in Proposition S5.1, we can bound J_2 via its upper bound (S5.38) as:

$$J_2 \leq O_P\left(\sqrt{C_g \lambda T}\right) + O_P\left(\sqrt{C_g T}\right) + O_P(\sqrt{c_{1,S}S}) + O_P\left(\sqrt{c_{1,S}\gamma_S^{-1}}\right). \tag{S5.39}$$

Finally, for J_3, we first notice that:

$$J_3 = \|\mathbf{W}(\widetilde{\mathbf{z}} \otimes \mathbf{K})\boldsymbol{\gamma}^*\|_{\mathrm{F}} \leq \|\mathbf{W}^{1/2}\|_s \cdot \|\mathbf{W}^{1/2}(\widetilde{\mathbf{z}} \otimes \mathbf{K})\boldsymbol{\gamma}^*\|_{\mathrm{F}} \leq \|\mathbf{W}^{1/2}(\widetilde{\mathbf{z}} \otimes \mathbf{K})\boldsymbol{\gamma}^*\|_{\mathrm{F}}.$$

The upper bound of J_3 above can be further bounded by:

$$\begin{aligned}\|\mathbf{W}^{1/2}(\widetilde{\mathbf{z}}\otimes\mathbf{K})\boldsymbol{\gamma}^*\|_{\mathrm{F}}^2 &= (\lambda T)[(\mathbf{I}_D\otimes\mathbf{K})\boldsymbol{\gamma}^*]^\top\left\{\mathbf{I}_{SD}-\left(\lambda^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\otimes\mathbf{K}+\mathbf{I}_{SD}\right)^{-1}\right\}\boldsymbol{\gamma}^* \\ &= (\lambda T)\left(\sum_{d=1}^{D}\|g_d^*\|_{\mathbb{H}_k}^2\right) \\ &\quad -(\lambda^2 T)\,(\boldsymbol{\gamma}^*)^\top\left[(\mathbf{I}_D\otimes\mathbf{K})\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\otimes\mathbf{K}+\lambda\mathbf{I}_{SD}\right)^{-1}\right]\boldsymbol{\gamma}^* \\ &\leq C_g\lambda T, \end{aligned} \tag{S5.40}$$

where $C_g=\sum_{d=1}^{D}\|g_d^*\|_{\mathbb{H}_k}^2$ is the norm of all the underlying functional parameters. The last inequality of (S5.40) follows from the fact that the quadratic form led by $\lambda^2 T$ is non-negative. To see why, first note that:

$$(\mathbf{I}_D\otimes\mathbf{K})\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\otimes\mathbf{K}+\lambda\mathbf{I}_{SD}\right)^{-1}=\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\otimes\mathbf{I}_S\right)^{-1}-\left[\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\otimes\mathbf{I}_S+\lambda^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}^2\otimes\mathbf{K}\right]^{-1}.$$

Then, we have the following lemma:

Lemma S5.5. *If $\mathbf{A},\mathbf{B}$ are symmetric, positive definite real matrices and $\mathbf{A}-\mathbf{B}$ is positive semi-definite, then $\mathbf{B}^{-1}-\mathbf{A}^{-1}$ is also positive semi-definite.*

We leave the proof to Section S6.6. Let $\mathbf{M}=\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\otimes\mathbf{I}_S+\lambda^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}^2\otimes\mathbf{K}$ and $\mathbf{N}=\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\otimes\mathbf{I}_S$, then both $\mathbf{M}$ and $\mathbf{N}$ are positive definite and $\mathbf{M}-\mathbf{N}$ is positive semi-definite. By Lemma S5.5, we have $\mathbf{N}^{-1}-\mathbf{M}^{-1}$ being positive semi-definite and thus (S5.40) holds.

Using the result in (S5.40), we eventually have $J_3\leq O_P(\sqrt{C_g\lambda T})$. Combining all the bounds for J_1,J_2,J_3, we end up with:

$$\begin{aligned}\frac{1}{\sqrt{TS}}\|(\widetilde{\mathbf{z}}\otimes\mathbf{K})(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\|_{\mathrm{F}} &\leq O_P\left(\frac{\sqrt{c_{1,S}}\sqrt{\gamma_S^{-1/2r_0}}}{\sqrt{T}\sqrt[4]{S}}\right)+O_P(\sqrt{\gamma_S}) \\ &\quad +O_P\left(\frac{1}{\sqrt{S}}\right)+O_P\left(\sqrt{\frac{c_{1,S}}{T}}\right)+O_P\left(\frac{\sqrt{c_{1,S}\gamma_S^{-1}}}{\sqrt{TS}}\right).\end{aligned}$$

□

S5.2 Proof of Proposition S5.1

Proof. The MARAC estimator $\widehat{\boldsymbol{\phi}}$ is the minimizer of $\ell_\lambda(\boldsymbol{\phi})$, defined in (S5.32), for all $\boldsymbol{\phi} \in \mathbb{F}_\phi$ and thus $\ell_\lambda(\widehat{\boldsymbol{\phi}}) \leq \ell_\lambda(\boldsymbol{\phi}^*)$. Equivalently, this means that:

$$\frac{1}{2}\left(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*\right)^\top\left(\frac{\widetilde{\mathbf{X}}^\top\mathbf{W}\widetilde{\mathbf{X}}}{T}\right)\left(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*\right) \leq \frac{1}{T}\left[\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^*+\boldsymbol{\mathcal{E}}\right]^\top\mathbf{W}\widetilde{\mathbf{X}}\left(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*\right).$$

Let $\boldsymbol{\delta} = \mathbf{W}^{1/2}\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*)/\sqrt{T}$ and $\boldsymbol{\omega} = \mathbf{W}^{1/2}\left[\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^*+\boldsymbol{\mathcal{E}}\right]/\sqrt{T}$, then the inequality can be simply written as $\boldsymbol{\delta}^\top\boldsymbol{\delta} \leq 2\boldsymbol{\delta}^\top\boldsymbol{\omega}$, and we can upper bound our quantity of interest, namely $\boldsymbol{\delta}^\top\boldsymbol{\delta}$, as:

$$\boldsymbol{\delta}^\top\boldsymbol{\delta} \leq 2(\boldsymbol{\delta}-\boldsymbol{\omega})^\top(\boldsymbol{\delta}-\boldsymbol{\omega}) + 2\boldsymbol{\omega}^\top\boldsymbol{\omega} \leq 4\boldsymbol{\omega}^\top\boldsymbol{\omega}.$$

Therefore, the bound of $\|\boldsymbol{\delta}\|_{\mathrm{F}}^2$ can be obtained via the bound of $\|\boldsymbol{\omega}\|_{\mathrm{F}}^2$. We have the following upper bound for $\|\boldsymbol{\omega}\|_{\mathrm{F}}^2$:

$$\begin{aligned}\|\boldsymbol{\delta}\|_{\mathrm{F}}^2 \leq 4\|\boldsymbol{\omega}\|_{\mathrm{F}}^2 &= \frac{4}{T}\left[\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^*+\boldsymbol{\mathcal{E}}\right]^\top\mathbf{W}\left[\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^*+\boldsymbol{\mathcal{E}}\right] \\ &\leq \frac{8}{T}\left[\underbrace{\|\mathbf{W}^{1/2}\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^*\|_{\mathrm{F}}^2}_{I_1} + \underbrace{\|\mathbf{W}^{1/2}\boldsymbol{\mathcal{E}}\|_{\mathrm{F}}^2}_{I_2}\right], \end{aligned} \tag{S5.41}$$

where the last inequality follows from the fact that $\mathbf{W}$ is positive semi-definite.

For I_1, it can be bounded by (S5.40) and thus $I_1 \leq C_g\lambda T$. To bound I_2, we utilize Lemma S5.4 and bound I_2 as $I_2 \asymp c_{1,S}\cdot\mathrm{tr}\left(\mathbf{W}\right) \leq c_{1,S}\cdot SD$. Combining the bounds for I_1 and I_2, we have:

$$\|\boldsymbol{\delta}\|_{\mathrm{F}}^2 = \left(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*\right)^\top\left(\frac{\widetilde{\mathbf{X}}^\top\mathbf{W}\widetilde{\mathbf{X}}}{T}\right)\left(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*\right) \lesssim O_P(C_g\lambda) + O_P(c_{1,S}\cdot SD/T),$$

which completes the proof. □

S6 Technical Lemmas & Proofs

In this section, we first introduce Lemma S6.1 on the consistency of the covariance matrix estimator for any stationary vector autoregressive process and then Corollary S6.2 on the consistency of the covariance estimator of our MARAC model, given the joint stationarity condition. Then we provide proof for Lemma S4.1 used in Section S4.2 when proving Theorem 2 on the asymptotic normality under fixed spatial dimension. Then we provide proofs for Lemma S5.2, S5.3, S5.4 and S5.5 used in Section S5 when proving the error bounds with high spatial dimensionality.

S6.1 Statement of Lemma S6.1

In Lemma S6.1, we restate the result of Propositions 6 and 7 of Li and Xiao (2021), which covers the general result of the consistency of the estimator for the lag-0 auto-covariance matrix of a stationary VAR(p) process.

Lemma S6.1. *Let $\mathbf{x}_t \in \mathbb{R}^N$ be a zero-meaned stationary VAR(p) process: $\mathbf{x}_t = \sum_{l=1}^{p} \mathbf{\Phi}_p \mathbf{x}_{t-p} + \boldsymbol{\xi}_t$, where $\boldsymbol{\xi}_t$ have independent sub-Gaussian entries. Let $\widehat{\mathbf{\Sigma}} = (1/T)\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t^\top$ and $\mathbf{\Sigma} = \mathrm{E}[\widehat{\mathbf{\Sigma}}]$, then we have:*

$$\mathrm{E}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_s \leq C\left(\sqrt{\frac{N\log N}{T}} + \frac{N\log N}{T}\right)\|\mathbf{\Sigma}\|_s, \tag{S6.42}$$

where C is an absolute constant.

We refer our readers to Appendix C.3 of Li and Xiao (2021) for the proof. As a corollary of Lemma S6.1, we have the following results:

Corollary S6.2. *Assume that $\{\mathbf{z}_t\}_{t=1}^{T}$ is generated by a stationary VAR($\widetilde{Q}$) process: $\mathbf{z}_t = \sum_{\widetilde{q}=1}^{\widetilde{Q}} \mathbf{C}_{\widetilde{q}} \mathbf{z}_{t-\widetilde{q}} + \boldsymbol{\nu}_t$, with $\boldsymbol{\nu}_t$ having independent sub-Gaussian entries, then with $\widehat{\mathbf{\Sigma}}_{\mathbf{z}} = (1/T)\sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t^\top$*

and $\mathbf{\Sigma}_{\mathbf{z}}^* = \mathrm{E}[\widehat{\mathbf{\Sigma}}_{\mathbf{z}}]$*, we have:*

$$\mathrm{P}\left(\left\|\widehat{\mathbf{\Sigma}}_{\mathbf{z}} - \mathbf{\Sigma}_{\mathbf{z}}^*\right\|_s \geq \epsilon\right) \leq C\epsilon^{-1}\left(\sqrt{\frac{D}{T}} + \frac{D}{T}\right), \tag{S6.43}$$

with C *being an absolute constant and* ϵ *being a fixed positive real number, and thus* $\left\|\widehat{\mathbf{\Sigma}}_{\mathbf{z}} - \mathbf{\Sigma}_{\mathbf{z}}^*\right\|_s \xrightarrow{p.} 0$.

Let $\{\mathbf{X}_t\}_{t=1}^T$ *be a zero-meaned matrix time series generated by the MARAC model with lag* P, Q *and* $\{\mathbf{z}_t\}_{t=1}^T$ *satisfies the assumption above and* $\{\mathbf{X}_t, \mathbf{z}_t\}_{t=1}^T$ *are jointly stationary in the sense of Theorem 1. Assume further that* $\mathbf{E}_t$ *has i.i.d. Gaussian entries with constant variance* σ^2*, then for* $\mathbf{y}_t = [\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top$*,* $\widehat{\mathbf{\Sigma}}_0 = (1/T)\sum_{t=1}^T \mathbf{y}_t\mathbf{y}_t^\top$ *and* $\mathbf{\Sigma}_0^* = \mathrm{E}[\mathbf{y}_t\mathbf{y}_t^\top]$*, we have:*

$$\mathrm{E}\left\|\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0^*\right\|_s \leq C\left(\sqrt{\frac{S\log S}{T}} + \frac{S\log S}{T}\right)\|\mathbf{\Sigma}_0^*\|_s, \tag{S6.44}$$

where C *is an absolute constant.*

Proof. The proof of (S6.43) is straightforward from Lemma S6.1 together with Markov inequality. The proof of (S6.44) also follows from Lemma S6.1 since $\{\mathbf{y}_t\}_{t=1}^T$ follows a stationary VAR($\max(P, Q, \widetilde{Q})$) process with i.i.d. sub-Gaussian noise (see (S3.8)) and $\mathrm{E}[(1/T)\sum_{t=1}^T \mathbf{y}_t\mathbf{y}_t^\top] = \mathrm{E}[\mathbf{y}_t\mathbf{y}_t^\top]$ due to stationarity. □

Note that the convergence of the variance estimator in spectral norm also indicates that each element of the variance estimator converges in probability. Also, the assumption that $\mathbf{E}_t$ has i.i.d. Gaussian entries can be relaxed to $\mathbf{E}_t$ having independent sub-Gaussian entries.

S6.2 Proof of Lemma S4.1

Proof. Without loss of generality, we fix P, Q as 1 and use the same notation as (S4.10) in Section S4.1, so the MARAC model can be written as $\mathbf{x}_t = \mathbf{y}_t\boldsymbol{\theta}^* + \mathbf{e}_t$. Correspondingly, the penalized log-likelihood $h(\boldsymbol{\theta}, \mathbf{\Omega})$ is specified by (S4.11) and given any $\bar{\mathbf{\Omega}}$, we have $\widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}) =$

$\arg\min_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \bar{\boldsymbol{\Omega}})$ as specified by (S4.13). Given the decomposition of $\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})$ in (S4.14), we have:

$$\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) - \boldsymbol{\theta}^* = -\lambda\widetilde{\mathbf{K}}\boldsymbol{\theta}^* + \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t}{T} + \lambda\widetilde{\mathbf{K}}\right)^{-1} \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{e}_t}{T}\right),$$

where $\|\lambda\widetilde{\mathbf{K}}\boldsymbol{\theta}^*\|_{\mathrm{F}} = o(T^{-1/2})$ since $\lambda = o(T^{-1/2})$ and the norm of the second term is $O_P(T^{-1/2})$. To show that the norm of the second term is $O_P(T^{-1/2})$, we first observe that:

$$\begin{aligned}
&\left\| \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t}{T} + \lambda\widetilde{\mathbf{K}}\right)^{-1} \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{e}_t}{T}\right) \right\|_{\mathrm{F}} \\
&\leq \left\| \underbrace{\left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t}{T} + \lambda\widetilde{\mathbf{K}}\right)^{-1}}_{\mathbf{L}_T^{-1}} \right\|_{\mathrm{F}} \cdot \left\| \underbrace{\left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{e}_t}{T}\right)}_{\mathbf{R}_T} \right\|_{\mathrm{F}}.
\end{aligned}$$

For the sequence of random matrices $\{\mathbf{L}_T\}_{T=1}^{\infty}$, we have:

$$\mathbf{L}_T = \frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t}{T} + \lambda\widetilde{\mathbf{K}} \xrightarrow{p} \begin{bmatrix} \mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_t) \otimes \bar{\boldsymbol{\Omega}} & \mathrm{Cov}(\mathbf{x}_t, \mathbf{z}_t) \otimes \bar{\boldsymbol{\Omega}}\mathbf{K} \\ \mathrm{Cov}(\mathbf{z}_t, \mathbf{x}_t) \otimes \mathbf{K}\bar{\boldsymbol{\Omega}} & \mathrm{Cov}(\mathbf{z}_t, \mathbf{z}_t) \otimes \mathbf{K}\bar{\boldsymbol{\Omega}}\mathbf{K} \end{bmatrix},$$

and we define the limiting matrix as $\mathbf{L}$. To show this, first note that the covariance estimator $\widehat{\mathrm{Var}}([\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top) = T^{-1}\sum_t [\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top [\mathbf{x}_t^\top, \mathbf{z}_t^\top]$ converges in probability to the true covariance $\mathrm{Var}([\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top)$, which we prove separately in Corollary S6.2. Secondly, notice that $\lambda = o(T^{-1/2})$, thus we have $\lambda\widetilde{\mathbf{K}} \to \mathbf{O}$ and thus we have the convergence in probability of $\mathbf{L}_T$ to $\mathbf{L}$ holds.

Notice that the limiting matrix $\mathbf{L}$ is invertible because the matrix $\mathbf{L}'$, defined as:

$$\mathbf{L}' = \begin{bmatrix} \mathbf{I}\otimes\mathbf{K} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} \end{bmatrix} \mathbf{L} \begin{bmatrix} \mathbf{I}\otimes\mathbf{K} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} \end{bmatrix} = \mathrm{Var}([\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top) \otimes (\mathbf{K}\bar{\boldsymbol{\Omega}}\mathbf{K}),$$

is invertible. To see why, firstly note that $\mathrm{Var}([\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top)$ is invertible because we can express $[\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top$ as $\sum_{j=0}^{\infty} \boldsymbol{\Phi}_j [\mathbf{e}_t^\top, \boldsymbol{\nu}_t^\top]^\top$, where $\{\boldsymbol{\Phi}_j\}_{j=0}^{\infty}$ is a sequence of matrices whose elements are absolutely summable and $\boldsymbol{\Phi}_0 = \mathbf{I}$, therefore, we have $\underline{\rho}(\mathrm{Var}([\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top)) \geq$

$\underline{\rho}(\text{Var}([\mathbf{e}_t^\top, \boldsymbol{\nu}_t^\top]^\top)) > 0$. Secondly, by Assumption 2, we have $\underline{\rho}(\mathbf{K}) > 0$ and we also have $\underline{\rho}(\bar{\boldsymbol{\Omega}}) > 0$ by definition, therefore we have $\mathbf{K}\bar{\boldsymbol{\Omega}}\mathbf{K}$ to be positive definite. The invertibility of $\mathbf{L}$ and the fact that $\mathbf{L}_T \xrightarrow{p} \mathbf{L}$ indicates that $\mathbf{L}_T^{-1} \xrightarrow{p} \mathbf{L}^{-1}$, since matrix inversion is a continuous function of the input matrix and the convergence in probability carries over under continuous transformations. Eventually, this leads to the conclusion that $\|\mathbf{L}_T^{-1}\|_{\text{F}} = O_P(1)$.

For the sequence of random matrices $\{\mathbf{R}_T\}_{T=1}^{\infty}$, we note that the sequence $\{\mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{e}_t\}_{t=1}^{\infty}$ is a martingale difference sequence (MDS) such that $\|\mathbf{R}_T\|_{\text{F}} = O_P(T^{-1/2})$ (see proposition 7.9 of Hamilton (2020) for the central limit theorem of martingale difference sequence). Combining the result of $\|\mathbf{L}_T\|_{\text{F}}$ and $\|\mathbf{R}_T\|_{\text{F}}$, we conclude that $\|\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) - \boldsymbol{\theta}^*\|_{\text{F}} = O_P(T^{-1/2})$.

Fix $\boldsymbol{\Omega} = \bar{\boldsymbol{\Omega}}$, we can decompose $h(\boldsymbol{\theta}, \bar{\boldsymbol{\Omega}})$ via the second-order Taylor expansion as follows:

$$\begin{aligned} h(\boldsymbol{\theta}, \bar{\boldsymbol{\Omega}}) &= h(\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}), \bar{\boldsymbol{\Omega}}) + \frac{1}{2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}))^\top \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t}{T} + \lambda \widetilde{\mathbf{K}} \right) (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})) \\ &\geq h(\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}), \bar{\boldsymbol{\Omega}}) + \frac{1}{2}\underline{\rho}(\mathbf{L}_T)\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\|_{\text{F}}^2, \end{aligned} \tag{S6.45}$$

and recall that $\mathbf{L}_T = T^{-1}\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t + \lambda \widetilde{\mathbf{K}}$. In the previous proof, we've shown that $\mathbf{L}_T \xrightarrow{p} \mathbf{L}$, with $\mathbf{L}$ being a positive definite matrix. Therefore, with probability approaching 1, we have $\underline{\rho}(\mathbf{L}_T) \geq \underline{\rho}(\mathbf{L})/2 > 0$.

With the lower bound on $\underline{\rho}(\mathbf{L}_T)$, we can claim that for some constant $C_1 > 0$:

$$\begin{aligned} &\inf_{\bar{\boldsymbol{\Omega}} \in \mathbb{F}_{\boldsymbol{\Omega}}: \|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\text{F}} \leq C_1} h(\boldsymbol{\theta}, \bar{\boldsymbol{\Omega}}) \\ &\geq \inf_{\bar{\boldsymbol{\Omega}} \in \mathbb{F}_{\boldsymbol{\Omega}}: \|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\text{F}} \leq C_1} \left\{ h(\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}), \bar{\boldsymbol{\Omega}}) + \frac{1}{4}\underline{\rho}(\mathbf{L}) \cdot \|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\|_{\text{F}}^2 \right\}, \end{aligned} \tag{S6.46}$$

with probability approaching 1. Now consider $\boldsymbol{\theta}$ belongs to the set $\{\boldsymbol{\theta} \in \mathbb{F}_{\boldsymbol{\theta}} | \sqrt{T}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\text{F}} \geq c_T\}$, where $c_T \to \infty$ is an arbitrary sequence that diverges to infinity. Within this set, we have:

$$\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\|_{\text{F}} \geq \frac{c_T}{\sqrt{T}} - \|\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\|_{\text{F}}, \tag{S6.47}$$

thus $\|\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\|_{\mathrm{F}} \gtrsim O_P(c_T'/\sqrt{T})$ for some sequence $c_T' \to \infty$ since $\|\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})-\boldsymbol{\theta}^*\|_{\mathrm{F}} = O_P(T^{-1/2})$. By the Taylor expansion in (S6.45), we can conclude that $h(\boldsymbol{\theta}^*, \bar{\boldsymbol{\Omega}}) = h(\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}), \bar{\boldsymbol{\Omega}}) + O_P(T^{-1})$, also using that $\|\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) - \boldsymbol{\theta}^*\|_{\mathrm{F}} = O_P(T^{-1/2})$. Combining this result together with the order of $\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\|_{\mathrm{F}}$, we have the following hold according to (S6.46):

$$\mathrm{P}\left(\inf_{\sqrt{T}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_{\mathrm{F}}\geq c_T}\inf_{\bar{\boldsymbol{\Omega}}\in\mathbb{F}_{\boldsymbol{\Omega}}:\|\bar{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^*\|_{\mathrm{F}}\leq C_1} h(\boldsymbol{\theta}, \bar{\boldsymbol{\Omega}}) > \inf_{\bar{\boldsymbol{\Omega}}\in\mathbb{F}_{\boldsymbol{\Omega}}:\|\bar{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^*\|_{\mathrm{F}}\leq C_1} h(\boldsymbol{\theta}^*, \bar{\boldsymbol{\Omega}})\right) \to 1. \tag{S6.48}$$

The result in (S6.48) indicates that for any $\boldsymbol{\theta}$ that lies outside of the set $\{\boldsymbol{\theta} \in \mathbb{F}_{\boldsymbol{\theta}} | \sqrt{T}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathrm{F}} < c_T\}$, the penalized log-likelihood is no smaller than a sub-optimal solution with probability approaching 1. Therefore, with probability approaching 1, one must have $\sqrt{T}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_{\mathrm{F}} \leq c_T$. And since the choice of c_T is arbitrary, we can conclude that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\mathrm{F}} = O_P(T^{-1/2})$ and thus each block of $\widehat{\boldsymbol{\theta}}$, namely $\widehat{\mathbf{A}}_p, \widehat{\mathbf{B}}_p, \widehat{\boldsymbol{\gamma}}_q$ converges to their ground truth value at the rate of $T^{-1/2}$.

The convergence rate of $\widehat{\mathbf{B}}_p \otimes \widehat{\mathbf{A}_p}$ can be derived from the following inequality:

$$\|\widehat{\mathbf{B}}_p \otimes \widehat{\mathbf{A}}_p - \mathbf{B}_p^* \otimes \mathbf{A}_p^*\|_{\mathrm{F}} \leq \|\widehat{\mathbf{B}}_p\|_{\mathrm{F}} \cdot \|\widehat{\mathbf{A}}_p - \mathbf{A}_p^*\|_{\mathrm{F}} + \|\widehat{\mathbf{B}}_p - \mathbf{B}_p^*\|_{\mathrm{F}} \cdot \|\mathbf{A}_p^*\|_{\mathrm{F}},$$

as well as the convergence rate of $\widehat{\mathbf{A}}_p$ and $\widehat{\mathbf{B}}_p$.

□

S6.3 Proof of Lemma S5.2

Proof. Based on the definition of $\mathbf{W}$ in equation (S5.33), we have

$$\begin{aligned}
\frac{\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}}{T} &= \widehat{\boldsymbol{\Sigma}}_{\mathbf{x},\mathbf{x}} \otimes \mathbf{I}_S - \left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \otimes \mathbf{K}\right)\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{z}} \otimes \mathbf{K} + \lambda \mathbf{I}_{SD}\right)^{-1}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}} \otimes \mathbf{I}_S\right) \\
&= \left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},\mathbf{x}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1} \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}}\right) \otimes \mathbf{I}_S \\
&\quad + \left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}} \otimes \mathbf{I}_S\right)^\top \left[\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{z}}^2 \otimes \lambda^{-1}\mathbf{K} + \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{z}} \otimes \mathbf{I}_S\right]^{-1}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}} \otimes \mathbf{I}_S\right),
\end{aligned} \tag{S6.49}$$

where the second term in (S6.49) is positive semi-definite since both $\underline{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}})$ and $\underline{\rho}(\mathbf{K})$ are non-negative and the whole term is symmetric. Therefore, by Weyl's inequality, one can lower bound $\underline{\rho}(\widetilde{\mathbf{X}}^\top \mathbf{W}\widetilde{\mathbf{X}}/T)$ by $\underline{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{x},\mathbf{x}} - \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}})$. For simplicity, we will use $\mathbf{A}, \mathbf{B}, \mathbf{C}$ to denote $\mathbf{\Sigma}^*_{\mathbf{x},\mathbf{x}}, \mathbf{\Sigma}^*_{\mathbf{z},\mathbf{x}}, (\mathbf{\Sigma}^*_{\mathbf{z},\mathbf{z}})^{-1}$, and $\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}$ to denote $\widehat{\mathbf{\Sigma}}_{\mathbf{x},\mathbf{x}}, \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}, \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1}$, respectively. We will use $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}^*$ to denote the estimated and true covariance matrix of $[\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top$. It is evident that $\|\mathbf{A}\|_s \leq \|\mathbf{\Sigma}^*\|_s$ and $\|\mathbf{B}\|_s \leq \|\mathbf{\Sigma}^*\|_s$, since both $\mathbf{A}$ and $\mathbf{B}$ are blocks of $\mathbf{\Sigma}^*$ and can thus be represented as $\mathbf{E}_1^\top \mathbf{\Sigma}^* \mathbf{E}_2$ with $\mathbf{E}_1, \mathbf{E}_2$ being two block matrices with unity spectral norm.

The rest of the proof focuses on showing that with $S \log S/T \to 0$, $\underline{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{x},\mathbf{x}} - \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}) \xrightarrow{p} \underline{\rho}(\mathbf{\Sigma}^*_{\mathbf{x},\mathbf{x}} - \left(\mathbf{\Sigma}^*_{\mathbf{z},\mathbf{x}}\right)^\top \left(\mathbf{\Sigma}^*_{\mathbf{z},\mathbf{z}}\right)^{-1} \mathbf{\Sigma}^*_{\mathbf{z},\mathbf{x}})$. For brevity, we omit the subscript s for the spectral norm notation and simply use $\|\cdot\|$ in this proof.

To start with, we have:

$$
\begin{aligned}
&\|\widehat{\mathbf{A}} - \widehat{\mathbf{B}}^\top \widehat{\mathbf{C}}\widehat{\mathbf{B}} - (\mathbf{A} - \mathbf{B}^\top \mathbf{C}\mathbf{B})\| \\
&\leq \|\widehat{\mathbf{A}} - \mathbf{A}\| + \|\widehat{\mathbf{B}}^\top \widehat{\mathbf{C}}\widehat{\mathbf{B}} - \mathbf{B}^\top \widehat{\mathbf{C}}\mathbf{B}\| + \|\mathbf{B}^\top \widehat{\mathbf{C}}\mathbf{B} - \mathbf{B}^\top \mathbf{C}\mathbf{B}\| \\
&\leq \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| + \|(\widehat{\mathbf{B}} - \mathbf{B})^\top \widehat{\mathbf{C}}\widehat{\mathbf{B}}\| + \|\mathbf{B}^\top \mathbf{C}(\widehat{\mathbf{B}} - \mathbf{B})\| + \|\mathbf{B}^\top (\widehat{\mathbf{C}} - \mathbf{C})\widehat{\mathbf{B}}\| \\
&\leq \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| + \|\widehat{\mathbf{B}} - \mathbf{B}\| \cdot \Big(\|\widehat{\mathbf{C}}\| \cdot \|\widehat{\mathbf{B}}\| + \|\mathbf{C}\| \cdot \|\mathbf{B}\|\Big) \\
&+ \|\mathbf{B}\| \cdot \|\widehat{\mathbf{B}}\| \cdot \|\widehat{\mathbf{C}} - \mathbf{C}\|. \qquad \text{(S6.50)}
\end{aligned}
$$

Based on Corollary S6.2, under the condition that $S \log S/T \to 0$ and the conditions that $\mathbf{z}_t$ follows a stationary VAR($\widetilde{Q}$) process and is jointly stationary with $\mathbf{x}_t$, we have $\|\widehat{\mathbf{C}} - \mathbf{C}\| \xrightarrow{p} 0$ and $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\| \xrightarrow{p} 0$. Therefore, with probability approaching 1, we have $\|\widehat{\mathbf{C}}\| \leq 2\|\mathbf{C}\|$, $\|\widehat{\mathbf{B}} - \mathbf{B}\| \leq \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\| \leq 2\|\mathbf{\Sigma}^*\|$ and $\|\widehat{\mathbf{B}}\| \leq 3\|\mathbf{\Sigma}^*\|$.

Combining these results and the upper bound in (S6.50), with probability approaching

1, we have:

$$\|\widehat{\mathbf{A}} - \widehat{\mathbf{B}}^\top \widehat{\mathbf{C}} \widehat{\mathbf{B}} - (\mathbf{A} - \mathbf{B}^\top \mathbf{C} \mathbf{B})\| \leq (1 + 7\|\mathbf{C}\| \cdot \|\mathbf{\Sigma}^*\|) \cdot \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\| + 3\|\mathbf{\Sigma}^*\|^2 \cdot \|\widehat{\mathbf{C}} - \mathbf{C}\|. \tag{S6.51}$$

The upper bound in (S6.51) can be arbitrarily small as $S, T \to \infty$ since $\|\widehat{\mathbf{C}} - \mathbf{C}\| \xrightarrow{p} 0$ and $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\| \xrightarrow{p} 0$.

Eventually, with probability approaching 1, we have:

$$\underline{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{x},\mathbf{x}} - \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1} \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}) \geq \frac{1}{2} \underline{\rho} \left(\mathbf{\Sigma}_{\mathbf{x},\mathbf{x}}^* - \left(\mathbf{\Sigma}_{\mathbf{z},\mathbf{x}}^*\right)^\top \left(\mathbf{\Sigma}_{\mathbf{z},\mathbf{z}}^*\right)^{-1} \mathbf{\Sigma}_{\mathbf{z},\mathbf{x}}^* \right) = \frac{c_{0,S}}{2}. \tag{S6.52}$$

This completes the proof.

□

S6.4 Proof of Lemma S5.3

Proof. By the definition of $\mathbf{W}$ in (S5.33), we have:

$$\begin{aligned} \operatorname{tr}(\mathbf{I} - \mathbf{W}) &= \operatorname{tr}\left[\left(\widehat{\mathbf{\Sigma}}_{\mathbf{z}} \otimes \mathbf{K} + \lambda \mathbf{I}_{SD} \right)^{-1} \left(\widehat{\mathbf{\Sigma}}_{\mathbf{z}} \otimes \mathbf{K} \right) \right] \\ &= \sum_{s=1}^{S} \sum_{d=1}^{D} \frac{\rho_d(\widehat{\mathbf{\Sigma}}_{\mathbf{z}}) \rho_s(\mathbf{K})}{\lambda + \rho_d(\widehat{\mathbf{\Sigma}}_{\mathbf{z}}) \rho_s(\mathbf{K})} \leq D \cdot \sum_{s=1}^{S} \frac{1}{1 + \lambda \bar{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{z}})^{-1} \rho_s(\mathbf{K})^{-1}}. \end{aligned} \tag{S6.53}$$

Using Lemma S6.1, we can bound $\bar{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{z}})$ by $2\bar{\rho}(\mathbf{\Sigma}_{\mathbf{z}}^*)$ with probability approaching 1 as $T \to \infty$. Conditioning on this high probability event and using the Assumption 3 that the kernel function is separable, the kernel Gram matrix $\mathbf{K}$ can be written as $\mathbf{K}_2 \otimes \mathbf{K}_1$ and thus (S6.53) can be bounded as:

$$D \cdot \sum_{s=1}^{S} \frac{1}{1 + \lambda \bar{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{z}})^{-1} \rho_s(\mathbf{K})^{-1}} \leq D \cdot \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{1}{1 + c_{\mathbf{z}} \lambda \rho_i(\mathbf{K}_1)^{-1} \rho_j(\mathbf{K}_2)^{-1}}, \tag{S6.54}$$

where $c_{\mathbf{z}} = 1/2\bar{\rho}(\mathbf{\Sigma}_{\mathbf{z}}^*)$. As $M, N \to \infty$, based on Assumption 4, we have $\rho_i(\mathbf{K}_1) \to Mi^{-r_0}$ and $\rho_j(\mathbf{K}_2) \to Nj^{-r_0}$. Consequently, we can find two constants $0 < c_1 < c_2$, with c_1 being

sufficiently small and c_2 being sufficiently large, such that:

$$\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{1}{1+c_2\lambda(ij)^{r_0}/S} \leq \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{1}{1+c_{\mathbf{z}}\lambda\rho_i(\mathbf{K}_1)^{-1}\rho_j(\mathbf{K}_2)^{-1}} \leq \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{1}{1+c_1\lambda(ij)^{r_0}/S}, \tag{S6.55}$$

where we, with a little abuse of notations, incorporate $c_{\mathbf{z}}$ into c_1, c_2. To estimate the order of the lower and upper bound in (S6.55), we first notice that for any constant $c>0$, one has:

$$\sum_{i=1}^{M\wedge N}\frac{1}{1+c\lambda i^{2r_0}/S} \leq \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{1}{1+c\lambda(ij)^{r_0}/S} \leq 2(M\vee N)\sum_{i=1}^{M\vee N}\frac{1}{1+c\lambda i^{2r_0}/S}. \tag{S6.56}$$

To approximate the sum in (S6.56), notice that:

$$\sum_{i=1}^{M\vee N}\frac{1}{1+c\lambda i^{2r_0}/S} = (S/c\lambda)^{1/2r_0}\cdot\sum_{i=1}^{M\vee N}\frac{1}{1+[\frac{i}{(S/c\lambda)^{1/2r_0}}]^{2r_0}}\cdot\frac{1}{(S/c\lambda)^{1/2r_0}},$$

and furthermore, we have:

$$\lim_{S\to\infty}\sum_{i=1}^{M\vee N}\frac{1}{1+[\frac{i}{(S/c\lambda)^{1/2r_0}}]^{2r_0}}\cdot\frac{1}{(S/c\lambda)^{1/2r_0}} = \int_0^C\frac{1}{1+x^{2r_0}}\mathrm{d}x<\infty,$$

where $C=\lim_{S\to\infty}c(M\vee N)^{2r_0}\cdot\gamma_S$. In the assumptions of Theorem 3, we assume that $M\vee N=O(\sqrt{S})$ and $\lim_{S\to\infty}\gamma_S\cdot S^{r_0}\to C_1$ where $0<C_1\leq\infty$. As a result, we have C being either a finite value or infinity, thus we have:

$$\lim_{S\to\infty}\sum_{i=1}^{M\vee N}\frac{1}{1+c\lambda i^{2r_0}/S} = \int_0^C\frac{1}{1+x^{2r_0}}\mathrm{d}x\cdot\lim_{S\to\infty}(S/c\lambda)^{1/2r_0} = O(\gamma_S^{-1/2r_0}). \tag{S6.57}$$

Combining (S6.53), (S6.54), (S6.55) and (S6.57), we have $\mathrm{tr}\,(\mathbf{I}-\mathbf{W})\lesssim O_P((M\vee N)\gamma_S^{-1/2r_0}) = O_P(\sqrt{S}\gamma_S^{-1/2r_0})$. To obtain the lower bound of $\mathrm{tr}\,(\mathbf{I}-\mathbf{W})$, we have:

$$\mathrm{tr}\,(\mathbf{I}-\mathbf{W})\geq D\cdot\sum_{s=1}^{S}\frac{1}{1+\lambda c'_{\mathbf{z}}\rho_s(\mathbf{K})^{-1}}\geq D\cdot\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{1}{1+c_3\lambda(ij)^{r_0}/S},$$

which holds with probability approaching 1 and $c'_{\mathbf{z}}=2/\underline{\rho}(\mathbf{\Sigma}^*_{\mathbf{z}})$ and the second inequality follows from (S6.55). To further lower bound the double summation, we have:

$$\cdot\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{1}{1+c_3\lambda(ij)^{r_0}/S}\geq\sum_{i=1}^{M\wedge N}\frac{1}{1+c_3\lambda(ij)^{r_0}/S}.$$

This new lower bound can be approximated with the same method as (S6.57) under the assumption that $M \wedge N = O(\sqrt{S})$. We can obtain the lower bound of $\mathrm{tr}\,(\mathbf{I} - \mathbf{W})$ as $O_P(\gamma_S^{-1/2r_0})$, which establishes the final result.

The upper bound of $\mathrm{tr}\,(\mathbf{W})$ is trivial since:

$$\mathrm{tr}\,(\mathbf{W}) = \sum_{s=1}^{S}\sum_{d=1}^{D} \frac{\lambda}{\lambda + \rho_d(\widehat{\mathbf{\Sigma}}_{\mathbf{z}})\rho_s(\mathbf{K})} \leq SD.$$

□

S6.5 Proof of Lemma S5.4

Proof. Let $\mathbf{W}' = (\mathbf{I}_T \otimes \mathbf{\Sigma}^{1/2})\mathbf{W}(\mathbf{I}_T \otimes \mathbf{\Sigma}^{1/2})$, then by the Hanson-Wright inequality (Rudelson and Vershynin, 2013), for any fixed $\mathbf{W}$, with $c, t > 0$ being constants and $K = \sqrt{8/3}$, we have:

$$\mathrm{P}\left[\left|\boldsymbol{\mathcal{E}}^\top \mathbf{W}\boldsymbol{\mathcal{E}} - \mathbf{E}\left[\boldsymbol{\mathcal{E}}^\top \mathbf{W}\boldsymbol{\mathcal{E}}\right]\right| > t \middle| \mathbf{W}\right] \leq 2\exp\left[-c \cdot \min\left(\frac{t^2}{K^4\|\mathbf{W}'\|_{\mathrm{F}}^2}, \frac{t}{K^2\|\mathbf{W}'\|_s}\right)\right]. \quad \text{(S6.58)}$$

We denote each of the $S \times S$ sub-matrix along the diagonal of $\mathbf{W}$ as $\mathbf{W}_1, \ldots, \mathbf{W}_T$, then for $\mathbf{E}\left[\boldsymbol{\mathcal{E}}^\top \mathbf{W}\boldsymbol{\mathcal{E}}|\mathbf{W}\right]$, we have:

$$\mathbf{E}\left[\boldsymbol{\mathcal{E}}^\top \mathbf{W}\boldsymbol{\mathcal{E}}|\mathbf{W}\right] \overset{(1)}{=} \sum_{t=1}^{T}\langle \mathbf{W}_t, \mathbf{\Sigma}\rangle \overset{(2)}{\leq} \sum_{t=1}^{T}\|\mathbf{\Sigma}\|_s \cdot \|\mathbf{W}_t\|_* \overset{(3)}{=} \|\mathbf{\Sigma}\|_s \cdot \mathrm{tr}\,(\mathbf{W}),$$

where $\langle \cdot, \cdot \rangle$ denotes the matrix inner product and $\|\cdot\|_*$ denotes the matrix nuclear norm. For (1), this is because of the definition of $\boldsymbol{\mathcal{E}}^\top \mathbf{W}'\boldsymbol{\mathcal{E}}$ as well as the independence between $\mathbf{W}$ and $\boldsymbol{\mathcal{E}}$. For (2), this inequality holds for the matrix/tensor inner product, and we refer our reader to Lemma 1 of Wang and Li (2020). Similarly, we also have $\mathbf{E}\left[\boldsymbol{\mathcal{E}}^\top \mathbf{W}\boldsymbol{\mathcal{E}}|\mathbf{W}\right] \geq \underline{\rho}(\mathbf{\Sigma}) \cdot \mathrm{tr}\,(\mathbf{W})$.

For (3), we can prove it via the semi-definiteness of $\mathbf{W}$:

$$\sum_{t=1}^{T}\|\mathbf{W}_t\|_* = \sum_{t=1}^{T}\left\|\mathbf{L}_t^\top \mathbf{W}\mathbf{L}_t\right\|_* = \mathrm{tr}\left(\mathbf{W} \cdot \left(\sum_{t=1}^{T}\mathbf{L}_t\mathbf{L}_t^\top\right)\right) = \mathrm{tr}\,(\mathbf{W}) = \|\mathbf{W}\|_*,$$

where $\mathbf{L}_t = [\overbrace{\mathbf{O},\ldots,\mathbf{O}}^{\text{t-1 blocks}},\mathbf{I},\overbrace{\mathbf{O},\ldots,\mathbf{O}}^{\text{T-t blocks}}]^\top$.

Letting $t = \mathbf{E}\left[\boldsymbol{\mathcal{E}}^\top\mathbf{W}\boldsymbol{\mathcal{E}}|\mathbf{W}\right]/2$, then we have:

$$\begin{aligned}
2\exp\left[-c\min\left(\frac{t^2}{K^4\|\mathbf{W}\|_{\mathrm{F}}^2},\frac{t}{K^2\|\mathbf{W}\|_s}\right)\right] &\leq 2\exp\left[-c\cdot\min\left(\frac{\underline{\rho}(\boldsymbol{\Sigma})^2\cdot\mathrm{tr}\left(\mathbf{W}\right)^{2\,2}}{K^4\|\mathbf{W}'\|_{\mathrm{F}}},\frac{\underline{\rho}(\boldsymbol{\Sigma})\cdot\mathrm{tr}\left(\mathbf{W}\right)}{K^2\cdot\|\mathbf{W}'\|_s}\right)\right] \\
&\leq 2\exp\left[-c\cdot\min\left(\frac{\underline{\rho}(\boldsymbol{\Sigma})^2\cdot\mathrm{tr}\left(\mathbf{W}\right)^{2\;\;2}}{K^4\bar{\rho}(\boldsymbol{\Sigma})^2\cdot\|\mathbf{W}\|_{\mathrm{F}}},\frac{\underline{\rho}(\boldsymbol{\Sigma})\cdot\mathrm{tr}\left(\mathbf{W}\right)}{K^2\cdot\bar{\rho}(\boldsymbol{\Sigma})}\right)\right] \\
&\leq 2\exp\left[-\frac{c}{C_1^2K^4}\cdot\mathrm{tr}\left(\mathbf{W}\right)\right] \qquad \text{(S6.59)}
\end{aligned}$$

We can lower bound the trace of $\mathbf{W}$ as follows. First, note that:

$$\mathrm{tr}\left(\mathbf{W}\right) = \sum_{s=1}^{S}\sum_{d=1}^{D}\frac{\lambda}{\lambda+\rho_d(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}})\rho_s(\mathbf{K})} \geq SD\cdot\frac{\lambda}{\lambda+\bar{\rho}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}})\bar{\rho}(\mathbf{K})}.$$

By the assumption that $\bar{\rho}(\mathbf{K})$ is bounded and that the fact that $\bar{\rho}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}) \leq 2\bar{\rho}(\boldsymbol{\Sigma}^*_{\mathbf{z}})$ with probability approaching 1 as $T\to\infty$, we have:

$$\mathrm{P}\left[\mathrm{tr}\left(\mathbf{W}\right)\geq\frac{SD\lambda}{\lambda+\bar{c}}\right]\to 1, \quad \text{as } T\to\infty, \qquad \text{(S6.60)}$$

where $\bar{c} = 2\bar{\rho}(\boldsymbol{\Sigma}^*_{\mathbf{z}})\bar{\rho}(\mathbf{K})$. Since $r_0 < 2$ and $\gamma_S\cdot S^{r_0}\to C_1$ as $S\to\infty$, with C_1 being either a positive constant or infinity, we have $\gamma_S\cdot S^2 = \lambda\cdot S\to\infty$. Therefore, we have $\mathrm{tr}\left(\mathbf{W}\right)\to\infty$ with probability approaching 1, as $S,T\to\infty$.

With these results, we can now upper bound the unconditional probability of the event $\{\left|\boldsymbol{\mathcal{E}}^\top\mathbf{W}\boldsymbol{\mathcal{E}}-\mathbf{E}\left[\boldsymbol{\mathcal{E}}^\top\mathbf{W}\boldsymbol{\mathcal{E}}\right]\right| > \mathbf{E}\left[\boldsymbol{\mathcal{E}}^\top\mathbf{W}\boldsymbol{\mathcal{E}}\right]/2\}$ as follows:

$$\begin{aligned}
&\mathrm{P}\left[\left|\boldsymbol{\mathcal{E}}^\top\mathbf{W}\boldsymbol{\mathcal{E}}-\mathbf{E}\left[\boldsymbol{\mathcal{E}}^\top\mathbf{W}\boldsymbol{\mathcal{E}}\right]\right| > \mathbf{E}\left[\boldsymbol{\mathcal{E}}^\top\mathbf{W}\boldsymbol{\mathcal{E}}\right]/2\right] \\
&\leq \mathrm{E}\left[2\exp\left[-\frac{c}{C_1^2K^4}\cdot\mathrm{tr}\left(\mathbf{W}\right)\right]\right] \\
&\leq 2\left\{1\cdot\mathrm{P}\left(\mathrm{tr}\left(W\right)<\frac{SD\lambda}{\lambda+\bar{c}}\right)+\exp\left[-\frac{c}{C_1^2K^4}\cdot\frac{SD\lambda}{\lambda+\bar{c}}\right]\cdot\mathrm{P}\left(\mathrm{tr}\left(W\right)\geq\frac{SD\lambda}{\lambda+\bar{c}}\right)\right\}\to 0.
\end{aligned} \qquad \text{(S6.61)}$$

This indicates that $\boldsymbol{\mathcal{E}}^\top \mathbf{W} \boldsymbol{\mathcal{E}}$ concentrates around its mean $\mathbf{E}\left[\boldsymbol{\mathcal{E}}^\top \mathbf{W} \boldsymbol{\mathcal{E}}\right]$ with high probability, and thus $\boldsymbol{\mathcal{E}}^\top \mathbf{W} \boldsymbol{\mathcal{E}} / \mathrm{tr}\left(\mathbf{W}\right) = O_P(\|\boldsymbol{\Sigma}\|_s) = O_P(c_{1,S})$. To establish $\boldsymbol{\mathcal{E}}^\top (\mathbf{I} - \mathbf{W})^2 \boldsymbol{\mathcal{E}} / \mathrm{tr}\left((\mathbf{I} - \mathbf{W})^2\right) = O_P(\|\boldsymbol{\Sigma}\|_s) = O_P(c_{1,S})$, we first note the unboundedness of $\mathrm{tr}\left((\mathbf{I} - \mathbf{W})^2\right)$ by following the same idea as the proof for Lemma S5.3, where we have:

$$\mathrm{tr}\left((\mathbf{I} - \mathbf{W})^2\right) \geq (S/c\lambda)^{1/2r_0} \cdot \sum_{i=1}^{M \wedge N} \left\{ \frac{1}{1 + \left[\frac{i}{(S/c\lambda)^{1/2r_0}}\right]^{2r_0}} \right\}^2 (S/c\lambda)^{-1/2r_0},$$

with probability approaching 1 and c is some constant. The remainder of the proof follows exactly the same steps, and we omit the rest of the details here. □

S6.6 Proof of Lemma S5.5

Proof. For any two arbitrary symmetric matrices $\mathbf{M}, \mathbf{N}$ with identical sizes, we use $\mathbf{M} \gtrsim \mathbf{N}$ to indicate that $\mathbf{M} - \mathbf{N}$ is positive semi-definite, and we use $\mathbf{M}^{1/2}$ to denote the symmetric, positive semi-definite square root matrix of $\mathbf{M}$.

Since $\mathbf{A} - \mathbf{B}$ is positive semi-definite, multiplying it by $\mathbf{B}^{-1/2}$ on both left and right sides of $\mathbf{A} - \mathbf{B}$, we have $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} \gtrsim \mathbf{I}$. Therefore, we have $\mathbf{B}^{-1/2}\mathbf{A}^{1/2}\mathbf{A}^{1/2}\mathbf{B}^{-1/2} \gtrsim \mathbf{I}$. Notice that the matrix $\mathbf{A}^{1/2}\mathbf{B}^{-1/2}$ is invertible and thus has no zero eigenvalues. As a result, all eigenvalues of $\mathbf{B}^{-1/2}\mathbf{A}^{1/2}\mathbf{A}^{1/2}\mathbf{B}^{-1/2}$ are the same as the eigenvalues of $\mathbf{A}^{1/2}\mathbf{B}^{-1/2}\mathbf{B}^{-1/2}\mathbf{A}^{1/2}$ and thus $\mathbf{A}^{1/2}\mathbf{B}^{-1/2}\mathbf{B}^{-1/2}\mathbf{A}^{1/2} \gtrsim \mathbf{I}$. Multiplying both sides by $\mathbf{A}^{-1/2}$ on both the left and right sides yields $\mathbf{B}^{-1} \gtrsim \mathbf{A}^{-1}$, which completes the proof. □

S7 Additional Details on Simulation and Algorithm

S7.1 Simulation Setup

We generate the simulated dataset according to the MARAC(P, Q) model specified by (2.1) and (2.3). We simulate the autoregressive coefficients $\mathbf{A}_p, \mathbf{B}_p$ such that they satisfy the stationarity condition specified in Theorem 1 and have a banded structure. We use a similar setup for generating $\mathbf{\Sigma}_r, \mathbf{\Sigma}_c$ with their diagonals fixed at unity. In Figure 1, we plot the simulated $\mathbf{A}_1, \mathbf{B}_1, \mathbf{\Sigma}_r, \mathbf{\Sigma}_c$ when $(M, N) = (20, 20)$.

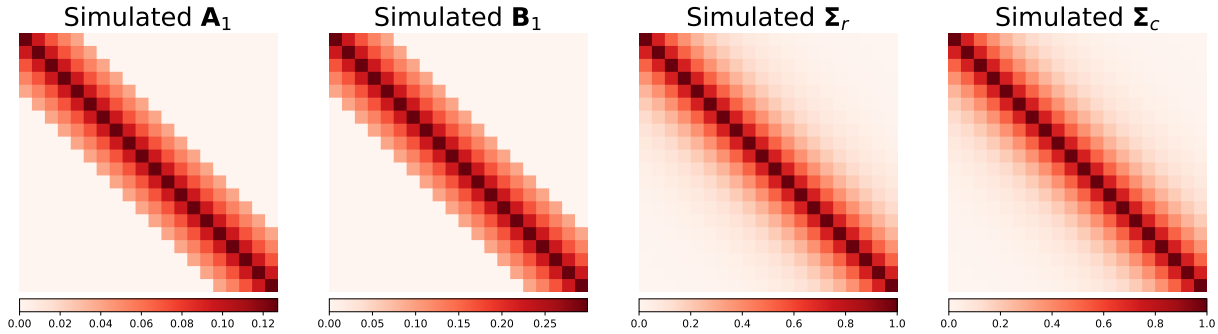

Figure 1: Visualization of the simulated $\mathbf{A}_1, \mathbf{B}_1, \mathbf{\Sigma}_r, \mathbf{\Sigma}_c$ with $M = N = 20$.

To generate $g_1, g_2, g_3 \in \mathbb{H}_k$ and mimic the spatial grid in our real data application in Section 6, we specify the 2-D spatial grid with the two dimensions being latitude and longitude of points on a unit sphere $\mathbb{S}^2$. Each of the evenly spaced $M \times N$ grid points has its polar-azimuthal coordinate pair as $(\theta_i, \phi_j) \in [0°, 180°] \times [0°, 360°], i \in [M], j \in [N]$, and one projects the sampled grid points on the sphere onto a plane to form an $M \times N$ matrix. The polar θ (co-latitude) and azimuthal ϕ (longitude) angles are very commonly used in the spherical coordinate system, with the corresponding Euclidean coordinates being $(x, y, z) = (\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta))$.

As for the spatial kernel, we choose the Lebedev kernel:

$$k_\eta(s_1, s_2) = \left(\frac{1}{4\pi} + \frac{\eta}{12\pi}\right) - \frac{\eta}{8\pi}\sqrt{\frac{1 - \langle s_1, s_2\rangle}{2}}, \quad s_1, s_2 \in \mathbb{S}^2, \tag{S7.62}$$

where $\langle \cdot, \cdot \rangle$ denotes the angle between two points on the sphere $\mathbb{S}^2$ and η is a hyperparameter of the kernel. In the simulation experiment as well as the real data application, we fix $\eta = 3$. The Lebedev kernel has the spherical harmonics functions as its eigenfunction:

$$k_\eta(s_1, s_2) = \frac{1}{4\pi} + \sum_{l=1}^{\infty} \frac{\eta}{(4l^2 - 1)(2l + 3)} \sum_{m=-l}^{l} Y_l^m(s_1) Y_l^m(s_2),$$

where $Y_l^m(\cdot)$ is a series of orthonormal real spherical harmonics bases defined on sphere $\mathbb{S}^2$:

$$Y_l^m(s) = Y_l^m(\theta, \phi) = \begin{cases} \sqrt{2} N_{lm} P_l^m(\cos(\theta)) \cos(m\phi) & \text{if } m > 0 \\ N_{l0} P_l^0(\cos(\theta)) & \text{if } m = 0 \\ \sqrt{2} N_{l|m|} P_l^{|m|}(\cos(\theta)) \sin(|m|\phi) & \text{if } m < 0 \end{cases},$$

with $N_{lm} = \sqrt{(2l + 1)(l - m)!/(4\pi(l + m)!)}$, and $P_l^m(\cdot)$ being the associated Legendre polynomials of order l. We refer our readers to Kennedy et al. (2013) for detailed information about the spherical harmonics functions and the associated isotropic kernels. Under our 2-D grid setup and the choice of kernel, we have found that empirically, the kernel Gram matrix $\mathbf{K}$ has its eigen spectrum decaying at a rate of $\rho_i(\mathbf{K}) \approx i^{-r}$ with $r \in [1.3, 1.5]$.

We randomly sample g_1, g_2, g_3 from Gaussian processes with a covariance kernel being the Lebedev kernel in (S7.62). Finally, we simulate the vector time series $\mathbf{z}_t$ using a VAR(1) process. In Figure 2, we visualize the simulated functional parameters as well as the vector time series from one random draw.

S7.2 Approximated Penalized MLE with Kernel Truncation

The iterative algorithm in Section 3.1 requires inverting an $MND \times MND$ matrix in (S1.4) when updating $\boldsymbol{\gamma}_q$, i.e., the coefficients of the representer functions $k(\cdot, s)$. One way to reduce

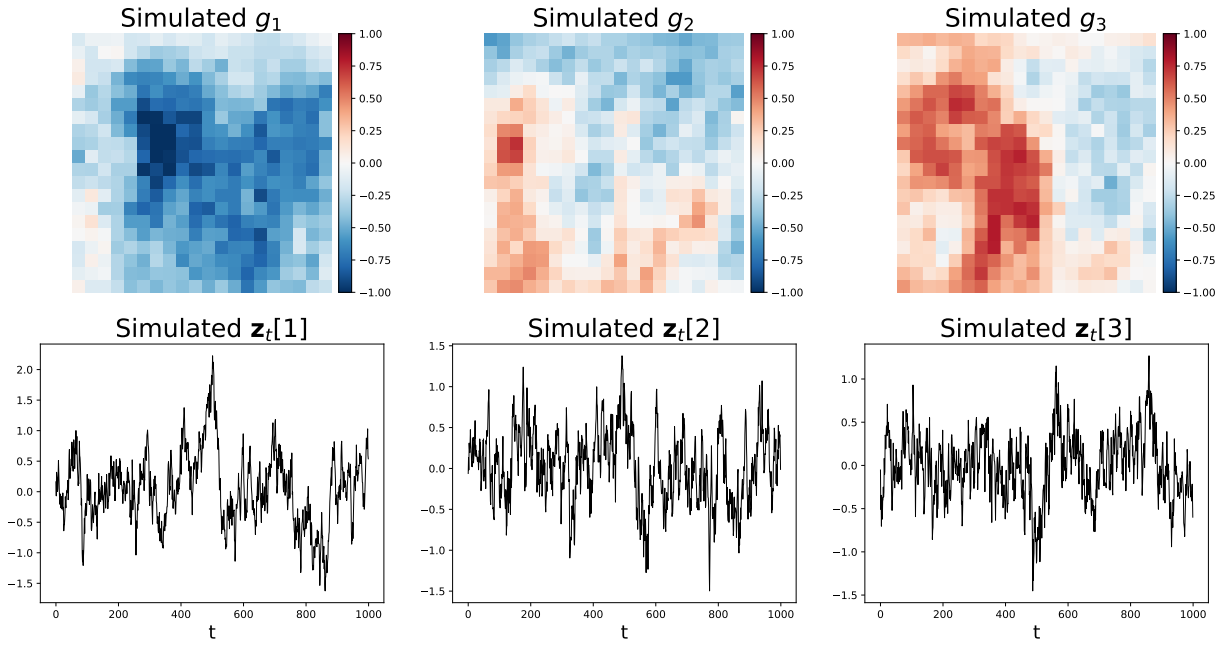

Figure 2: Simulated functional parameters g_1, g_2, g_3 evaluated on a 20×20 spatial grid (top row) and the corresponding auxiliary vector time series (bottom row).

the computational complexity without any approximation is to divide the step of updating $\boldsymbol{\gamma}_q = [\boldsymbol{\gamma}_{q,1}^\top : \cdots : \boldsymbol{\gamma}_{q,D}^\top]^\top$ to updating one block of parameters at a time following the order of $\boldsymbol{\gamma}_{q,1} \to \cdots \to \boldsymbol{\gamma}_{q,D}$. However, such a procedure requires inverting a matrix of size $MN \times MN$, which could still be high-dimensional.

To circumvent the issue of inverting large matrices, we can approximate the linear combination of all MN representers using a set of $R << MN$ basis functions, i.e., $\mathbf{K}\boldsymbol{\gamma}_{q,d} \approx \mathbf{K}_R\boldsymbol{\theta}_{q,d}$, where $\mathbf{K}_R \in \mathbb{R}^{MN \times R}, \boldsymbol{\theta}_{q,d} \in \mathbb{R}^R$. For example, one can reduce the spatial resolution by subsampling a fraction of the rows and columns of the matrix and only use the representers at the subsampled "knots" as the basis functions. In this subsection, we consider an alternative approach by truncating the Mercer decomposition in (3.4). A similar technique can be found in Kang et al. (2018).

Given the eigen-decomposition of $k(\cdot, \cdot)$ in (3.4), one can truncate the decomposition at

the R^{th} largest eigenvalue λ_R and get an approximation: $k(\cdot,\cdot) \approx \sum_{r\leq R} \lambda_r \psi_r(\cdot)\psi_r(\cdot)$. We will use the set of eigen-functions $\{\psi_1(\cdot),\ldots,\psi_R(\cdot)\}$ for faster computation. The choice of R depends on the decaying rate of the eigenvalue sequence $\{\lambda_r\}_{r=1}^{\infty}$ (thus the smoothness of the underlying functional parameters) and can be done via cross-validation in practice. Our simulation result shows that the estimation and prediction errors shrink monotonically as $R \to \infty$. Therefore, R can be chosen based on the computational resources available. The kernel truncation speeds up the computation at the cost of providing an overly-smoothed estimator, as we demonstrate later in this section.

Given the kernel truncation, any functional parameter $g_{q,d}(\cdot)$ is now approximated as: $g_{q,d}(\cdot) \approx \sum_{r\in[R]} [\boldsymbol{\theta}_{q,d}]_r \psi_r(\cdot)$. The parameter to be estimated now is $\boldsymbol{\Theta}_q = [\boldsymbol{\theta}_{q,1};\cdots;\boldsymbol{\theta}_{q,D}] \in \mathbb{R}^{R\times D}$, whose dimension is much lower than before ($\boldsymbol{\Gamma}_q \in \mathbb{R}^{MN\times D}$). Estimating $\boldsymbol{\Theta}_q$ requires solving a ridge regression problem, and the updating formula for $\mathbf{vec}\left(\boldsymbol{\Theta}_q\right) = \boldsymbol{\theta}_q$ can be written as:

$$\boldsymbol{\theta}_q^{(l+1)} \leftarrow \left[\boldsymbol{\Phi}\left(\mathbf{z}_{t-q}^{\top}\otimes \mathbf{K}_R, \boldsymbol{\Sigma}^{(l)}\right) + \lambda T\left(\mathbf{I}_D \otimes \boldsymbol{\Lambda}_R^{-1}\right)\right]^{-1} \boldsymbol{\Phi}\left(\mathbf{z}_{t-q}^{\top}\otimes \mathbf{K}_R, \widetilde{\mathbf{x}}_{t,-q}, \boldsymbol{\Sigma}^{(l)}\right),$$

where $\mathbf{K}_R \in \mathbb{R}^{MN\times R}$ satisfies $[\mathbf{K}_R]_{ur} = \psi_r(s_{ij}), u = i+(j-1)M$, and $\boldsymbol{\Lambda}_r = \mathrm{diag}(\lambda_1,\ldots,\lambda_R)$, with λ_r being the r^{th} largest eigenvalue of the Mercer decomposition of $k(\cdot,\cdot)$. Now we only need to invert a matrix of size $RD \times RD$, which speeds up the computation.

In Figure 3, we visualize the ground truth of g_3 and both its penalized MLE and truncated penalized MLE estimators. It is evident that the truncated penalized MLE estimators give a smooth approximation to g_3, and the approximation gets better when R gets larger. The choice of R should be as large as possible for accuracy, so one can determine R based on the computational resources available.

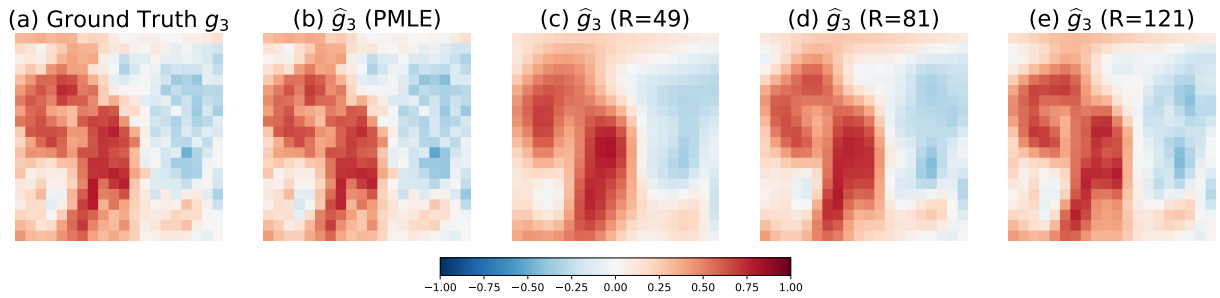

Figure 3: Ground truth g_3 (panel (a)) against the penalized MLE estimator $\widehat{g}_3$ (panel (b)) and the truncated penalized MLE estimator $\widehat{g}_3$ using $R \in \{49, 81, 121\}$ basis functions. $M = 20$.

Bibliography

Chen, R., H. Xiao, and D. Yang (2021). Autoregressive Models for Matrix-valued Time Series. *Journal of Econometrics 222*(1), 539–560.

Hall, P. and C. C. Heyde (2014). *Martingale Limit Theory and its Application*. Academic press.

Hamilton, J. D. (2020). *Time Series Analysis*. Princeton University Press.

Kang, J., B. J. Reich, and A.-M. Staicu (2018). Scalar-on-Image Regression via the Soft-Thresholded Gaussian Process. *Biometrika 105*(1), 165–184.

Kennedy, R. A., P. Sadeghi, Z. Khalid, and J. D. McEwen (2013). Classification and Construction of Closed-form Kernels for Signal Representation on the 2-sphere. In *Wavelets and Sparsity XV*, Volume 8858, pp. 169–183. SPIE.

Li, Z. and H. Xiao (2021). Multi-linear Tensor Autoregressive Models. *arXiv preprint arXiv:2110.00928*.

Rudelson, M. and R. Vershynin (2013). Hanson-Wright Inequality and Sub-Gaussian Concentration. *Electronic Communications in Probability 18*, 1–9.

Wang, M. and L. Li (2020). Learning from Binary Multiway Data: Probabilistic Tensor Decomposition and its Statistical Optimality. *The Journal of Machine Learning Research 21*(154), 1–38.