Matrix Autoregressive Model with Vector Time Series Covariates for Spatio-Temporal Data

Hu Sun
Department of Statistics, University of Michigan, Ann Arbor
Zuofeng Shang
Department of Mathematical Sciences, New Jersey Institute of Technology
and
Yang Chen
Department of Statistics, University of Michigan, Ann Arbor

May 17, 2024

Abstract

We develop a new methodology for forecasting matrix-valued time series with historical matrix data and auxiliary vector time series data. We focus on a time series of matrices defined on a static 2-D spatial grid and an auxiliary time series of non-spatial vectors. The proposed model, Matrix AutoRegression with Auxiliary Covariates (MARAC), contains an autoregressive component for the historical matrix predictors and an additive component that maps the auxiliary vector predictors to a matrix response via tensor-vector product. The autoregressive component adopts a bi-linear transformation framework following Chen et al. (2021), significantly reducing the number of parameters. The auxiliary component posits that the tensor coefficient, which maps non-spatial predictors to a spatial response, contains slices of spatially smooth matrix coefficients that are discrete evaluations of smooth functions from a Reproducible Kernel Hilbert Space (RKHS). We propose to estimate the model parameters under a penalized maximum likelihood estimation framework coupled with an alternating minimization algorithm. We establish the joint asymptotics of the autoregressive and tensor parameters under fixed and high-dimensional regimes. Extensive simulations and a geophysical application for forecasting the global Total Electron Content (TEC) are conducted to validate the performance of MARAC.

Keywords: Auxiliary covariates, matrix autoregression, reproducing kernel Hilbert space (RKHS), spatio-temporal forecast, tensor data model

1 Introduction

Matrix-valued time series data have received increasing attention in multiple scientific fields, such as economics (Wang et al., 2019), geophysics (Sun et al., 2022), and environmental science (Dong et al., 2020), where scientists are interested in modeling the joint dynamics of data observed on a 2-D grid across time. This paper focuses on the matrix-valued data defined on a 2-D spatial grid that contains the geographical information of the individual observations. As a concrete example, we visualize the global Total Electron Content (TEC) distribution in Figure 1. TEC is the density of electrons in the Earth's ionosphere along the vertical pathway connecting a radio transmitter and a ground-based receiver. An accurate prediction of the global TEC is critical since it can foretell the impact of space weather on the positioning, navigation, and timing (PNT) service (Wang et al., 2021; Younas et al., 2022). Every image in panel (A)-(C) is a 71×73 matrix, distributed on a spatial grid with 2.5°-latitude-by-5°-longitude resolution.

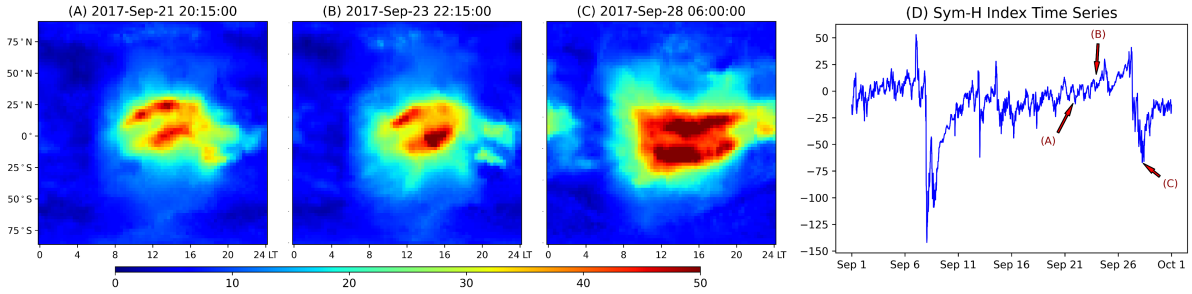

Figure 1: An example of matrix time series with auxiliary vector time series. Panels (A)-(C) show the global Total Electron Content (TEC) distribution at three timestamps on the latitude-local-time grid (source: the IGS TEC database (Hernández-Pajares et al., 2009)). Panel (D) plots the auxiliary Sym-H index time series, which measures the impact of solar eruptions on Earth. We highlight the time of panels (A)-(C) in (D) with arrows.

The matrix-valued time series, such as the TEC time series, is often associated with auxiliary vector time series that measures the same object, such as the Earth's ionosphere, from a different data modality. In panel (D) of Figure 1, we plot the global SYM-H index, which measures the geomagnetic activity caused by the solar eruptions that can finally impact the global TEC distribution. These non-spatial auxiliary covariates carry additional information related to the matrix time series data dynamics.

In this paper, we investigate the problem of forecasting future matrix data jointly with the historical matrices and the vector time series covariates. There are two major challenges in this modeling context. From the perspective of building a matrix-variate regression model, we need to integrate the information of predictors with non-uniform modes, namely both matrices and vectors. Adding the auxiliary vector covariates benefits the prediction and enables domain scientists to understand the interplay between different data modalities but at the same time, it complicates the modeling and the subsequent theoretical analysis. From the perspective of spatio-temporal analysis (Cressie and Wikle, 2015), we need to properly leverage the spatial information of the data and transform the classical spatial statistics framework to accommodate the grid geometry of matrix-valued data. In the remainder of this section, we briefly review the related literature that can shed light on these challenges and then summarize our unique contributions.

A naive yet straightforward prediction model is to vectorize the matrices as vectors and make predictions via the Vector Autoregression (VAR) (Stock and Watson, 2001). In this approach, the auxiliary vector covariates can be incorporated easily by concatenating them with the vectorized matrix predictors. However, vectorizing matrix data leads to the loss of spatial information and also requires a significant amount of parameters given the high dimensionality of the data. To avoid vectorizing the matrix data, scalar-on-tensor regression (Zhou et al., 2013; Guhaniyogi et al., 2017; Li et al., 2018; Papadogeorgou et al., 2021) tackles the problem by using matrix predictors directly. However, these models are built for *scalar* responses while in our setting we are dealing with *matrix* responses. Dividing the matrix response into individual scalar responses and fitting scalar-on-tensor regression still requires a significant number of parameters and more importantly, it fails to take the spatial information of the response into account.

The statistical framework that can incorporate matrices as both predictors and response is the tensor-on-tensor regression (Lock, 2018; Liu et al., 2020; Luo and Zhang, 2022) and more specifically for time series data, the matrix/tensor autoregression (Chen et al., 2021; Li and Xiao, 2021; Hsu et al., 2021; Wang et al., 2024). The matrix/tensor predictors are mapped to matrix/tensor responses via multi-linear transformations that greatly reduce the number of parameters. Our work builds on this framework and incorporates the non-spatial vector predictors at the same time.

To incorporate the vector predictor in the same model, we need to map vector predictors to matrix responses. Tensor-on-scalar regression (Rabusseau and Kadri, 2016; Sun and Li, 2017; Li and Zhang, 2017; Guha and Guhaniyogi, 2021) illustrates a way of mapping low-order scalar/vector predictors to high-order matrix/tensor responses via taking the tensor-vector product of the predictor with a high-order tensor coefficient. In this paper, we take a similar approach and introduce a 3-D tensor coefficient for the vector predictors such that our model can take predictors with non-uniform modes, which is a key distinction of our model compared to existing works.

The other distinction of our model is that our model leverages the spatial information of the matrix response. In our model, a key assumption is that the vector predictor has similar predictive effects on neighboring locations in the matrix response. This is equivalent to saying that the tensor coefficient is spatially smooth. Typically, the estimation of spatially smooth tensor coefficients in such regression models is done via adding a total-variation (TV) penalty (Wang et al., 2017; Shen et al., 2022; Sun et al., 2023) to the unknown tensor. The TV penalty leads to piecewise smooth estimators with sharp edges and enables feature selections. However, the estimation with the TV penalty requires solving non-convex optimization problems, making the subsequent theoretical analysis difficult. In our model, we utilize a simpler approach by assuming that the tensor coefficients are discrete evaluations of functional parameters from a Reproducing Kernel Hilbert Space (RKHS). Such a kernel method has been widely used in scalar-on-image regressions (Kang et al., 2018) where the regression coefficients of the image predictor are constrained to be spatially smooth.

We facilitate the estimation of the unknown functional parameters with the functional norm penalty. Functional norm penalties have been widely used for estimating smooth functions in classic semi/non-parametric learning in which data variables are either scalar/vector-valued (see Hastie et al., 2009; Gu, 2013; Yuan and Cai, 2010; Cai and Yuan, 2012; Shang and Cheng, 2013, 2015; Cheng and Shang, 2015; Yang et al., 2020). To the best of our knowledge, the present article is the first to consider functional norm penalty for tensor coefficient estimation in a matrix autoregressive setting.

To encapsulate, our paper has two major contributions. Firstly, we build a unified matrix autoregression framework for spatio-temporal data that incorporates lower-order

scalar/vector time series covariates. Such a framework has strong application motivation where domain scientists are curious about integrating the information of spatial and non-spatial data for predictions and inference. The framework also bridges regression methodologies with tensor predictors and responses of non-uniform modes, making the theoretical investigation itself an interesting topic. Secondly, we propose to estimate coefficients of the auxiliary covariates, together with the autoregressive coefficients, in a single penalized maximum likelihood estimation (MLE) framework with the RKHS functional norm penalty. The RKHS framework builds spatial continuity into the regression coefficients. We establish the joint asymptotics of the autoregressive coefficients and the functional parameters under fixed/high matrix dimensionality regimes and propose an efficient alternating minimization algorithm for estimation and validate it with extensive simulations and real applications.

The remainder of the paper is organized as follows. We introduce our model formally in Section 2 and provide model interpretations and comparisons in sufficient detail. Section 3 introduces the penalized MLE framework and describes the exact and approximate estimation algorithms. Large sample properties of the estimators under fixed and high matrix dimensionality are established in Section 4. Section 5 provides extensive simulation studies for validating the consistency of the estimators, demonstrating BIC-based model selection results, and comparing our method with various competitors. We apply our method to the global TEC data in Section 6 and make conclusions in Section 7. Technical proofs and additional details of the simulation and real data applications are deferred to supplemental materials.

2 Model

2.1 Notation

We adopt the following notations throughout the article. We use calligraphic bold-face letters (e.g. $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{G}}$) for tensors with at least three modes, uppercase bold-face letters (e.g. $\mathbf{X}, \mathbf{G}$) for matrix and lowercase bold-face letters (e.g. $\mathbf{x}, \mathbf{z}$) for vector and blackboard bold-faced letters for sets (e.g. $\mathbb{R}, \mathbb{S}$). To subscript any tensor/matrix/vector, we use square

brackets with subscripts such as $[\boldsymbol{\mathcal{G}}]_{ijd}, [\mathbf{z}_t]_d, [\mathbf{X}_t]_{ij}$, and we keep the subscript t inside the square bracket to index time. Any fibers and slices of tensor are subscript-ed with colons such as $[\boldsymbol{\mathcal{G}}]_{ij:}$, $[\boldsymbol{\mathcal{G}}]_{::d}$ and thus any row and column of a matrix is denoted as $[\mathbf{X}_t]_{i:}$ and $[\mathbf{X}_t]_{:j}$. If the slices of tensor/matrix are based on the last mode such as $[\boldsymbol{\mathcal{G}}]_{::d}$ and $[\mathbf{X}_t]_{:j}$, we will often omit the colons and write as $[\boldsymbol{\mathcal{G}}]_d$ and $[\mathbf{X}_t]_j$ for brevity. For any tensor $\boldsymbol{\mathcal{X}}$, we use $\mathbf{vec}\,(\boldsymbol{\mathcal{X}})$ to denote the vectorized tensor. For any two tensors $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}$ with identical size, we define their inner product as: $\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle = \mathbf{vec}\,(\boldsymbol{\mathcal{X}})^\top \mathbf{vec}\,(\boldsymbol{\mathcal{Y}})$, and we use $\|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}}$ to denote the Frobenius norm of a tensor and one has $\|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}} = \sqrt{\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \rangle}$.

Following Li and Zhang (2017), the *tensor-vector product* between a tensor $\boldsymbol{\mathcal{G}}$ of size $d_1 \times \cdots \times d_{K+1}$ and a vector $\mathbf{z} \in \mathbb{R}^{d_{K+1}}$, denoted as $\boldsymbol{\mathcal{G}} \bar{\times}_{(K+1)} \mathbf{z}$, or simply $\boldsymbol{\mathcal{G}} \bar{\times} \mathbf{z}$, is a tensor of size $d_1 \times \cdots \times d_K$ with $[\boldsymbol{\mathcal{G}} \bar{\times} \mathbf{z}]_{i_1 \ldots i_K} = \sum_{i_{K+1}} [\boldsymbol{\mathcal{G}}]_{i_1 \ldots i_K i_{K+1}} \cdot [\mathbf{z}]_{i_{K+1}}$. For tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, we use $\mathbf{X}_{(k)} \in \mathbb{R}^{d_k \times \prod_{m \neq k} d_m}$ to denote its k-mode matricization. The Kronecker product between matrices is denoted via $\mathbf{A} \otimes \mathbf{B}$ and the trace of a square matrix $\mathbf{A}$ is denoted as $\mathrm{tr}\,(\mathbf{A})$. We use $\bar{\rho}(\cdot), \underline{\rho}(\cdot), \rho_i(\cdot)$ to denote the maximum, minimum and i^{th} largest eigenvalue of a matrix. We use $diag(\mathbf{C}_1, \ldots, \mathbf{C}_d)$ to denote a block diagonal matrix with $\mathbf{C}_1, \ldots, \mathbf{C}_d$ along the diagonal. More details on the related tensor algebra can be found in Kolda and Bader (2009).

For the matrix time series $\mathbf{X}_t \in \mathbb{R}^{M \times N}$ in our modeling context, we assume that all $S = MN$ grid locations are points on an $M \times N$ grid within the domain $\bar{\mathbb{S}} = [0, 1]^2$. The collection of all the spatial locations is denoted as $\mathbb{S}$ and any particular element of $\mathbb{S}$ corresponding to the $(i, j)^{\text{th}}$ entry of the matrix is denoted as s_{ij}. We will often index the $(i, j)^{\text{th}}$ entry of the matrix $\mathbf{X}_t$ with a single index $u = i + (j - 1)M$ and thus s_{ij} will be denoted as s_u. We use $[N]$ to denote index set, i.e. $[N] = \{1, 2, \ldots, N\}$. Finally, we use $k(\cdot, \cdot) : \bar{\mathbb{S}} \times \bar{\mathbb{S}} \mapsto \mathbb{R}$ to denote a spatial kernel function and $\mathbb{H}_k$ to denote the corresponding Reproducing Kernel Hilbert Space (RKHS).

2.2 Matrix AutoRegression with Auxiliary Covariates (MARAC)

Let $\{\mathbf{X}_t, \mathbf{z}_t\}_{t=1}^T$ be a joint observation of the matrix and the auxiliary vector time series with $\mathbf{X}_t \in \mathbb{R}^{M \times N}, \mathbf{z}_t \in \mathbb{R}^D$. To forecast $\mathbf{X}_t$, we propose our Matrix AutoRegression with

Auxiliary Covariates, or MARAC, as:

$$\mathbf{X}_t = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{X}_{t-p} \mathbf{B}_p^\top + \sum_{q=1}^{Q} \boldsymbol{\mathcal{G}}_q \bar{\times} \mathbf{z}_{t-q} + \mathbf{E}_t, \tag{1}$$

where $\mathbf{A}_p \in \mathbb{R}^{M\times M}, \mathbf{B}_p \in \mathbb{R}^{N\times N}$ are the autoregressive coefficients for the lag-p matrix predictor and $\boldsymbol{\mathcal{G}}_q \in \mathbb{R}^{M\times N\times D}$ is the tensor coefficient for the lag-q vector predictor, and $\mathbf{E}_t$ is a noise term whose distribution will be specified later. The lag parameters P, Q are hyper-parameters of the model and we often refer to the model (1) as MARAC(P, Q).

Based on model (1), for the $(i, j)^{\text{th}}$ element of $\mathbf{X}_t$, the MARAC(P, Q) specifies the following model:

$$[\mathbf{X}_t]_{ij} = \sum_{p=1}^{P} \left\langle [\mathbf{A}_p]_{i:}^\top [\mathbf{B}_p]_{j:}, \mathbf{X}_{t-p} \right\rangle + \sum_{q=1}^{Q} [\boldsymbol{\mathcal{G}}_q]_{ij:}^\top \mathbf{z}_{t-q} + [\mathbf{E}_t]_{ij}, \tag{2}$$

where each autoregressive term is associated with a rank-1 coefficient matrix determined by the specific rows from $\mathbf{A}_p, \mathbf{B}_p$ and each non-spatial auxiliary covariate is associated with a coefficient vector that is location-specific, i.e. $[\boldsymbol{\mathcal{G}}_q]_{ij:}$. It now becomes more evident from (2) that the auxiliary vector covariates enter the model via an elementwise linear model. The autoregressive term utilizes $\mathbf{A}_p, \mathbf{B}_p$ to transform each lag-p predictor in a bi-linear form. Using such bi-linear transformation greatly reduces the total amount of parameters in that each lagged predictor that required M^2N^2 parameters previously now only requires M^2+N^2 parameters.

For the tensor coefficient $\boldsymbol{\mathcal{G}}_q$, we assume that it is spatially smooth. More specifically, we assume that $[\boldsymbol{\mathcal{G}}_q]_{ijd}$ and $[\boldsymbol{\mathcal{G}}_q]_{uvd}$ are similar if s_{ij}, s_{uv} are spatially close. Formally, we assume that each $[\boldsymbol{\mathcal{G}}_q]_d$, i.e. the coefficient matrix for the d^{th} covariate at lag-q, is a discrete evaluation of a function $g_{q,d}(\cdot) : [0, 1]^2 \mapsto \mathbb{R}$ on $\mathbb{S}$. Furthermore, each $g_{q,d}(\cdot)$ comes from an RKHS $\mathbb{H}_k$ endowed with the spatial kernel function $k(\cdot, \cdot)$. The spatial kernel function specifies the spatial smoothness of the functional parameters $g_{q,d}(\cdot)$ and thus the tensor coefficient $\boldsymbol{\mathcal{G}}_q$.

An alternative formulation for $\boldsymbol{\mathcal{G}}_q$ would be a low-rank form (Li and Zhang, 2017). We choose locally smooth over low rank to explicitly model the spatial smoothness of the coefficients and avoid tuning the tensor rank of $\boldsymbol{\mathcal{G}}_q$. We leave the low-rank model for future research and focus on the RKHS framework for the current paper.

Finally, for the additive noise term $\mathbf{E}_t$, we assume that it is i.i.d. from a multivariate normal distribution with a separable Kronecker-product covariance:

$$\mathbf{vec}\left(\mathbf{E}_t\right) \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r\right), \quad t \in [T] \tag{3}$$

where $\mathbf{\Sigma}_r \in \mathbb{R}^{M\times M}, \mathbf{\Sigma}_c \in \mathbb{R}^{N\times N}$ are the row/column covariance components. Such a Kronecker-product covariance is commonly seen in the covariance models for multi-way data (Hoff, 2011; Fosdick and Hoff, 2014) with the merit of reducing the number of parameters significantly.

Compared to existing models that can only deal with either matrix or vector predictors, our model (1) can incorporate predictors with non-uniform modes. If one redefines $\mathbf{E}_t$ in our model as $\sum_{q=1}^{Q} \boldsymbol{\mathcal{G}}_q \bar{\times} \mathbf{z}_{t-q} + \mathbf{E}_t$, i.e. all terms except the autoregressive term, then our model ends up specifying:

$$\mathrm{Cov}(\mathbf{vec}\left(\mathbf{E}_t\right), \mathbf{vec}\left(\mathbf{E}_{t'}\right)) = \mathbb{1}_{\{t=t'\}} \cdot \mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r + \mathbf{F}\mathbf{M}\mathbf{F}^\top$$
$$\mathbf{F} = [(\boldsymbol{\mathcal{G}}_1)_{(3)}^\top : \cdots : (\boldsymbol{\mathcal{G}}_Q)_{(3)}^\top], \quad \mathbf{M} = [\mathrm{Cov}(\mathbf{z}_{t-q_1}, \mathbf{z}_{t'-q_2})]_{q_1,q_2\in[Q]}$$

where $(\boldsymbol{\mathcal{G}}_q)_{(3)}$ is the mode-3 matricization of $\boldsymbol{\mathcal{G}}_q$ and we will use $\mathbf{G}_q$ to denote it for the rest of the paper. This new formulation reveals how our model differs from other autoregression models with matrix predictors. The covariance of $\mathbf{E}_t, \mathbf{E}_{t'}$ in our model contains a separable covariance matrix $\mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r$ that is based on the matrix grid geometry, a locally smooth coefficient matrix $\mathbf{F}$ that captures the local spatial dependency and an auto-covariance matrix $\mathbf{M}$ that captures the temporal dependency. Consequently, entries of $\mathbf{E}_t$ are more correlated if either they are spatially/temporally close or they share the same row/column index and are thus more flexible for spatial data distributed on a matrix grid.

As a comparison, in the kriging framework (Cressie, 1986), the covariance of $\mathbf{E}_t, \mathbf{E}_{t'}$ is characterized by a spatio-temporal kernel that captures the dependencies among spatial and temporal neighbors. Such kernel method can account for the local dependency but not the spatial dependency based on the matrix grid geometry. In the matrix autoregression model (Chen et al., 2021), the authors do not consider the local spatial dependencies among entries of $\mathbf{E}_t$ nor the temporal dependency across different t. In Hsu et al. (2021), the matrix autoregression model is generalized to adapt to spatial data via fixed-rank co-kriging (FRC) (Cressie and Johannesson, 2008) with $\mathrm{Cov}(\mathbf{vec}\left(\mathbf{E}_t\right), \mathbf{vec}\left(\mathbf{E}_{t'}\right)) = \mathbb{1}_{\{t=t'\}} \cdot \mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r +$

$\mathbf{FMF}^\top$, where $\mathbf{M}$ is a $k \times k$ coefficient matrix and $\mathbf{F}$ is a pre-specified $MN \times k$ spatial basis matrix. Such a co-kriging framework does not account for the temporal dependency of the noises nor does it consider the auxiliary covariates. Our model generalizes these previous works to allow for temporally dependent noise with both local and grid spatial dependency.

The combination of (1) and (3) specifies the complete MARAC(P, Q) model. Vectorizing both sides of (1) yields the vectorized MARAC(P, Q) model:

$$\mathbf{x}_t = \sum_{p=1}^{P} (\mathbf{B}_p \otimes \mathbf{A}_p)\, \mathbf{x}_{t-p} + \sum_{q=1}^{Q} \mathbf{G}_q^\top \mathbf{z}_{t-q} + \mathbf{e}_t, \quad \mathbf{e}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r) \tag{4}$$

where $\mathbf{x}_t = \mathbf{vec}(\mathbf{X}_t), \mathbf{e}_t = \mathbf{vec}(\mathbf{E}_t)$, and recall that $\mathbf{G}_q = (\boldsymbol{\mathcal{G}}_q)_{(3)}$. It is now more evident that the Kronecker-product structure of the autoregressive coefficient matrix and the noise covariance matrix greatly reduce the number of parameters, making the regression estimation feasible given finite samples. The spatially smooth structure of $\mathbf{G}_q$ leverages the spatial information of the spatial data. In the next section, we will discuss the estimating algorithm of the model parameters of MARAC.

3 Estimating Algorithm

In this section, we discuss the parameter estimation for the MARAC(P, Q) model (1). We first propose a penalized maximum likelihood estimator (MLE) in Section 3.1 for exact parameter estimation. Then, we propose an approximation to the penalized MLE in Section 3.2 for faster computation when dealing with high-dimensional matrix data. Finally, in Section 3.3, we outline the model selection criterion for selecting the lag hyper-parameters whose consistency will be validated empirically in Section 5.

3.1 Penalized Maximum Likelihood Estimation (MLE)

To estimate the parameters of the MARAC(P, Q) model, which we denote collectively as $\mathbf{\Theta}$, we propose a penalized maximum likelihood estimation (MLE) approach. Following the distribution assumption on $\mathbf{E}_t$ in (3), we can write the negative log-likelihood of $\{\mathbf{X}_t\}_{t=1}^T$ with a squared RKHS functional norm penalty, after dropping the constants, as:

$$\mathfrak{L}_\lambda(\mathbf{\Theta}) = -\frac{1}{T} \sum_{t \in [T]} \ell\left(\mathbf{X}_t | \mathbf{X}_{t-1:P}, \mathbf{z}_{t-1:Q}; \mathbf{\Theta}\right) + \frac{\lambda}{2} \sum_{q \in [Q]} \sum_{d \in [D]} \|g_{q,d}\|_{\mathbb{H}_k}^2, \tag{5}$$

where $\ell(\cdot)$ is the conditional log-likelihood of $\mathbf{X}_t$:

$$\ell\left(\mathbf{X}_t|\mathbf{X}_{t-1:P}, \mathbf{z}_{t-1:Q}; \boldsymbol{\Theta}\right) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r| - \frac{1}{2}\mathbf{r}_t^\top \left(\boldsymbol{\Sigma}_c^{-1} \otimes \boldsymbol{\Sigma}_r^{-1}\right)\mathbf{r}_t, \tag{6}$$

and $\mathbf{r}_t = \mathbf{x}_t - \sum_p (\mathbf{B}_p \otimes \mathbf{A}_p)\mathbf{x}_{t-p} - \sum_q \mathbf{G}_q^\top \mathbf{z}_{t-q}$ is the vectorized residual at t. To estimate the parameters, one needs to solve a constrained minimization problem:

$$\min_{\boldsymbol{\Theta}} \mathfrak{L}_\lambda(\boldsymbol{\Theta}), \quad \text{s.t. } g_{q,d}(s_{ij}) = [\boldsymbol{\mathcal{G}}_q]_{ijd}, \text{ for all } s_{ij} \in \mathbb{S}. \tag{7}$$

We now define the functional norm penalty in (5) explicitly and derive a *finite-dimensional equivalent* of the optimization problem above. We assume that the spatial kernel function $k(\cdot,\cdot)$ is continuous and square-integrable, thus it has an eigen-decomposition following the Mercer's Theorem (Williams and Rasmussen, 2006):

$$k(s_{ij}, s_{uv}) = \sum_{r=1}^{\infty} \lambda_r \psi_r(s_{ij})\psi_r(s_{uv}), \quad s_{ij}, s_{uv} \in [0,1]^2, \tag{8}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots$ is a sequence of non-negative eigenvalues and $\psi_1, \psi_2, \ldots$ is a set of orthonormal eigen-functions on $[0,1]^2$. The functional norm of function g from the RKHS $\mathbb{H}_k$ endowed with kernel $k(\cdot,\cdot)$ is defined as:

$$\|g\|_{\mathbb{H}_k} = \sqrt{\sum_{r=1}^{\infty} \frac{\beta_r^2}{\lambda_r}}, \qquad \text{where } g(\cdot) = \sum_{r=1}^{\infty} \beta_r \psi_r(\cdot), \tag{9}$$

following van Zanten and van der Vaart (2008).

Given any $\lambda > 0$ in (5), the generalized representer theorem (Schölkopf et al., 2001) suggests that the solution of the functional parameters, denoted as $\{\widetilde{g}_{q,d}\}_{q=1,d=1}^{Q,D}$, of the minimization problem (7), with all other parameters held fixed, is a linear combination of the representers $\{k(\cdot,s)\}_{s\in\mathbb{S}}$ plus a linear combination of the basis functions $\{\phi_1, \ldots, \phi_J\}$ of the null space of $\mathbb{H}_k$, i.e.,

$$\widetilde{g}_{q,d}(\cdot) = \sum_{s\in\mathbb{S}} \gamma_s k(\cdot, s) + \sum_{j=1}^{J} \alpha_j \phi_j(\cdot), \quad \|\phi_j\|_{\mathbb{H}_k} = 0, \tag{10}$$

where we omit the subscript (q,d) for the coefficient γ_s, α_j for brevity but they are essentially different for each (q,d). We assume that the null space of $\mathbb{H}_k$ contains only the zero function for the remainder of the paper. As a consequence of (10), the minimization problem in (7) can be reduced to a finite-dimensional Kernel Ridge Regression (KRR) problem. We summarize the discussion above in the proposition below:

Proposition 1 *If $\lambda > 0$, the constrained minimization problem in (7) is equivalent to the following unconstrained kernel ridge regression problem:*

$$\min_{\mathbf{\Theta}} \left\{ \frac{1}{2} \log |\mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r| + \frac{1}{2T} \sum_{t \in [T]} \mathbf{r}_t^\top \left(\mathbf{\Sigma}_c^{-1} \otimes \mathbf{\Sigma}_r^{-1} \right) \mathbf{r}_t + \frac{\lambda}{2} \sum_{q \in [Q]} tr \left(\mathbf{\Gamma}_q^\top \mathbf{K} \mathbf{\Gamma}_q \right) \right\}, \tag{11}$$

where $\mathbf{r}_t = \mathbf{x}_t - \sum_p (\mathbf{B}_p \otimes \mathbf{A}_p) \mathbf{x}_{t-p} - \sum_q \mathbf{K} \mathbf{\Gamma}_q \mathbf{z}_{t-q}$ is the vectorized residual, $\mathbf{K} \in \mathbb{R}^{MN \times MN}$ is the kernel Gram matrix with $[\mathbf{K}]_{u_1 u_2} = k(s_{i_1 j_1}, s_{i_2 j_2}), s_{i_l j_l} \in \mathbb{S}, u_l = i_l + (j_l - 1)M, l = 1, 2$ and $\mathbf{\Gamma}_q \in \mathbb{R}^{MN \times D}$ contains the coefficients of the representers with $[\mathbf{\Gamma}_q]_{ud}$ being the coefficient for the u^{th} representer $k(\cdot, s_u)$ and the d^{th} auxiliary covariate at lag q.

We give the proof in the supplemental material. After introducing the functional norm penalty, the original tensor coefficient is now converted to a linear combination of the representer functions with the relationship that $[\boldsymbol{\mathcal{G}}_q]_{ijd} = \langle [\mathbf{K}]_{u:}^\top, [\mathbf{\Gamma}_q]_{:d} \rangle$ where $u = i + (j - 1)M$.

We attempt to solve the minimization problem in (11) with an alternating minimization algorithm (Attouch et al., 2013) where we update one block of parameters at a time while keeping the others fixed. We update the parameters following the order of: $\mathbf{A}_1 \to \mathbf{B}_1 \to \cdots \to \mathbf{A}_P \to \mathbf{B}_P \to \mathbf{\Gamma}_1 \to \cdots \to \mathbf{\Gamma}_Q \to \mathbf{\Sigma}_r \to \mathbf{\Sigma}_c \to \mathbf{A}_1 \to \cdots$. We choose the alternating minimization algorithm for its simplicity and efficiency. Each step of the algorithm conducts exact minimization over one block of the parameters, leading to a non-increasing sequence of the objective function, which guarantees the convergence of the algorithm towards a local stationary point.

To solve the optimization problem in (11) for $\mathbf{A}_p$ at the $(l+1)^{\text{th}}$ iteration, it suffices to solve the following least-square problem:

$$\min_{\mathbf{A}_p} \left\{ \sum_{t \in [T]} \text{tr} \left(\widetilde{\mathbf{X}}_t(\mathbf{A}_p)^\top \left(\mathbf{\Sigma}_r^{(l)} \right)^{-1} \widetilde{\mathbf{X}}_t(\mathbf{A}_p) \left(\mathbf{\Sigma}_c^{(l)} \right)^{-1} \right) \right\}, \tag{12}$$

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 $[\boldsymbol{\mathcal{G}}_q^{(l)}]_{ijd} = \langle [\mathbf{K}]_{u:}^\top, [\boldsymbol{\Gamma}_q^{(l)}]_{:d} \rangle$, with $u = i + (j-1)M$. The superscript l represents the value at the l^{th} iteration. To simplify the notation, we define $\boldsymbol{\Phi}(\mathbf{A}_t, \mathbf{B}_t, \boldsymbol{\Sigma}) = \sum_t \mathbf{A}_t^\top \boldsymbol{\Sigma}^{-1} \mathbf{B}_t$, where $\boldsymbol{\Sigma}, \mathbf{A}_t, \mathbf{B}_t$ are arbitrary matrices/vectors with conformal matrix sizes and we simply write $\boldsymbol{\Phi}(\mathbf{A}_t, \boldsymbol{\Sigma})$ if $\mathbf{A}_t = \mathbf{B}_t$. Solving (12) yields the following updating formula for $\mathbf{A}_p^{(l+1)}$:

$$\mathbf{A}_p^{(l+1)} \leftarrow \boldsymbol{\Phi}\left(\widetilde{\mathbf{X}}_{t,-p}^\top, \mathbf{B}_p^{(l)} \mathbf{X}_{t-p}^\top, \boldsymbol{\Sigma}_c^{(l)}\right) \boldsymbol{\Phi}\left(\mathbf{B}_p^{(l)} \mathbf{X}_{t-p}^\top, \boldsymbol{\Sigma}_c^{(l)}\right)^{-1} \tag{13}$$

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

$$\mathbf{B}_p^{(l+1)} \leftarrow \boldsymbol{\Phi}\left(\widetilde{\mathbf{X}}_{t,-p}, \mathbf{A}_p^{(l+1)} \mathbf{X}_{t-p}, \boldsymbol{\Sigma}_r^{(l)}\right) \boldsymbol{\Phi}\left(\mathbf{A}_p^{(l+1)} \mathbf{X}_{t-p}, \boldsymbol{\Sigma}_r^{(l)}\right)^{-1} \tag{14}$$

For updating $\boldsymbol{\Gamma}_q$, or its vectorized version $\boldsymbol{\gamma}_q = \mathbf{vec}\,(\boldsymbol{\Gamma}_q)$, 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 $\boldsymbol{\gamma}_q^{(l+1)}$:

$$\boldsymbol{\gamma}_q^{(l+1)} \leftarrow \left[\left(\sum_{t \in [T]} \mathbf{z}_{t-q} \mathbf{z}_{t-q}^\top \right) \otimes \mathbf{K} + \lambda T \left(\mathbf{I}_D \otimes \boldsymbol{\Sigma}^{(l)} \right) \right]^{-1} \left[\sum_{t \in [T]} \left(\mathbf{z}_{t-q} \otimes \widetilde{\mathbf{x}}_{t,-q} \right) \right]. \tag{15}$$

The step in (15) can be slow since one needs to invert a square matrix of size $MND \times MND$. In Section 3.2, we propose an approximation to (15) to speed up the computation under high dimensionality.

The updating rule of $\boldsymbol{\Sigma}_r^{(l+1)}$ and $\boldsymbol{\Sigma}_c^{(l+1)}$ can be easily derived by taking their derivative in (11) and setting it to zero. Specifically, we have:

$$\boldsymbol{\Sigma}_r^{(l+1)} \leftarrow \frac{1}{NT} \boldsymbol{\Phi}\left(\widetilde{\mathbf{X}}_t^\top, \boldsymbol{\Sigma}_c^{(l)}\right) \tag{16}$$

$$\boldsymbol{\Sigma}_c^{(l+1)} \leftarrow \frac{1}{MT} \boldsymbol{\Phi}\left(\widetilde{\mathbf{X}}_t, \boldsymbol{\Sigma}_r^{(l+1)}\right). \tag{17}$$

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

The algorithm cycles through (13), (14), (15), (16) and (17) 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 make two additional remarks on the algorithm:

Remark 2 *(Identifiability Constraint) The MARAC(P, Q) model specified in (1) is scale-unidentifiable in that one can re-scale each pair of $(\mathbf{A}_p, \mathbf{B}_p)$ by a non-zero constant c and obtain $(c \cdot \mathbf{A}_p, c^{-1} \cdot \mathbf{B}_p)$ without changing their Kronecker product. To enforce scale identifiability, we re-scale the algorithm output for each pair of $(\mathbf{A}_p, \mathbf{B}_p)$ such that $\|\mathbf{A}_p\|_{\mathrm{F}} = 1$, $sign(tr(\mathbf{A}_p)) = 1$. The identifiability constraint is enforced before outputting the estimators.*

Remark 3 *(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 instead:*

$$\|\mathbf{B}_p^{(l+1)} - \mathbf{B}_p^{(l)}\|_{\mathrm{F}} \cdot \|\mathbf{A}_p^{(l+1)}\|_{\mathrm{F}} + \|\mathbf{B}_p^{(l)}\|_{\mathrm{F}} \cdot \|\mathbf{A}_p^{(l+1)} - \mathbf{A}_p^{(l)}\|_{\mathrm{F}}, \tag{18}$$

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

3.2 Approximated Penalized MLE with Kernel Truncation

The iterative algorithm in Section 3.1 requires inverting an $MND \times MND$ matrix in (15) when updating $\boldsymbol{\gamma}_q$, i.e., the coefficients of the representer functions $k(\cdot, s)$. One way to reduce 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 (8). A similar technique can be found in Kang et al. (2018).

Given the eigen-decomposition of $k(\cdot, \cdot)$ in (8), one can truncate the decomposition at the R^{th} largest eigenvalue λ_R and get an approximation: $k(\cdot, \cdot) \approx \sum_{r \le R} \lambda_r \psi_r(\cdot)\psi_r(\cdot)$. 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). 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 empirically in the supplemental material.

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 generating $\mathbf{vec}\left(\boldsymbol{\Theta}_q\right)$ at the $(l+1)^{\text{th}}$ iteration can be written as:

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

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 $\boldsymbol{\Lambda}_r = \text{diag}(\lambda_1, \ldots, \lambda_R)$, 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.

3.3 Lag Selection

The MARAC(P, Q) model (1) has three hyper-parameters: the autoregressive lag P, the auxiliary covariate lag Q, and the RKHS norm penalty weight λ. In practice, λ can be chosen by cross-validation while choosing P and Q requires a more formal model selection criterion. We propose to select P and Q by using information criterion, including the Akaike Information Criterion (AIC) (Akaike, 1998) and the Bayesian Information Criterion (BIC) (Schwarz, 1978). We formally define the AIC and BIC for the MARAC(P, Q) model here and empirically validate their consistency via simulation experiments in Section 5.

Let $\widehat{\boldsymbol{\Theta}}$ be the set of the estimated parameters of the MARAC(P, Q) model, and $\mathbf{df}_{P,Q,\lambda}$ be the *effective degrees of the freedom* of the model. We can then define the AIC and the BIC as follows:

$$\text{AIC}(P, Q, \lambda) = -2\sum_{t\in[T]} \ell(\mathbf{X}_t|\mathbf{X}_{t-1:P}, \mathbf{z}_{t-1:Q}, \widehat{\boldsymbol{\Theta}}) + 2 \cdot \mathbf{df}_{P,Q,\lambda}, \tag{19}$$

$$\text{BIC}(P, Q, \lambda) = -2\sum_{t\in[T]} \ell(\mathbf{X}_t|\mathbf{X}_{t-1:P}, \mathbf{z}_{t-1:Q}, \widehat{\boldsymbol{\Theta}}) + \log(T) \cdot \mathbf{df}_{P,Q,\lambda}. \tag{20}$$

To calculate $\mathbf{df}_{P,Q,\lambda}$, we decompose it into the sum of three components: 1) for each pair of the autoregressive coefficient $\widehat{\mathbf{A}}_p, \widehat{\mathbf{B}}_p$, the model has (M^2+N^2-1) degrees of freedom; 2) for the noise covariance $\widehat{\mathbf{\Sigma}}_r, \widehat{\mathbf{\Sigma}}_c$, the model has (M^2+N^2) degrees of freedom; and 3) for the auxiliary covariate functional parameters $\widehat{g}_{q,1}, \ldots, \widehat{g}_{q,D}$, inspired by the kernel ridge regression estimator in (15), we define the sum of their degrees of freedom as:

$$\mathbf{df}_q(\widehat{g}) = \operatorname{tr}\left\{\left[\widetilde{\mathbf{K}} + \lambda\left(\mathbf{I}_D \otimes \widehat{\mathbf{\Sigma}}_c \otimes \widehat{\mathbf{\Sigma}}_r\right)\right]^{-1}\widetilde{\mathbf{K}}\right\},$$

where $\widetilde{\mathbf{K}} = \left(T^{-1}\sum_{t\in[T]}\mathbf{z}_{t-q}\mathbf{z}_{t-q}^\top\right)\otimes\mathbf{K}$. As $\lambda \to 0$, we have $\mathbf{df}_q(\widehat{g}) \to MND$; namely each covariate has MN free parameters, which then reduces to the element-wise linear regression model. Empirically, we find that the BIC is a consistent lag selection criterion for our model.

4 Theoretical Analysis

This section presents the theoretical analyses of the MARAC model. We first formulate the condition under which the matrix and vector time series are *jointly stationary*. Under this condition, we then establish the consistency and asymptotic normality of the penalized MLE under *fixed* matrix dimensionality as $T \to \infty$. Finally, we consider the case where the matrix size goes to infinity as $T \to \infty$ and derive the convergence rate of the penalized MLE estimator and also the optimal order of the functional norm penalty tuning parameter λ. Without loss of generality, we assume that the matrix and vector time series have zero means, and we use $S = MN$ to denote the spatial dimensionality of the matrix data. All proofs are deferred to the supplemental material.

4.1 Stationarity Condition

To facilitate the theoretical analysis, we make an assumption for the vector time series $\mathbf{z}_t$, which greatly simplifies the theoretical analysis.

Assumption 4 *The D-dimensional auxiliary vector time series $\{\mathbf{z}_t\}_{t=-\infty}^{\infty}$ follows a stationary VAR($\widetilde{Q}$) process:*

$$\mathbf{z}_t = \sum_{\widetilde{q}=1}^{\widetilde{Q}} \mathbf{C}_{\widetilde{q}}\mathbf{z}_{t-\widetilde{q}} + \boldsymbol{\nu}_t, \tag{21}$$

where $\mathbf{C}_{\widetilde{q}} \in \mathbb{R}^{D\times D}$ *is the lag-*$\widetilde{q}$ *transition matrix and* $\boldsymbol{\nu}_t$ *has independent sub-Gaussian entries and is independent of* $\mathbf{E}_t$.

Remark 5 *With Assumption 4, we can combine the MARAC model for* $\mathbf{X}_t$ *in (4) and the VAR process for* $\mathbf{z}_t$ *in (21) as a single VAR*(L) *model, with* $L = \max(P, Q, \widetilde{Q})$, *as:*

$$\begin{bmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{bmatrix} = \sum_{l=1}^{L} \begin{bmatrix} (\mathbf{B}_l \otimes \mathbf{A}_l) \odot \mathbf{1}_{\{l\le P\}} & \mathbf{G}_l^\top \odot \mathbf{1}_{\{l\le Q\}} \\ \mathbf{O}_{D\times S} & \mathbf{C}_l \odot \mathbf{1}_{\{l\le \widetilde{Q}\}} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{t-l} \\ \mathbf{z}_{t-l} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_t \\ \boldsymbol{\nu}_t \end{bmatrix}, \tag{22}$$

with $\odot$ *being the Hadamard product between matrices and* $\mathbf{1}_{\{l\le C\}}$ *being a matrix with all elements taking values from* $\mathbb{1}_{\{l\le C\}}$ *and* $\mathbf{O}$ *is zero matrix. As (22) shows,* $\mathbf{z}_{t'}$ *can help forecast* $\mathbf{x}_t$ *where* $t > t'$ *but not the opposite, indicating that* $\mathbf{z}_t$ *is an exogenous predictor for* $\mathbf{x}_t$.

Given the joint vector autoregressive model in (22), we derive the conditions for $\mathbf{x}_t$ and $\mathbf{z}_t$ to be *jointly stationary* in Theorem 6.

Theorem 6 (MARAC Stationarity Condition) *Assume that Assumption 4 holds for the auxiliary time series* $\{\mathbf{z}_t\}_{t=-\infty}^{\infty}$, *and that the matrix time series* $\{\mathbf{X}_t\}_{t=-\infty}^{\infty}$ *is generated by the MARAC*(P, Q) *model in (1), then* $\{\mathbf{X}_t, \mathbf{z}_t\}_{t=-\infty}^{\infty}$ *are jointly stationary if and only if for any* $y \in \mathbb{C}$ *in the complex plane such that* $|y| \le 1$, *we have*

$$\det\left[\mathbf{I}_S - \sum_{p=1}^{P} (\mathbf{B}_p \otimes \mathbf{A}_p)\, y^p\right] \neq 0, \quad \det\left[\mathbf{I}_D - \sum_{\widetilde{q}=1}^{\widetilde{Q}} \mathbf{C}_{\widetilde{q}} y^{\widetilde{q}}\right] \neq 0. \tag{23}$$

The proof of Theorem 6 follows the proof of the stationarity of the joint autoregressive model in (22). As a special case where $P = \widetilde{Q} = 1$, the stationarity condition in (23) is equivalent to $\bar{\rho}(\mathbf{A}_1)\cdot\bar{\rho}(\mathbf{B}_1) < 1$ and $\bar{\rho}(\mathbf{C}_1) < 1$, where $\bar{\rho}(\cdot)$ is the spectral radius of a square matrix.

Based on Theorem 6, the stationarity of the matrix and vector time series relies on the stationarity of the autoregressive coefficients of the MARAC(P, Q) and VAR$(\widetilde{Q})$ models and the tensor coefficients $\boldsymbol{\mathcal{G}}_1, \ldots, \boldsymbol{\mathcal{G}}_Q$ do not affect the stationarity. The MARAC model can be extended to the joint autoregressive process (22) where the dynamics of $\mathbf{X}_t$ and $\mathbf{z}_t$ are modeled jointly. We stick to the simpler case here and leave the joint autoregression to future work.

4.2 Finite Spatial Dimension Asymptotics

In this subsection, we establish the consistency and asymptotic normality of the MARAC model estimators under the scenario that M, N are *fixed*. Given a fixed matrix dimensionality, the functional parameters $g_{q,d} \in \mathbb{H}_k$ can only be estimated at $S = MN$ fixed locations, and thus the asymptotic normality result is established for the corresponding tensor coefficient $\widehat{\boldsymbol{\mathcal{G}}}_q$. In Section 4.3, we will discuss the *double* asymptotics when both $S, T \to \infty$. For the remainder of the paper, we denote the true model coefficient with an asterisk superscript, such as $\mathbf{A}_1^*, \mathbf{B}_1^*, \boldsymbol{\mathcal{G}}_1^*$ and $\boldsymbol{\Sigma}^*$.

To start with, we make an assumption on the Gram matrix $\mathbf{K}$:

Assumption 7 *The minimum eigenvalue of* $\mathbf{K}$ *is bounded by a positive constant* $\underline{c}$*, i.e.* $\underline{\rho}(\mathbf{K}) = \underline{c} > 0$.

As a result of Assumption 7, every $\boldsymbol{\mathcal{G}}_q^*$ has a unique kernel decomposition: $\mathbf{vec}\left(\boldsymbol{\mathcal{G}}_q^*\right) = (\mathbf{I}_D \otimes \mathbf{K})\boldsymbol{\gamma}_q^*$. With this additional assumption, the first theoretical result we establish is the consistency of the covariance matrix estimator $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_c \otimes \widehat{\boldsymbol{\Sigma}}_r$, which we summarize in Proposition 8.

Proposition 8 (Covariance Consistency) *Assume that* $\lambda \to 0$ *as* $T \to \infty$ *and* S *is fixed, and Assumption 4, 7 and the stationarity condition in Theorem 6 hold, , then* $\widehat{\boldsymbol{\Sigma}} \xrightarrow{p} \boldsymbol{\Sigma}^*$.

Given this result, we can further establish the asymptotic normality of the other model estimators:

Theorem 9 (Asymptotic Normality) *Assume that the matrix time series* $\{\mathbf{X}_t\}_{t=-\infty}^{\infty}$ *follows the MARAC*(P, Q) *model (1) with i.i.d. noise* $\{\mathbf{E}_t\}_{t=-\infty}^{\infty}$ *following (3) and Assumption 4, 7 and the stationarity condition in Theorem 6 hold and* $\lambda = o(T^{-1/2})$*. Additionally, assume that* $\underline{\rho}(\mathrm{Var}([\mathbf{vec}\left(\mathbf{X}_t\right)^\top, \mathbf{z}_t^\top]^\top)) = \underline{c}^{'} > 0$*. Then suppose* M, N *are fixed and* P, Q *are known and denote* $\mathbf{A}_p, \mathbf{B}_p^\top$ *as* $\boldsymbol{\alpha}_p$ *and* $\boldsymbol{\beta}_p$ *for any* $p \in [P]$*, the penalized MLE of the*

MARAC(P, Q) *model is asymptotically normal:*

$$\sqrt{T}\begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 \otimes \widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\beta}_1^* \otimes \boldsymbol{\alpha}_1^* \\ \vdots \\ \widehat{\boldsymbol{\beta}}_P \otimes \widehat{\boldsymbol{\alpha}}_P - \boldsymbol{\beta}_P^* \otimes \boldsymbol{\alpha}_P^* \\ \mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{G}}}_1 - \boldsymbol{\mathcal{G}}_1^*\right) \\ \vdots \\ \mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{G}}}_Q - \boldsymbol{\mathcal{G}}_Q^*\right) \end{bmatrix} \xrightarrow{d.} \mathcal{N}\left(\mathbf{0}, \mathbf{V}\boldsymbol{\Xi}\mathbf{V}^\top\right), \tag{24}$$

where $\mathbf{V}$ *is:*

$$\mathbf{V} = \begin{bmatrix} diag(\mathbf{V}_1, \ldots, \mathbf{V}_P) & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{QD} \otimes \mathbf{K} \end{bmatrix}, \quad \mathbf{V}_p = [\boldsymbol{\beta}_p^* \otimes \mathbf{I}_{M^2}, \mathbf{I}_{N^2} \otimes \boldsymbol{\alpha}_p^*],$$

and $\boldsymbol{\Xi} = \mathbf{H}^{-1}\mathrm{E}\left[\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t\right]\mathbf{H}^{-1}$*, and* $\mathbf{W}_t$ *is defined as:*

$$\mathbf{W}_t = \left[\mathbf{W}_{0,t} \otimes \mathbf{I}_M, \mathbf{I}_N \otimes \mathbf{W}_{1,t}, [\mathbf{z}_{t-1}^\top, \ldots, \mathbf{z}_{t-Q}^\top] \otimes \mathbf{K}\right],$$

where $\mathbf{H} = \mathrm{E}[\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t] + \boldsymbol{\zeta}\boldsymbol{\zeta}^\top$ *with* $\boldsymbol{\zeta}^\top = [(\boldsymbol{\alpha}_1^*)^\top, \cdots, (\boldsymbol{\alpha}_P^*)^\top, \mathbf{0}^\top]$ *and:*

$$\mathbf{W}_{0,t} = [\mathbf{B}_1^*\mathbf{X}_{t-1}^\top, \ldots, \mathbf{B}_P^*\mathbf{X}_{t-P}^\top], \quad \mathbf{W}_{1,t} = [\mathbf{A}_1^*\mathbf{X}_{t-1}, \ldots, \mathbf{A}_P^*\mathbf{X}_{t-P}].$$

The asymptotic distribution (24) indicates that all parameters are $\sqrt{T}$-consistent. Under a fixed matrix dimensionality S, the functional parameters $g_{q,d} \in \mathbb{H}_k$ are estimated only at fixed locations. Hence, the convergence is at a parametric rate just like the autoregressive coefficient.

4.3 High Spatial Dimension Asymptotics

The previous section presents the asymptotic normality of the MARAC estimators under a *fixed* matrix dimensionality S. In this section, we relax this assumption and establish the convergence rate of the MARAC estimators when $S, T \to \infty$. For technical reasons, we assume that all entries of $\mathbf{E}_t$ are i.i.d. normally-distributed random variables following $\mathcal{N}(0, \sigma^2)$.

To establish the convergence rate of the MARAC estimators when $S, T \to \infty$, we need to make several additional assumptions.

Assumption 10 *The spatial locations of the rows and columns of* $\mathbf{X}_t$ *are sampled independently from a uniform distribution on* $[0, 1]$.

Assumption 11 *The spatial kernel function* $k(\cdot, \cdot)$ *can be decomposed into the product of a row kernel* $k_1(\cdot, \cdot)$ *and a column kernel* $k_2(\cdot, \cdot)$ *that satisfies* $k((u, v), (s, t)) = k_1(u, s)k_2(v, t)$. *Both* k_1, k_2 *have their eigenvalues decaying at a polynomial rate:* $\lambda_j(k_1) \asymp j^{-r_0}, \lambda_j(k_2) \asymp j^{-r_0}$ *with* $r_0 \in (1/2, 2)$.

Assumption 11 elicits a simple eigen-spectrum characterization of the spatial kernel $k(\cdot, \cdot)$, whose eigenvalue can be written as $\lambda_i(k_1)\lambda_j(k_2)$. Also, the Gram matrix $\mathbf{K}$ is separable, i.e. $\mathbf{K} = \mathbf{K}_2 \otimes \mathbf{K}_1$ and all eigenvalues of $\mathbf{K}$ have the form: $\rho_i(\mathbf{K}_1)\rho_j(\mathbf{K}_2)$, where $\mathbf{K}_1 \in \mathbb{R}^{M\times M}, \mathbf{K}_2 \in \mathbb{R}^{N\times N}$ are the Gram matrix for the kernel k_1, k_2, respectively.

Under Assumption 10, we further have $\rho_i(\mathbf{K}_1) \to M\lambda_i(k_1)$ and $\rho_j(\mathbf{K}_2) \to N\lambda_j(k_2)$, as $M, N \to \infty$. Combined with Assumption 11, we can characterize the eigenvalues of $\mathbf{K}$ as $S(ij)^{-r_0}$. We refer our readers to Koltchinskii and Giné (2000); Braun (2006) for more references about the eigen-analysis of the kernel Gram matrix. One can generalize Assumption 10 to non-uniform sampling, but here, we stick to this simpler setting.

In Assumption 11, we assume the kernel separability to accommodate the grid structure of the spatial locations. We do not restrict r_0 to be an integer but just a parameter that characterizes the smoothness of the functional parameters. With these assumptions, we are now ready to present the main result in Theorem 12.

Theorem 12 (Asymptotics for High-Dimensional MARAC) *Assume that Assumptions 4, 10 and 11 hold and* $\mathbf{X}_t$ *is generated by the MARAC*(P, Q) *model (1) with* $\mathbf{E}_t$ *having i.i.d.* $\mathcal{N}(0, \sigma^2)$ *entries. Then as* $S, T \to \infty$ *(*D *is fixed) and* $S \log S/T \to 0$, *and under the additional assumptions that:*

1. $M = O(\sqrt{S}), N = O(\sqrt{S})$;
2. $\gamma_S \coloneqq \lambda/S \to 0$ *and* $\gamma_S \cdot S^{r_0} \to C_1$ *as* $S \to \infty$, *with* $0 < C_1 \leq \infty$;
3. $\underline{\rho}(\mathbf{\Sigma}^*_{\mathbf{x},\mathbf{x}} - (\mathbf{\Sigma}^*_{\mathbf{z},\mathbf{x}})^\top(\mathbf{\Sigma}^*_{\mathbf{z},\mathbf{z}})^{-1}\mathbf{\Sigma}_{\mathbf{z},\mathbf{x}}) = c_{0,S} > 0$ *as* $S, T \to \infty$, *where* $\mathbf{\Sigma}^*_{\mathbf{x},\mathbf{x}}, \mathbf{\Sigma}^*_{\mathbf{z},\mathbf{z}}, \mathbf{\Sigma}^*_{\mathbf{z},\mathbf{x}}$ *are* $\mathrm{Var}(\mathbf{x}_t)$, $\mathrm{Var}(\mathbf{z}_t)$ *and* $\mathrm{Cov}(\mathbf{z}_t, \mathbf{x}_t)$, *respectively.* $c_{0,S}$ *is a constant that only relates to* S;

4. For any S, we have $0 < \underline{\rho}(\mathbf{K}) < \bar{\rho}(\mathbf{K}) \leq C_0$, where C_0 is a finite constant,

then we have:

$$\frac{1}{\sqrt{P}S}\sqrt{\sum_{p=1}^{P}\left\|\widehat{\mathbf{B}}_p \otimes \widehat{\mathbf{A}}_p - \mathbf{B}_p^* \otimes \mathbf{A}_p^*\right\|_{\mathrm{F}}^2} \lesssim O_P\left(\sqrt{\frac{C_g \cdot \gamma_S}{c_{0,S} \cdot S}}\right) + O_P\left(\sqrt{\frac{D}{c_{0,S} \cdot TS}}\right), \tag{25}$$

where $C_g = \sum_{q=1}^{Q}\sum_{d=1}^{D}\|g_{q,d}\|_{\mathbb{H}_k}^2$. Furthermore, we also have:

$$\begin{aligned}&\sqrt{(TS)^{-1}\sum_{t=1}^{T}\left\|\sum_{q=1}^{Q}\left(\widehat{\boldsymbol{\mathcal{G}}}_q - \boldsymbol{\mathcal{G}}_q^*\right)\bar{\times}\mathbf{z}_{t-q}\right\|_{\mathrm{F}}^2}\\ &\lesssim O_P\left(\frac{\sqrt{\gamma_S^{-1/2r_0}}}{\sqrt{T}\sqrt[4]{S}}\right) + O_P(\sqrt{\gamma_S}) + O_P\left(\frac{1}{\sqrt{S}}\right) + O_P\left(\frac{\sqrt{\gamma_S^{-1}}}{\sqrt{TS}}\right).\end{aligned} \tag{26}$$

In Theorem 12, (25) gives the error bound of the autoregressive coefficients and (26) gives the error bound of the prediction made by the auxiliary time series, which contains the functional parameter estimators. As a special case of (25) where $\gamma_S = 0$ and S is fixed, the convergence rate for the autoregressive coefficients is $O_P(T^{-1/2})$, which reproduces the result in Theorem 9. For the discussion below, we use AR_{err} and AC_{err} as acronyms for the quantity on the left-hand side of (25) and (26).

Remark 13 (Optimal Choice of λ and Phase Transition) *According to our proof, the error bound (26) can be decomposed into the sum of:*

- *nonparametric error:* $O_P\left(\sqrt{\frac{\gamma_S^{-1/2r_0}}{T\sqrt{S}}}\right) + O_P(\sqrt{\gamma_S})$,
- *autoregressive error:* $O_P\left(\sqrt{\gamma_S}\right) + O_P\left(S^{-1/2}\right) + O_P\left(T^{-1/2}\right) + O_P\left(\sqrt{\frac{\gamma_S^{-1}}{TS}}\right)$,

where the autoregressive error stems from the estimation error of $\widehat{\mathbf{B}}_p \otimes \widehat{\mathbf{A}}_p$. The nonparametric error resembles the result of nonparametric regression with RKHS norm penalty (Cui et al., 2018), where if the number of data points is n and penalty tuning parameter is λ, then the nonparametric error is bounded by $O_P(\sqrt{\lambda^{-1/2r_0}/n}) + O_P(\sqrt{\lambda})$ with an optimal $\lambda \asymp n^{-2r_0/(2r_0+1)}$. In our model, if there is no autoregressive error, the optimal tuning parameter satisfies $\gamma_S \asymp (T\sqrt{S})^{-2r_0/(2r_0+1)}$. The number of data points in our case is TS, and we are short of $\sqrt{S}$ in the optimal tuning parameter due to Assumption 11, where the

eigenvalues of $\mathbf{K}$*, ordered as* $\rho_1(\mathbf{K}) \geq \cdots \rho_i(\mathbf{K}) \geq \cdots$*, decay slower than* i^{-2r_0}*. This is a special result for matrix-shaped data. It is also noteworthy that under the condition that* $S \log S/T \to 0$*, the autoregressive error dominates the nonparametric error.*

To simplify the discussion of the optimal order of γ_S*, we assume that* $S = T^c$*, where* $c < 1$ *is a constant. Under this condition, when* $P, Q \geq 1$*, the optimal tuning parameter* $\gamma_S = \lambda/S$ *shows an interesting phase transition phenomenon under different spatial smoothness* r_0 *and matrix dimensionality* $c = \log_T S$*, which we summarize in Table 1.*

r_0	$\log_T S$	Optimal γ_S	Estimator Error
$[1, 2)$	$[\frac{1}{2r_0-1}, 1)$	$O((TS)^{-\frac{1}{2}})$	$\mathrm{AR}_{err} = O_P(T^{-\frac{1}{4}} S^{-\frac{3}{4}})$ $\mathrm{AC}_{err} = O_P(S^{-\frac{1}{2}})$
$[1, 2)$	$(0, \frac{1}{2r_0-1})$	$O(S^{-r_0})$	$\mathrm{AR}_{err} = O_P(S^{-\frac{r_0+1}{2}})$ $\mathrm{AC}_{err} = O_P(S^{-\frac{1}{2}})$
$(\frac{1}{2}, 1)$	$[2r_0 - 1, 1)$	$O(S^{-r_0(2r_0-1)})$	$\mathrm{AR}_{err} = O_P(S^{-\frac{r_0(2r_0-1)+1}{2}})$ $\mathrm{AC}_{err} = O_P(S^{-\frac{1}{2}})$
$(\frac{1}{2}, 1)$	$(0, 2r_0 - 1)$	$O((T\sqrt{S})^{-\frac{2r_0}{2r_0+1}})$	$\mathrm{AR}_{err} = O_P((TS)^{-\frac{1}{2}}) + O_P((T\sqrt{S})^{-\frac{r_0}{2r_0+1}} S^{-\frac{1}{2}})$ $\mathrm{AC}_{err} = O_P(S^{-\frac{1}{2}}) + O_P((T\sqrt{S})^{-\frac{r_0}{2r_0+1}})$

Table 1: Summary of optimal tuning parameter γ_S and estimators error following (25) and (26), under the assumption that $c_{0,S} \geq c_0 > 0$, for all S and $S = T^c$ for some constant $0 < c < 1$ such that $S \log S/T \to 0$. AR_{err} and AC_{err} are the quantity on the left-hand side of (25) and (26).

Based on the results in Table 1, the faster S *grows with respect to* T*, the smaller the optimal tuning parameter* γ_S *is. This is an intuitive result since when one has more spatial locations, the observations are denser, and thus less smoothing is needed. Furthermore, we achieve an optimal tuning order of* γ_S *that is close to the classic nonparametric optimal rate at* $(TS)^{-2r_0/(2r_0+1)}$ *only under the regime where* $1/2 < r_0 < 1$ *and* $\log_T S < 2r_0 - 1$*. This regime specifies the scenario where the functional parameter is relatively unsmooth and the spatial dimensionality grows slowly with respect to* T*. Only under this regime will the discrepancy between the nonparametric error and the autoregressive error remain small, leading to an optimal tuning parameter close to the result of nonparametric regression.*

In (25), the constant $c_{0,S}$ appears in the error bound of the autoregressive term. This constant characterizes the spatial correlation of the matrix time series $\mathbf{X}_t$, conditioning on the auxiliary vector time series $\mathbf{z}_t$ and can vary across different assumptions made on the covariances of $\mathbf{E}_t$ and $\boldsymbol{\nu}_t$. In Table 1, we assume that $c_{0,S} \geq c_0 > 0$ for some universal constant c_0. Unfortunately, in practice, it is common to have $c_{0,S} \to 0$ as $S \to \infty$, which makes the autoregressive coefficient converge at a slower rate but does not affect the functional parameter convergence. We leave the constant $c_{0,S}$ here in (25) to give a general result and leave the characterization of $c_{0,S}$ under specific assumptions for future works.

5 Simulation Experiments

5.1 Consistency and Convergence Rate

In this section, we validate the consistency and convergence rate of the MARAC estimators. We consider a simple setup with $P = Q = 1$ and $D = 3$ and simulate the autoregressive coefficients $\mathbf{A}_1^*, \mathbf{B}_1^*$ such that they satisfy the stationarity condition in Theorem 6. We specify both $\mathbf{A}_1^*, \mathbf{B}_1^*$ and $\boldsymbol{\Sigma}_r^*, \boldsymbol{\Sigma}_c^*$ to have symmetric banded structures. To simulate g_1, g_2, g_3 (we drop the lag subscript) from the RKHS $\mathbb{H}_k$, we choose $k(\cdot, \cdot)$ to be the Lebedev kernel (Kennedy et al., 2013) and generate g_1, g_2, g_3 randomly from Gaussian processes with the Lebedev kernel as the covariance kernel. Finally, we simulate the auxiliary vector time series $\mathbf{z}_t \in \mathbb{R}^3$ from a VAR(1) process. We include more details and visualizations of the simulation setups in the supplemental material.

The evaluation metric is the rooted mean squared error (RMSE), defined as $\text{RMSE}(\widehat{\boldsymbol{\Theta}}) = \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\text{F}}/\sqrt{d(\boldsymbol{\Theta}^*)}$, where $d(\boldsymbol{\Theta}^*)$ is the number of elements in $\boldsymbol{\Theta}^*$. We consider $\boldsymbol{\Theta} \in \{\mathbf{B}_1 \otimes \mathbf{A}_1, \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r, \boldsymbol{\mathcal{G}}_1, \boldsymbol{\mathcal{G}}_2, \boldsymbol{\mathcal{G}}_3\}$ and we report the average RMSE for $\boldsymbol{\mathcal{G}}_1, \boldsymbol{\mathcal{G}}_2, \boldsymbol{\mathcal{G}}_3$. The dataset is configured with $M \in \{5, 10, 20, 40\}$ and $N = M$. For each M, we train the MARAC model with $P = Q = 1$ over $T_{\text{train}} \in \{1, 5, 10, 20, 40, 80, 160\} \times 10^2$ frames of the matrix time series and choose the tuning parameter λ based on the prediction RMSE over a held-out validation set with $T_{\text{val}} = T_{\text{train}}/2$ and we validate the prediction performance over a $5,000$-frame testing set. We simulate a sufficiently long time series and choose the training set starting from the first frame and the validation set right after the training set.

The testing set is always fixed as the last $5,000$ frames of the time series. All results are reported with 20 repetitions in Figure 2.

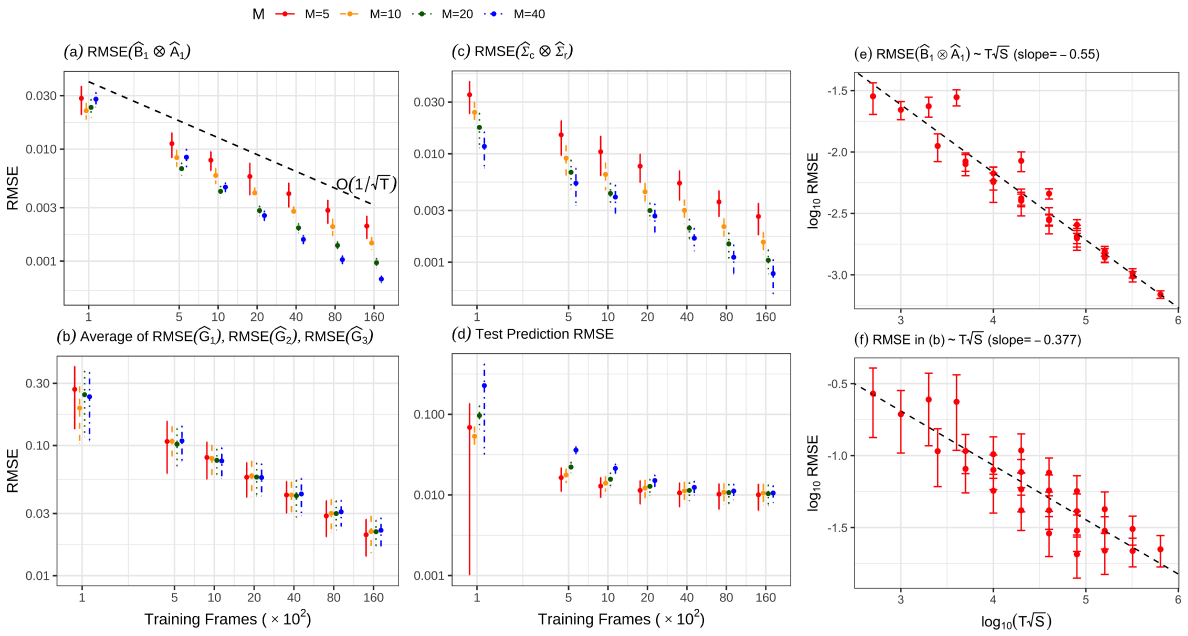

Figure 2: Panel (a), (b), (c) show the RMSE of the penalized MLE of the MARAC model. Panel (d) shows the testing set prediction RMSE subtracted by 1, where 1 is the noise variance of the simulated time series. Panels (a)-(d) have both axes plotted in $\log_{10}$ scale. (e) and (f) are RMSE of the autoregressive parameters and auxiliary covariates parameters under different $T\sqrt{S}$, plotted with both axes in $\log_{10}$ scale together with a fitted linear regression line.

The result shows that all model estimators are consistent and the convergence rate, under a fixed spatial dimensionality, is close to $1/\sqrt{T}$ (the black line in panel (a) shows a reference line of $O(1/\sqrt{T})$), echoing the result in Theorem 9. As the spatial dimensionality S increases, the RMSE for $\widehat{\mathbf{B}}_1 \otimes \widehat{\mathbf{A}}_1$ becomes even smaller, echoing the result in (25) and Table 1. The RMSE of the nonparametric estimators $\widehat{g}_1, \widehat{g}_2, \widehat{g}_3$, under a fixed spatial dimensionality, also decay at a rate of $1/\sqrt{T}$, echoing the result in Theorem 9 as well. The RMSE of the covariance matrix estimator $\widehat{\mathbf{\Sigma}}_c \otimes \widehat{\mathbf{\Sigma}}_r$ suggests that it is consistent, confirming the result of Proposition 8 and shows a convergence rate similar to $\widehat{\mathbf{B}}_1 \otimes \widehat{\mathbf{A}}_1$, though we did not provide the exact convergence rate theoretically.

In this simulation, we fix the variance of each element of $\mathbf{vec}\,(\mathbf{E}_t)$ to be unity. There-

fore, the optimal testing set prediction RMSE should be unity. When plotting the test prediction RMSE in (d), we subtract 1 from all RMSE results and thus the RMSE should be interpreted as the RMSE for the *signal* part of the matrix time series. The test prediction RMSE for all cases converges to zero, and for matrices of higher dimensionality, we typically require more training frames to reach the same prediction performance.

To validate the theoretical result of the high-dimensional MARAC in Theorem 12, we also plot the RMSE of $\widehat{\mathbf{B}}_1 \otimes \widehat{\mathbf{A}}_1$ and $\widehat{g}_1, \widehat{g}_2, \widehat{g}_3$ against $T\sqrt{S}$ in panel (e) and (f) of Figure 2. The trend line is fitted by linear regression, and it shows that $\widehat{\mathbf{B}}_1 \otimes \widehat{\mathbf{A}}_1$ converges roughly at the rate of $1/\sqrt{T}\sqrt[4]{S}$, which indicates that $c_{0,S} \asymp 1/\sqrt{S}$ under this specific simulation setup. It also shows that the functional parameter's convergence rate is around $(T\sqrt{S})^{-3/8}$, which coincides with our simulation setup where $r_0 \approx 3/4$ and the theoretical result in the last row of Table 1.

All the results reported in Figure 2 are based on the penalized MLE framework without the kernel truncation introduced in Section 3.2. Kernel truncation speeds up the computation, especially when the matrix dimensionality is high, at the cost of over-smoothing the functional parameter estimates. We illustrate the performance of the kernel truncation method in the supplemental material.

5.2 Lag Selection Consistency

In Section 3.3, we propose to select the lag parameters P and Q of the MARAC model using information criteria such as AIC and BIC. To validate the consistency of these model selection criteria, we simulate data from a MARAC$(2, 2)$ model with 5×5 matrix dimensionality. We consider a candidate model class with $1 \leq P, Q \leq 4$ and each model is fitted with $T \in \{1, 2, 4, 8\} \times 10^3$ frames with λ being chosen from a held-out validation set. In Table 2, we report the proportion of times that AIC and BIC select the correct P, Q individually (first two numbers in each parenthesis), and (P, Q) jointly (last number in each parenthesis) from 100 repetitions.

From Table 2, we find that AIC tends to select the model with more autoregressive lags but BIC performs consistently better under large sample sizes. This coincides with the findings in Hsu et al. (2021) for the matrix autoregression model.

	$T = 1 \times 10^3$	$T = 2 \times 10^3$	$T = 4 \times 10^3$	$T = 8 \times 10^3$
AIC	(.54, .99, .53)	(.55, .97, .53)	(.59, .96, .55)	(.65, .94, .59)
BIC	(1.00, .09, .09)	(.99, .56, .56)	(.97, .97, .94)	(.96, .99, .95)

Table 2: Probability that AIC and BIC select the correct P (first number), Q (second number) and (P, Q) (third number) from 100 repetitions.

5.3 Comparison with Alternative Methods

We compare our MARAC model against other competing methods for the matrix autoregression task. We simulate the matrix time series $\mathbf{X}_t$ from a MARAC(P, Q) model, with $P = Q \in \{1, 2, 3\}$, and the vector time series $\mathbf{z}_t \in \mathbb{R}^3$ from VAR(1). The dataset is generated with $T_{\text{train}} = T_{\text{val}} = T_{\text{test}} = 2000$. Under each (P, Q), we simulate with varying matrix dimensionality with $M = N \in \{5, 10, 20, 40\}$. We evaluate the performance of each method via the testing set prediction RMSE. Each simulation scenario is repeated 20 times.

Under each P, Q, M, N specification, we consider the following five competing methods besides our own MARAC(P, Q) model.

1. MAR (Chen et al., 2021):

$$\mathbf{X}_t = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{X}_{t-p} \mathbf{B}_p^\top + \mathbf{E}_t, \mathbf{vec}\,(\mathbf{E}_t) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r).$$

2. MAR with fixed-rank co-kriging (MAR+FRC) (Hsu et al., 2021):

$$\mathbf{X}_t = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{X}_{t-p} \mathbf{B}_p^\top + \mathbf{E}_t, \mathbf{vec}\,(\mathbf{E}_t) \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2 \mathbf{I} + \mathbf{F}\mathbf{M}\mathbf{F}^\top),$$

   where $\mathbf{F} \in \mathbb{R}^{MN \times QD}$ is the multi-resolution spline basis (Tzeng and Huang, 2018).

3. MAR followed by a tensor-on-scalar linear model (MAR+LM) (Li and Zhang, 2017):

$$\mathbf{X}_t - \sum_{p=1}^{P} \widehat{\mathbf{A}}_p \mathbf{X}_{t-p} \widehat{\mathbf{B}}_p^\top = \sum_{q=1}^{Q} \boldsymbol{\mathcal{G}}_q \bar{\times} \mathbf{z}_{t-q} + \mathbf{E}_t, \mathbf{vec}\,(\mathbf{E}_t) \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2 \mathbf{I}), \tag{27}$$

   where $\widehat{\mathbf{A}}_p, \widehat{\mathbf{B}}_p$ come from a pre-trained MAR model and $\boldsymbol{\mathcal{G}}_q$ can be a low-rank tensor. The MAR+LM model can be considered as a two-step procedure for fitting the MARAC model.

4. Pixel-wise autoregression (Pixel-AR): for each $i \in [M], j \in [N]$, we have:

$$[\mathbf{X}_t]_{ij} = \alpha_{ij} + \sum_{p=1}^{P} \beta_{ijp}[\mathbf{X}_{t-p}]_{ij} + \sum_{q=1}^{Q} \boldsymbol{\gamma}_{ijq}^{\top}\mathbf{z}_{t-q} + [\mathbf{E}_t]_{ij}, \quad [\mathbf{E}_t]_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2).$$

5. Vector Autoregression with Exogenous Predictor (VARX), which vectorizes the matrix time series and stacks them up with the vector time series as predictors.

The results of the average prediction RMSE obtained from the 20 repeated runs are plotted in Figure 3. Overall, our MARAC model outperforms the other competing methods under varying matrix dimensionality and lags. We make two additional remarks. First, when the matrix size is small (e.g., 5×5), the vector autoregression model (VARX) performs almost as well as the MARAC model and is better than other methods. However, the performance of the VARX model gets worse quickly as the matrix becomes larger, indicating that sufficient dimension reduction is needed for dealing with large matrix time series. The MARAC model is a parsimonious version of VARX for such purposes. Secondly, the MAR, MAR with fixed-rank co-kriging (MAR+FRC), and two-step MARAC (MAR+LM) all perform worse than MARAC. This shows that when the auxiliary time series predictors are present, it is sub-optimal to remove them from the model (MAR), incorporate them implicitly in the covariance structure (MAR+FRC), or fit them separately in a tensor-on-scalar regression model (MAR+LM). Putting both matrix predictors and vector predictors in a unified framework like MARAC can be beneficial for improving prediction performances.

6 Application to Global Total Electron Content Forecast

For real data applications, we consider the problem of predicting the global total electron content (TEC) distribution, which we briefly introduce in Section 1. The TEC data we use is the IGS (International GNSS Service) TEC data, which are freely available from the National Aeronautics and Space Administration (NASA) Crustal Dynamics Data Information System (Hernández-Pajares et al., 2009). The spatial-temporal resolution of the data is 2.5°(latitude) $\times$ 5°(longitude) $\times$ 15(minutes). We downloaded the data for September 2017, and the whole month of data form a matrix time series with $T = 2880$ and

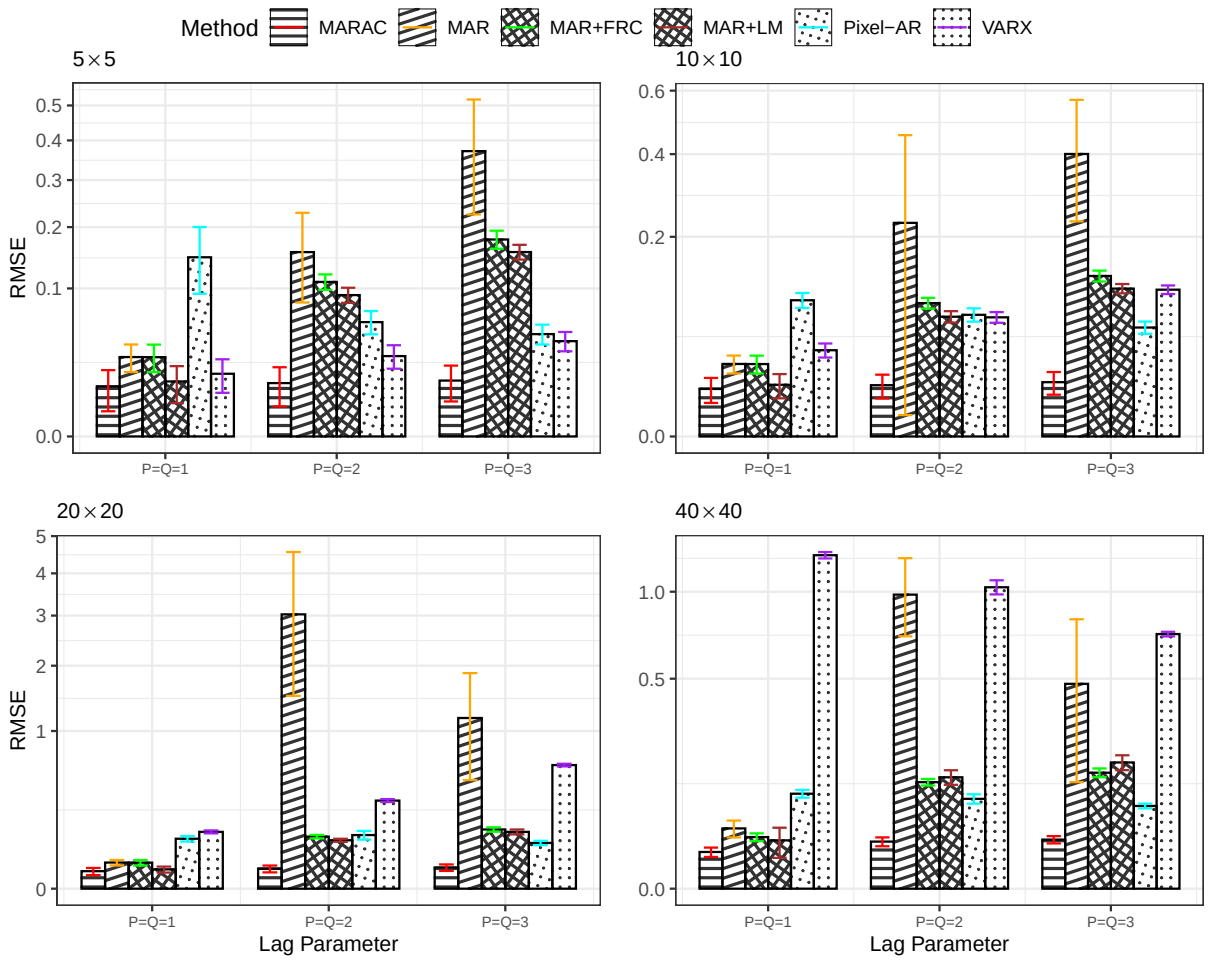

Figure 3: Testing set prediction RMSE comparison across six competing methods on the matrix autoregression task. Four panels correspond to four different matrix dimensionality (labeled on the top-left corner of each panel). Test prediction RMSE is subtracted by 1 for better visualization, where 1 is the noise variance of the simulated data. Error bar shows 95% CI of the 20 repeated runs. We rearrange the spacing between ticks along the y-axis using a square root transformation for better visualization.

$M = 71, N = 73$. For the auxiliary covariates, we download the 15-minute resolution IMF Bz and Sym-H time series, which are parameters related to the near-Earth magnetic field and plasma (Papitashvili et al., 2014). We also download the daily F10.7 index, which measures the solar radio flux at 10.7 cm, as an additional auxiliary predictor. The IMF Bz and Sym-H time series are accessed from the OMNI dataset (Papitashvili and King, 2020) and the F10.7 index is accessed from the NOAA data repository (Tapping, 2013). These covariates measure the solar wind strengths. Strong solar wind might lead to geomagnetic

storms that could increase the global TEC significantly.

We formulate our MARAC model for the TEC prediction problem as:

$$\text{TEC}_{t+h} = \sum_{p=1}^{P} \mathbf{A}_p \text{TEC}_{t-p} \mathbf{B}_p^\top + \sum_{q=1}^{Q} \boldsymbol{\mathcal{G}}_q \bar{\times} \mathbf{z}_{t-q} + \mathbf{E}_t, \tag{28}$$

where h is the forecast latency time and $\mathbf{z}_t \in \mathbb{R}^3$ includes the IMF Bz, Sym-H and F10.7 indices at time t. We consider the forecasting scenario with $h \in \{1, 2, \ldots, 24\}$, which corresponds to making forecasts from 15 minutes ahead up to 6 hours ahead. At each latency time, we fit our MARAC(P, Q) model following (28) with $1 \leq P, Q \leq 3$. We fit the MARAC model with kernel truncation approximation using $R = 121$ basis functions from the truncated Lebedev kernel. As a comparison, we also fit the MAR model with $1 \leq P \leq 3$ and the MAR+LM model with $1 \leq P, Q \leq 3$, see the definition of MAR+LM model in (27). As a benchmark, we consider using TEC_{t-1} to predict TEC_{t+h} and name it the *persistence model*.

The $2,880$ frames of matrix data are split into a 70% training set, 15% validation set, and a 15% testing set following the chronological order. We choose the tuning parameter λ for MARAC based on the validation set prediction RMSE. The lag parameters P, Q are chosen for all models based on the BIC. To increase computational speed, we assume that matrices $\boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_c$ are diagonal when fitting all models. We zero-meaned all sets of data using the mean of the matrix and vector time series of the training set.

In Figure 4(A), we report the pixel-wise prediction RMSE on the testing set. The result shows that when the latency time is low, the matrix autoregressive (MAR) model is sufficient for making the TEC prediction. As the latency time increases to around 4 to 5 hours, the auxiliary time series helps improve the prediction performance as compared to the MAR model. This coincides with the domain intuition that the disturbances from the solar wind to Earth's ionosphere will affect the global TEC distribution but with a delay in time of up to several hours. The additional prediction gain from incorporating the auxiliary covariates vanishes as one further increases the latency time, indicating that the correlation of the solar wind and global TEC is weak beyond a 6-hour separation.

In Figure 4(B), we visualize an example of the TEC prediction across the competing methods under the 4-hour latency time scenario (i.e., h=16). The MAR and MAR+LM results are similar and do not resemble the ground truth very well. The global TEC typically

has two peaks located symmetrically around the equator, and both models fail to capture this as they provide a single patch in the middle. The MARAC model, however, can capture this fine-scale structure in its prediction. To further showcase the MARAC model prediction result, we decompose the prediction from the autoregressive component and the auxiliary covariates component and visualize them separately. The auxiliary covariate component highlights a sub-region in the southern hemisphere with high TEC values, complementing the prediction made by the autoregressive component.

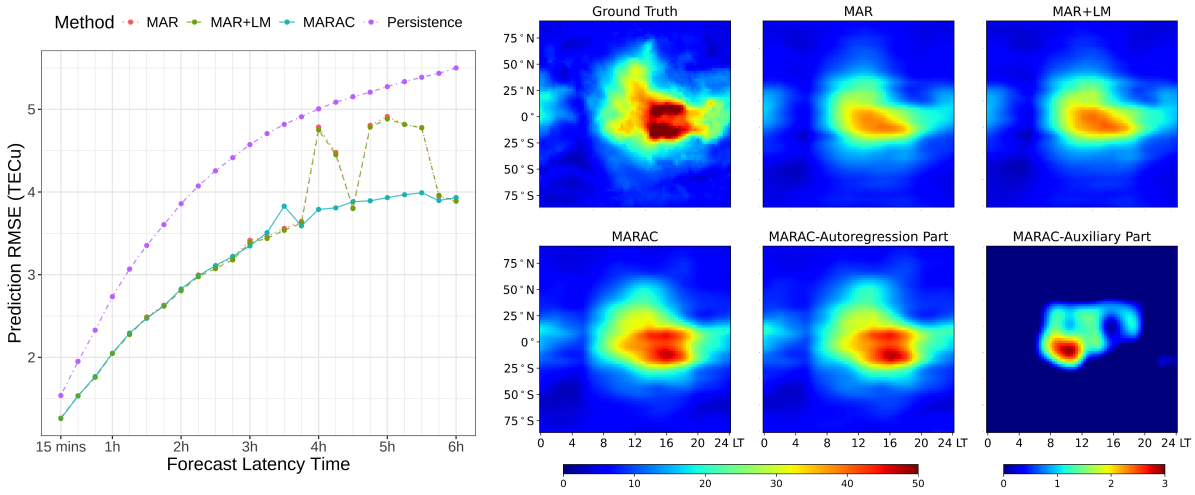

Figure 4: IGS TEC prediction results. Panel (A) shows the testing set prediction RMSE across four competing methods under 24 different latency times. Panel (B) shows an example of the predicted TEC at 10:45:00 UT, 2017-Sep-28, under the 4-hour latency time scenario. Note that the "MARAC-Auxiliary Part" plot has a different color bar underneath it and that color bar applies to it exclusively.

7 Summary

In this paper, we propose a new methodology for spatial-temporal matrix autoregression with non-spatial exogenous vector covariates. The model has an autoregressive component with bi-linear transformations on the lagged matrix predictors and an additive auxiliary covariate component with tensor-vector product between a tensor coefficient and the lagged vector covariates. We propose a penalized MLE estimation approach with a squared RKHS

norm penalty and establish the estimator asymptotics under fixed and high matrix dimensionality. The model efficacy has been validated using both numerical experiments and an application to the global TEC forecast.

The application of our model can be extended to other spatial data with exogenous, non-spatial predictors and is not restricted to matrix-valued data but can be generalized to the tensor setting and potentially data without grid structure or containing missing data. Furthermore, our model nests a simpler model that does not contain the autoregressive term, i.e. $P = 0$, and thus can be applied to matrix-on-scalar regression with spatial data. We leave the discussions for these setups to future research.

Supplementary Materials

The supplemental material contains technical proofs of all theorems and propositions of the paper and additional details and results of the simulation experiments.

Acknowledgements

The authors thank Shasha Zou, Zihan Wang, and Yizhou Zhang for helpful discussions on the TEC data. YC acknowledges support from NSF DMS 2113397 and NSF PHY 2027555.

References

Hirotogu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. *Selected papers of hirotugu akaike*, pages 199–213, 1998.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of Descent Methods for Semi-Algebraic and Tame Problems: Proximal Algorithms, Forward–Backward Splitting, and Regularized Gauss–Seidel Methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

Mikio L Braun. Accurate Error Bounds for the Eigenvalues of the Kernel Matrix. *The Journal of Machine Learning Research*, 7:2303–2328, 2006.

T Tony Cai and Ming Yuan. Minimax and Adaptive Prediction for Functional Linear Regression. *Journal of the American Statistical Association*, 107(499):1201–1216, 2012.

Rong Chen, Han Xiao, and Dan Yang. Autoregressive Models for Matrix-valued Time Series. *Journal of Econometrics*, 222(1):539–560, 2021.

Guang Cheng and Zuofeng Shang. Joint Asymptotics for Semi-nonparametric Regression Models with Partially Linear Structure. *The Annals of Statistics*, 43:1351–1390, 2015.

Noel Cressie. Kriging Nonstationary Data. *Journal of the American Statistical Association*, 81(395):625–634, 1986.

Noel Cressie and Gardar Johannesson. Fixed Rank Kriging for Very Large Spatial Data Sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 209–226, 2008.

Noel Cressie and Christopher K Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, 2015.

Wenquan Cui, Haoyang Cheng, and Jiajing Sun. An RKHS-based Approach to Double-Penalized Regression in High-dimensional Partially Linear Models. *Journal of Multivariate Analysis*, 168:201–210, 2018.

Mingwang Dong, Linfu Huang, Xueqin Wu, and Qingguang Zeng. Application of Least-Squares Method to Time Series Analysis for 4dpm Matrix. *IOP Conference Series: Earth and Environmental Science*, 455(1):012200, feb 2020. doi: 10.1088/1755-1315/455/1/012200. URL https://dx.doi.org/10.1088/1755-1315/455/1/012200.

BK Fosdick and PD Hoff. Separable Factor Analysis with Applications to Mortality Data. *The Annals of Applied Statistics*, 8(1):120–147, 2014.

Chong Gu. *Smoothing Spline ANOVA models, 2nd edition*. Springer, New York, 2013.

Sharmistha Guha and Rajarshi Guhaniyogi. Bayesian Generalized Sparse Symmetric Tensor-on-Vector Regression. *Technometrics*, 63(2):160–170, 2021.

Rajarshi Guhaniyogi, Shaan Qamar, and David B Dunson. Bayesian Tensor Regression. *The Journal of Machine Learning Research*, 18(1):2733–2763, 2017.

Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

James D Hamilton. *Time Series Analysis*. Princeton University Press, 2020.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition*. Springer, New York, 2009.

Manuel Hernández-Pajares, JM Juan, J Sanz, R Orus, A Garcia-Rigo, J Feltens, A Komjathy, SC Schaer, and A Krankowski. The IGS VTEC Maps: a Reliable Source of Ionospheric Information since 1998. *Journal of Geodesy*, 83:263–275, 2009.

Peter D Hoff. Separable Covariance Arrays via the Tucker Product, with Applications to Multivariate Relational Data. *Bayesian Analysis*, 6(2):179–196, 2011.

Nan-Jung Hsu, Hsin-Cheng Huang, and Ruey S Tsay. Matrix Autoregressive Spatio-Temporal Models. *Journal of Computational and Graphical Statistics*, 30(4):1143–1155, 2021.

Jian Kang, Brian J Reich, and Ana-Maria Staicu. Scalar-on-Image Regression via the Soft-Thresholded Gaussian Process. *Biometrika*, 105(1):165–184, 2018.

Rodney A Kennedy, Parastoo Sadeghi, Zubair Khalid, and Jason D McEwen. Classification and Construction of Closed-form Kernels for Signal Representation on the 2-sphere. In *Wavelets and Sparsity XV*, volume 8858, pages 169–183. SPIE, 2013.

Tamara G Kolda and Brett W Bader. Tensor Decompositions and Applications. *SIAM review*, 51(3):455–500, 2009.

Vladimir Koltchinskii and Evarist Giné. Random Matrix Approximation of Spectra of Integral Operators. *Bernoulli*, pages 113–167, 2000.

Lexin Li and Xin Zhang. Parsimonious Tensor Response Regression. *Journal of the American Statistical Association*, 112(519):1131–1146, 2017.

Xiaoshan Li, Da Xu, Hua Zhou, and Lexin Li. Tucker Tensor Regression and Neuroimaging Analysis. *Statistics in Biosciences*, 10(3):520–545, 2018.

Zebang Li and Han Xiao. Multi-linear Tensor Autoregressive Models. *arXiv preprint arXiv:2110.00928*, 2021.

Yipeng Liu, Jiani Liu, and Ce Zhu. Low-rank Tensor Train Coefficient Array Estimation for Tensor-on-Tensor Regression. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5402–5411, 2020.

Eric F Lock. Tensor-on-Tensor Regression. *Journal of Computational and Graphical Statistics*, 27(3):638–647, 2018.

Yuetian Luo and Anru R Zhang. Tensor-on-Tensor Regression: Riemannian Optimization, Over-Parameterization, Statistical-Computational gap, and their Interplay. *arXiv preprint arXiv:2206.08756*, 2022.

Georgia Papadogeorgou, Zhengwu Zhang, and David B Dunson. Soft Tensor Regression. *J. Mach. Learn. Res.*, 22:219–1, 2021.

Natalia E. Papitashvili and Joseph H. King. Omni 5-min Data [Data set]. NASA Space Physics Data Facility, 2020. https://doi.org/10.48322/gbpg-5r77.

Natasha Papitashvili, Dieter Bilitza, and Joseph King. OMNI: a Description of Near-Earth Solar Wind Environment. *40th COSPAR scientific assembly*, 40:C0–1, 2014.

Guillaume Rabusseau and Hachem Kadri. Low-rank Regression with Tensor Responses. *Advances in Neural Information Processing Systems*, 29, 2016.

Mark Rudelson and Roman Vershynin. Hanson-Wright Inequality and Sub-Gaussian Concentration. *Electronic Communications in Probability*, pages 1–9, 2013.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A Generalized Representer Theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, pages 461–464, 1978.

Zuofeng Shang and Guang Cheng. Local and Global Asymptotic Inference in Smoothing Spline Models. *The Annals of Statistics*, 41:2608–2638, 2013.

Zuofeng Shang and Guang Cheng. Nonparametric Inference in Generalized Functional Linear Models. *The Annals of Statistics*, 43:1742–1773, 2015.

Bo Shen, Weijun Xie, and Zhenyu Kong. Smooth Robust Tensor Completion for Background/Foreground Separation with Missing Pixels: Novel Algorithm with Convergence Guarantee. *The Journal of Machine Learning Research*, 23(1):9757–9796, 2022.

James H Stock and Mark W Watson. Vector Autoregressions. *Journal of Economic perspectives*, 15(4):101–115, 2001.

Hu Sun, Zhijun Hua, Jiaen Ren, Shasha Zou, Yuekai Sun, and Yang Chen. Matrix Completion Methods for the Total Electron Content Video Reconstruction. *The Annals of Applied Statistics*, 16(3):1333–1358, 2022.

Hu Sun, Ward Manchester, Meng Jin, Yang Liu, and Yang Chen. Tensor Gaussian Process with Contraction for Multi-Channel Imaging Analysis. In *International Conference on Machine Learning*, pages 32913–32935. PMLR, 2023.

Will Wei Sun and Lexin Li. STORE: Sparse Tensor Response Regression and Neuroimaging Analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944, 2017.

KF Tapping. The 10.7 cm Solar Radio Flux (F10. 7). *Space weather*, 11(7):394–406, 2013.

ShengLi Tzeng and Hsin-Cheng Huang. Resolution Adaptive Fixed Rank Kriging. *Technometrics*, 60(2):198–208, 2018.

JH van Zanten and Aad W van der Vaart. Reproducing Kernel Hilbert Spaces of Gaussian Priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pages 200–222. Institute of Mathematical Statistics, 2008.

Di Wang, Yao Zheng, and Guodong Li. High-Dimensional Low-rank Tensor Autoregressive Time Series Modeling. *Journal of Econometrics*, 238(1):105544, 2024.

Dong Wang, Xialu Liu, and Rong Chen. Factor Models for Matrix-valued High-dimensional Time Series. *Journal of econometrics*, 208(1):231–248, 2019.

Xiao Wang, Hongtu Zhu, and Alzheimer's Disease Neuroimaging Initiative. Generalized Scalar-on-Image Regression Models via Total Variation. *Journal of the American Statistical Association*, 112(519):1156–1168, 2017.

Zihan Wang, Shasha Zou, Lei Liu, Jiaen Ren, and Ercha Aa. Hemispheric Asymmetries in the Mid-latitude Ionosphere During the September 7–8, 2017 Storm: Multi-instrument Observations. *Journal of Geophysical Research: Space Physics*, 126:e2020JA028829, 4 2021. ISSN 2169-9402. doi: 10.1029/2020JA028829.

Christopher K Williams and Carl Edward Rasmussen. *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA, 2006.

Yun Yang, Zuofeng Shang, and Guang Cheng. Non-asymptotic Analysis for Nonparametric Testing. In *33rd Annual Conference on Learning Theory*, pages 1–47. ACM, 2020.

Waqar Younas, Majid Khan, C. Amory-Mazaudier, Paul O. Amaechi, and R. Fleury. Middle and Low Latitudes Hemispheric Asymmetries in $\Sigma O/N2$ and TEC during Intense Magnetic Storms of Solar Cycle 24. *Advances in Space Research*, 69:220–235, 1 2022.

Ming Yuan and T Tony Cai. A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor Regression with Applications in Neuroimaging Data Analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

SUPPLEMENTARY MATERIAL

This supplemental material is organized as follows. In Section A, we prove Proposition 1 on the equivalence of the estimation problem of MARAC to a kernel ridge regression problem. In Section B, we prove Theorem 6 on the joint stationarity condition of the matrix and auxiliary vector time series. Then in Section C, we provide proofs of the theoretical results under fixed spatial dimensionality, including Proposition 8 and Theorem 9. In Section D, we present proofs of the theoretical results under high spatial dimensionality, namely Theorem 12. All essential lemmas used throughout the proofs are presented and proved in Section E. Finally, we include additional details and results of the simulation experiments in Section F.

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.

A 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 $h_{q,d}(\cdot) \notin \text{span}(\{k(\cdot, s)|s \in \mathbb{S}\})$, we have:

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

and the equality holds only if $h_{q,d}(\cdot) = 0$. Consequently, the global minimizer for the constrained optimization problem (7) 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. ■

B Proof of Theorem 1

Proof Under Assumption 4 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 follows a VAR($\max(P, Q, \widetilde{Q})$) process, namely,

$$\begin{bmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{bmatrix} = \sum_{l=1}^{\max(P,Q,\widetilde{Q})} \begin{bmatrix} (\mathbf{B}_l \otimes \mathbf{A}_l) \odot \mathbf{1}_{\{l \le P\}} & \mathbf{G}_l^\top \odot \mathbf{1}_{\{l \le Q\}} \\ \mathbf{O}_{D \times S} & \mathbf{C}_l \odot \mathbf{1}_{\{l \le \widetilde{Q}\}} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{t-l} \\ \mathbf{z}_{t-l} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_t \\ \boldsymbol{\nu}_t \end{bmatrix}. \tag{29}$$

Let $L = \max(P, Q, \widetilde{Q})$ and $\mathbf{y}_t = [\mathbf{x}_t^\top, \mathbf{z}_t^\top]$. Denote the transition matrix in (29) 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 (29) as a VAR(1) process as:

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-L+1} \end{bmatrix} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{J}_2 & \cdots & \mathbf{J}_{L-1} & \mathbf{J}_L \\ \mathbf{I}_{S+D} & \mathbf{O}_{S+D} & \cdots & \cdots & \mathbf{O}_{S+D} \\ \mathbf{O}_{S+D} & \mathbf{I}_{S+D} & \mathbf{O}_{S+D} & \cdots & \mathbf{O}_{S+D} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{O}_{S+D} & \mathbf{O}_{S+D} & \cdots & \mathbf{I}_{S+D} & \mathbf{O}_{S+D} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-L} \end{bmatrix} + \begin{bmatrix} \mathbf{u}_t \\ \mathbf{0}_{S+D} \\ \vdots \\ \mathbf{0}_{S+D} \end{bmatrix}, \tag{30}$$

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 (30). The determinant $\det(\lambda \mathbf{I} - \mathbf{J})$ can be simplified by column

operations as:

$$\begin{aligned}
&\det\left(\lambda\mathbf{I}-\mathbf{J}\right)\\
&=\det\begin{bmatrix}\lambda^{L}\mathbf{I}_{S}-\sum_{l=1}^{L}\lambda^{L-l}\left(\mathbf{B}_{l}\otimes\mathbf{A}_{l}\right)\odot\mathbf{1}_{\{l\leq P\}} & -\sum_{l=1}^{L}\lambda^{L-l}\mathbf{G}_{l}^{\top}\odot\mathbf{1}_{\{l\leq Q\}}\\ \mathbf{O} & \lambda^{L}\mathbf{I}_{D}-\sum_{l=1}^{L}\lambda^{L-l}\mathbf{C}_{l}\odot\mathbf{1}_{\{l\leq\widetilde{Q}\}}\end{bmatrix}\\
&=\lambda^{2L}\det\left[\mathbf{\Phi}_{1}(\lambda)\right]\det\left[\mathbf{\Phi}_{2}(\lambda)\right],
\end{aligned}$$

where $\mathbf{\Phi}_{1}(\lambda)=\mathbf{I}_{S}-\sum_{p=1}^{P}\lambda^{-p}\left(\mathbf{B}_{p}\otimes\mathbf{A}_{p}\right)$ 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. ■

C Theory under Fixed Spatial Dimension

C.1 Proof of Proposition 8

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 (4), we can equivalently write it as:

$$\mathbf{x}_{t}=\mathbf{y}_{t}\boldsymbol{\theta}+\mathbf{e}_{t}, \tag{31}$$

where $\mathbf{y}_{t}=[\mathbf{x}_{t-1}^{\top}\otimes\mathbf{I}_{S};\mathbf{z}_{t-1}^{\top}\otimes\mathbf{K}]$ and $\boldsymbol{\theta}=[\mathbf{vec}\left(\mathbf{B}_{1}\otimes\mathbf{A}_{1}\right)^{\top},\boldsymbol{\gamma}_{1}^{\top}]^{\top}$. Using $\mathbf{\Omega}=\mathbf{\Sigma}^{-1}$ to denote the precision matrix for $\mathbf{e}_{t}$, we can rewrite the penalized likelihood in (11) for $(\boldsymbol{\theta},\mathbf{\Omega})$ as:

$$h(\boldsymbol{\theta},\mathbf{\Omega})=-\frac{1}{2}\log\left|\mathbf{\Omega}\right|+\frac{1}{2}\mathrm{tr}\left(\mathbf{\Omega}\mathbf{S}(\boldsymbol{\theta})\right)+\frac{\lambda}{2}\boldsymbol{\theta}^{\top}\widetilde{\mathbf{K}}\boldsymbol{\theta}, \tag{32}$$

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:

$$\begin{aligned}
\mathbb{F}_{\boldsymbol{\theta}}&=\{[\mathbf{vec}\left(\mathbf{B}_{1}\otimes\mathbf{A}_{1}\right)^{\top},\boldsymbol{\gamma}_{1}^{\top}]^{\top}|\|\mathbf{A}_{1}\|_{\mathrm{F}}=1,\mathrm{sign}(\mathrm{tr}\left(\mathbf{A}_{1}\right))=1\}\\
\mathbb{F}_{\mathbf{\Omega}}&=\{\mathbf{\Sigma}_{c}^{-1}\otimes\mathbf{\Sigma}_{r}^{-1}|\mathbf{\Sigma}_{r}\in\mathbb{R}^{M\times M},\mathbf{\Sigma}_{c}\in\mathbb{R}^{N\times N},\underline{\rho}(\mathbf{\Sigma}_{r}),\underline{\rho}(\mathbf{\Sigma}_{c})>0\}.
\end{aligned}$$

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

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

$$\mathrm{P}\left(\inf_{\|\bar{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^*\|_{\mathrm{F}} \geq c} \inf_{\bar{\boldsymbol{\theta}}} h(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\Omega}}) \leq h(\boldsymbol{\theta}^*, \boldsymbol{\Omega}^*)\right) \to 0, \text{ as } T \to \infty. \tag{33}$$

This is because if (33) 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 (33), 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:

$$\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) = \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t}{T} + \lambda \widetilde{\mathbf{K}}\right)^{-1} \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{x}_t}{T}\right), \tag{34}$$

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:

$$\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) = (\mathbf{I} - \lambda \widetilde{\mathbf{K}})\boldsymbol{\theta}^* + \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t}{T} + \lambda \widetilde{\mathbf{K}}\right)^{-1} \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{e}_t}{T}\right), \tag{35}$$

and the first term converges to $\boldsymbol{\theta}^*$ since $\lambda = o(1)$. In the second term of (35), we have:

$$\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t}{T} + \lambda \widetilde{\mathbf{K}} \xrightarrow{p} \begin{bmatrix} \boldsymbol{\Sigma}^*_{\mathbf{x},\mathbf{x}} \otimes \bar{\boldsymbol{\Omega}} & \boldsymbol{\Sigma}^*_{\mathbf{x},\mathbf{z}} \otimes \bar{\boldsymbol{\Omega}}\mathbf{K} \\ \boldsymbol{\Sigma}^*_{\mathbf{z},\mathbf{x}} \otimes \mathbf{K}\bar{\boldsymbol{\Omega}} & \boldsymbol{\Sigma}^*_{\mathbf{z},\mathbf{z}} \otimes \mathbf{K}\bar{\boldsymbol{\Omega}}\mathbf{K} \end{bmatrix}, \tag{36}$$

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 (36) 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 (36), we conclude that the second term in (35) 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 (33), it suffices to show that:

$$\mathrm{P}\left(\inf_{\|\bar{\mathbf{\Omega}}-\mathbf{\Omega}^*\|_{\mathrm{F}}\geq c}\ell(\bar{\mathbf{\Omega}})\leq\ell(\mathbf{\Omega}^*)\right)\to 0,\ \text{as } T\to\infty, \tag{37}$$

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

$$\begin{aligned}\ell(\bar{\mathbf{\Omega}})&=-\frac{1}{2}\log|\bar{\mathbf{\Omega}}|+\frac{1}{2}\mathrm{tr}\left(\bar{\mathbf{\Omega}}\frac{\sum_t\mathbf{x}_t\mathbf{e}_t^\top}{T}\right)-\frac{1}{2}\mathrm{tr}\left(\left(\frac{\sum_t\mathbf{x}_t^\top\bar{\mathbf{\Omega}}\mathbf{y}_t}{T}\right)\boldsymbol{\zeta}\right)\\&=\widetilde{\ell}(\bar{\mathbf{\Omega}})-\frac{1}{2}\mathrm{tr}\left(\left(\frac{\sum_t\mathbf{x}_t^\top\bar{\mathbf{\Omega}}\mathbf{y}_t}{T}\right)\boldsymbol{\zeta}\right),\end{aligned} \tag{38}$$

where we define the first two terms in (38) as $\widetilde{\ell}(\bar{\mathbf{\Omega}})$.

By the Cauchy-Schwartz inequality, we have:

$$\left|\frac{1}{2}\mathrm{tr}\left(\left(\frac{\sum_t\mathbf{x}_t^\top\bar{\mathbf{\Omega}}\mathbf{y}_t}{T}\right)\boldsymbol{\zeta}\right)\right|\leq\frac{1}{2}\left\|\frac{\sum_t\mathbf{x}_t^\top\bar{\mathbf{\Omega}}\mathbf{y}_t}{T}\right\|_{\mathrm{F}}\cdot\|\boldsymbol{\zeta}\|_{\mathrm{F}}. \tag{39}$$

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

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

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 (39) being $o_P(1)$.

Given this argument, proving (37) is now equivalent to proving:

$$\mathrm{P}\left(\inf_{\|\bar{\mathbf{\Omega}}-\mathbf{\Omega}^*\|_{\mathrm{F}}\geq c}\widetilde{\ell}(\bar{\mathbf{\Omega}})\leq\widetilde{\ell}(\mathbf{\Omega}^*)\right)\to 0,\ \text{as } T\to\infty. \tag{40}$$

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

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

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:

$$\widetilde{\ell}(\bar{\mathbf{\Omega}})=\widetilde{\ell}(\widetilde{\mathbf{\Omega}})+\frac{1}{4}\mathbf{vec}\left(\bar{\mathbf{\Omega}}-\widetilde{\mathbf{\Omega}}\right)^\top\left[\check{\mathbf{\Omega}}^{-1}\otimes\check{\mathbf{\Omega}}^{-1}\right]\mathbf{vec}\left(\bar{\mathbf{\Omega}}-\widetilde{\mathbf{\Omega}}\right), \tag{41}$$

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 (41), we end up having:

$$\mathrm{P}\left[\widetilde{\ell}(\bar{\mathbf{\Omega}}) \geq \widetilde{\ell}(\widetilde{\mathbf{\Omega}}) + \frac{1}{16} \cdot \left(\frac{\kappa}{3+\kappa}\right)^2\right] \to 1, \text{ as } T \to \infty, \tag{42}$$

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 (40) and thereby completes the proof. ■

C.2 Proof of Theorem 9

To prove Theorem 9, we first establish the consistency and the convergence rate of the estimators in Lemma 14 below.

Lemma 14 *Under the same assumption as Theorem 9, 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 14 to Section E.2. With this lemma, we are now ready to present the proof of Theorem 9.

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 (13). By plugging in the data-generating model for $\mathbf{X}_t$ according to MARAC(1, 1) model, we can transform (13) 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:

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

On the left-hand side of the equation above, we replace $\widehat{\mathbf{B}}_1, \widehat{\boldsymbol{\Sigma}}_r, \widehat{\boldsymbol{\Sigma}}_c$ with their true values $\mathbf{B}_1^*, \boldsymbol{\Sigma}_r^*, \boldsymbol{\Sigma}_c^*$, 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:

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

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 $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\mathbf{B}}_1$, we can replace $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\mathbf{B}}_1$ with their true values and incorporate the remainders into $o_P(\sqrt{T})$.

Similar transformations can be applied to (14) and (15), 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:

$$\left(\frac{1}{T} \sum_{t \in [T]} \mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t \right) \mathbf{vec}\left(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \right) = \frac{1}{T} \sum_{t \in [T]} \mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{vec}\left(\mathbf{E}_t \right) + o_P(T^{-1/2}), \tag{43}$$

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

In (43), we first establish that:

$$(1/T) \sum_{t \in [T]} \mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t \xrightarrow{p.} \mathrm{E}\left[\mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t \right]. \tag{44}$$

To prove 44, 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 20 and Corollary 21, where $\widetilde{\mathbf{x}}_t = [\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top$. See details of Lemma 20 and Corollary 21 in Section E.1. Then since each element of $\mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t$ is a linear combination of terms in $\widetilde{\mathbf{x}}_t \widetilde{\mathbf{x}}_t^\top$ (thus a continuous mapping), it is straightforward that (44) holds elementwise.

Given (44) and the fact that $\widehat{\mathbf{\Theta}}$ is $\sqrt{T}$-consistent, we can rewrite (43) as:

$$\mathrm{E}\left[\mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t \right] \mathbf{vec}\left(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \right) = \frac{1}{T} \sum_{t \in [T]} \mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{vec}\left(\mathbf{E}_t \right) + o_P(T^{-1/2}), \tag{45}$$

For the term on the right-hand side of (45), first notice that the sequence $\{\boldsymbol{\eta}_t\}_{t=1}^T$, where $\boldsymbol{\eta}_t = \mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{vec}\left(\mathbf{E}_t \right)$, 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:

$$\frac{1}{\sqrt{T}} \sum_{t \in [T]} \mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{vec}\left(\mathbf{E}_t \right) \xrightarrow{d.} \mathcal{N}(\mathbf{0}, \mathrm{E}\left[\mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t \right]). \tag{46}$$

Combining (45) and (46), we end up having:

$$\mathrm{E}\left[\mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t \right] \sqrt{T} \mathbf{vec}\left(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \right) \xrightarrow{d.} \mathcal{N}(\mathbf{0}, \mathrm{E}\left[\mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t \right]). \tag{47}$$

The asymptotic distribution of $\sqrt{T}\mathbf{vec}\left(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*\right)$ can thus be derived by multiplying both sides of (47) 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:

$$\sqrt{T}\boldsymbol{\zeta}^\top \mathbf{vec}\left(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*\right) \xrightarrow{p} 0. \tag{48}$$

Combining (47) and (48) 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:

$$\sqrt{T}\mathbf{vec}(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{H}^{-1}\mathbf{L}\mathbf{H}^{-1}). \tag{49}$$

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 (49) with multivariate delta method and we omit the details here. ■

D Theory under High Spatial Dimension

D.1 Proof of Theorem 12

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 $\boldsymbol{\phi}^* = \mathbf{vec}\,(\mathbf{B}^* \otimes \mathbf{A}^*)$, 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 $\mathbb{F}_{\boldsymbol{\phi}} = \{\mathbf{vec}\,(\mathbf{B} \otimes \mathbf{A}) \,|\, \|\mathbf{A}\|_{\mathrm{F}} = \mathrm{sign}(\mathrm{tr}\,(\mathbf{A})) = 1, \mathbf{A} \in \mathbb{R}^{M\times M}, \mathbf{B} \in \mathbb{R}^{N\times N}\}$. Using these new notations, the MARAC estimator is obtained by solving the following penalized least square problem:

$$\min_{\boldsymbol{\phi}\in\mathbb{F}_{\boldsymbol{\phi}}, \boldsymbol{\gamma}\in\mathbb{R}^{SD}} \mathfrak{L}_\lambda(\boldsymbol{\phi}, \boldsymbol{\gamma}) \coloneqq \left\{ \frac{1}{2T}\|\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\phi} - (\widetilde{\mathbf{z}} \otimes \mathbf{K})\,\boldsymbol{\gamma}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\boldsymbol{\gamma}^\top (\mathbf{I}_D \otimes \mathbf{K})\,\boldsymbol{\gamma} \right\}. \tag{50}$$

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:

$$\widehat{\boldsymbol{\gamma}}(\boldsymbol{\phi}) = T^{-1}\left[\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}} \otimes \mathbf{K} + \lambda \cdot \mathbf{I}_{SD}\right]^{-1} (\widetilde{\mathbf{z}}^\top \otimes \mathbf{I}_S)\left(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\phi}\right). \tag{51}$$

Plugging (51) into (50) yields the profile likelihood for $\boldsymbol{\phi}$:

$$\ell_\lambda(\boldsymbol{\phi}) = \mathfrak{L}_\lambda(\boldsymbol{\phi}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\phi})) = \frac{1}{2T}\left(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\phi}\right)^\top \mathbf{W}\left(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\phi}\right), \tag{52}$$

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

$$\mathbf{W} = \left\{ \mathbf{I} - \frac{(\widetilde{\mathbf{z}} \otimes \mathbf{K})\left[\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}} \otimes \mathbf{K} + \lambda \cdot \mathbf{I}_{SD}\right]^{-1} (\widetilde{\mathbf{z}}^\top \otimes \mathbf{I}_S)}{T} \right\} = \left(\mathbf{I} + \frac{\widetilde{\mathbf{z}}\widetilde{\mathbf{z}}^\top}{\lambda T} \otimes \mathbf{K}\right)^{-1}, \tag{53}$$

and the second equality in (53) 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 15 *Under the assumptions of Theorem 12, we have:*

$$\left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\right)^\top \left(\frac{\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}}{T}\right)\left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\right) \lesssim O_P(C_g\lambda) + O_P(SD/T), \tag{54}$$

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 16 *Under the assumptions of Theorem 12 and the requirement that* $S\log S/T \to 0$*, it holds that:*

$$\underline{\rho}\left(\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}/T\right) \geq \frac{c_{0,S}}{2} > 0, \tag{55}$$

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

The proof of Proposition 15 and Lemma 16 are relegated to Section D.2 and E.3, respectively. Combining Proposition 15 and Lemma 16, we can derive the error bound of $\widehat{\boldsymbol{\phi}}$ as:

$$\frac{1}{S}\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\|_{\mathrm{F}} \lesssim O_P(\sqrt{\frac{C_g \gamma_S}{c_{0,S} S}}) + O_P(\sqrt{\frac{D}{c_{0,S} T S}}). \tag{56}$$

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 17 *Given the definition of* $\mathbf{W}$ *in* (53) *and under the assumptions of Theorem 12, 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 18 *Given the definition of* $\mathbf{W}$ *in* (53) *and under the assumptions of Theorem 12, we have that:*

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

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

We leave the proof of Lemma 17 and Lemma 18 to Section E.4 and E.5. By Lemma 18, we have:

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

And by Lemma 17, we have $J_1 \leq O_P(S^{1/4}\gamma_S^{-1/4r_0})$.

For J_2, we have the following bound:

$$J_2 = \|(\mathbf{I} - \mathbf{W})\mathbf{W}^{-1/2}\mathbf{W}^{1/2}\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*)\|_{\mathrm{F}} \tag{57}$$

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

$$\leq \|\mathbf{W}^{-1/2}\|_s \cdot \|\mathbf{W}^{1/2}\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*)\|_{\mathrm{F}}. \tag{58}$$

To bound $\|\mathbf{W}^{-1/2}\|_s$, we can take advantage of the simpler form of $\mathbf{W}$ using the Woodbury matrix identity in (53) 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 20, which we state later in Section E.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 15, we can bound J_2 via its upper bound (58) as:

$$J_2 \leq O_P\left(\sqrt{C_g \lambda T}\right) + O_P\left(\sqrt{C_g T}\right) + O_P(\sqrt{S}) + O_P\left(\sqrt{\gamma_S^{-1}}\right). \tag{59}$$

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:

$$\|\mathbf{W}^{1/2}(\widetilde{\mathbf{z}} \otimes \mathbf{K})\boldsymbol{\gamma}^*\|_{\mathrm{F}}^2 = (\lambda T)[(\mathbf{I}_D \otimes \mathbf{K})\boldsymbol{\gamma}^*]^\top \left\{\mathbf{I}_{SD} - \left(\lambda^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}} \otimes \mathbf{K} + \mathbf{I}_{SD}\right)^{-1}\right\}\boldsymbol{\gamma}^*$$

$$= (\lambda T)\left(\sum_{d=1}^{D} \|g_d^*\|_{\mathbb{H}_k}^2\right)$$

$$- (\lambda^2 T)\left(\boldsymbol{\gamma}^*\right)^\top \left[(\mathbf{I}_D \otimes \mathbf{K})\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}} \otimes \mathbf{K} + \lambda \mathbf{I}_{SD}\right)^{-1}\right]\boldsymbol{\gamma}^*$$

$$\leq C_g \lambda T, \tag{60}$$

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 (60) 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 19 *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 E.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 19, we have $\mathbf{N}^{-1} - \mathbf{M}^{-1}$ being positive semi-definite and thus (60) holds.

Using the result in (60), 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:

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

where we drop the term $O_P(T^{-1/2})$ as it is a higher order term of $O_P(S^{-1/2})$ under the condition that $S \log S / T \to 0$. ■

D.2 Proof of Proposition 15

Proof The MARAC estimator $\widehat{\boldsymbol{\phi}}$ is the minimizer of $\ell_\lambda(\boldsymbol{\phi})$, defined in (52), for all $\boldsymbol{\phi} \in \mathbb{F}_\phi$ and thus $\ell_\lambda(\widehat{\boldsymbol{\phi}}) \leq \ell_\lambda(\boldsymbol{\phi}^*)$. Equivalently, this means that:

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

Let $\boldsymbol{\delta} = \mathbf{W}^{1/2} \widetilde{\mathbf{X}} (\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*) / \sqrt{T}$ and $\boldsymbol{\omega} = \mathbf{W}^{1/2} \left[(\widetilde{\mathbf{z}} \otimes \mathbf{K}) \boldsymbol{\gamma}^* + \boldsymbol{\mathcal{E}}\right] / \sqrt{T}$, 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 $\|\boldsymbol{\omega}\|_{\mathrm{F}}^2$:

$$\begin{aligned}\|\boldsymbol{\delta}\|_{\mathrm{F}}^2 \leq 4\|\boldsymbol{\omega}\|_{\mathrm{F}}^2 &= \frac{4}{T}\left[\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^* + \boldsymbol{\mathcal{E}}\right]^\top \mathbf{W}\left[\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^* + \boldsymbol{\mathcal{E}}\right] \\ &\leq \frac{8}{T}\left[\underbrace{\|\mathbf{W}^{1/2}\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^*\|_{\mathrm{F}}^2}_{I_1} + \underbrace{\|\mathbf{W}^{1/2}\boldsymbol{\mathcal{E}}\|_{\mathrm{F}}^2}_{I_2}\right], \end{aligned} \tag{61}$$

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

For I_1, it can be bounded by (60) and thus $I_1 \leq C_g\lambda T$. To bound I_2, we utilize Lemma 18 and bound I_2 as $I_2 \asymp \mathrm{tr}\left(\mathbf{W}\right) \leq SD$. Combining the bounds for I_1 and I_2, we have:

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

which completes the proof. ■

E Technical Lemmas & Proofs

In this section, we first introduce Lemma 20 on the consistency of the covariance matrix estimator for any stationary vector autoregressive process and then Corollary 21 on the consistency of the covariance estimator of our MARAC model, given the joint stationarity condition. Then we provide proof for Lemma 14 used in Section C.2 when proving Theorem 9 on the asymptotic normality under fixed spatial dimension. Then we provide proofs for Lemma 16, 17, 18 and 19 used in Section D when proving the error bounds with high spatial dimensionality.

E.1 Statement of Lemma 20

In Lemma 20, we restate the result of Proposition 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 20 *Let $\mathbf{x}_t \in \mathbb{R}^N$ be a zero-meaned stationary VAR(p) process: $\mathbf{x}_t = \sum_{l=1}^p \boldsymbol{\Phi}_p\mathbf{x}_{t-p} + \boldsymbol{\xi}_t$, where $\boldsymbol{\xi}_t$ have independent sub-Gaussian entries. Let $\widehat{\boldsymbol{\Sigma}} = (1/T)\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t^\top$ and $\boldsymbol{\Sigma} =$*

$\mathrm{E}[\widehat{\mathbf{\Sigma}}]$, *then we have:*

$$\mathrm{E}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_s \leq C\left(\sqrt{\frac{N\log N}{T}} + \frac{N\log N}{T}\right)\|\mathbf{\Sigma}\|_s, \tag{62}$$

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 20, we have the following results:

Corollary 21 *Assume that $\{\mathbf{z}_t\}_{t=1}^T$ is generated by a stationary VAR($\widetilde{Q}$) process: $\mathbf{z}_t = \sum_{\tilde{q}=1}^{\widetilde{Q}} \mathbf{C}_{\tilde{q}}\mathbf{z}_{t-\tilde{q}} + \boldsymbol{\nu}_t$, 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:*

$$\mathrm{P}\left(\left\|\widehat{\mathbf{\Sigma}}_{\mathbf{z}} - \mathbf{\Sigma}_{\mathbf{z}}^*\right\|_s \geq \epsilon\right) \leq C\epsilon^{-1}\left(\sqrt{\frac{D}{T}} + \frac{D}{T}\right), \tag{63}$$

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 6. 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:

$$\mathrm{E}\left\|\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0^*\right\|_s \leq C\left(\sqrt{\frac{S\log S}{T}} + \frac{S\log S}{T}\right)\|\mathbf{\Sigma}_0^*\|_s, \tag{64}$$

where C is an absolute constant.

Proof The proof of (63) is straightforward from Lemma 20 together with Markov inequality. The proof of (64) also follows from Lemma 20 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 (29)) 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.

E.2 Proof of Lemma 14

Proof Without loss of generality, we fix P, Q as 1 and use the same notation as (31) in Section C.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 $h(\boldsymbol{\theta}, \boldsymbol{\Omega})$ is specified by (32) and given any $\bar{\boldsymbol{\Omega}}$, we have $\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) = \arg\min_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \bar{\boldsymbol{\Omega}})$ as specified by (34). Given the decomposition of $\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})$ in (35), 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 21. 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 7, 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:

$$\begin{aligned} h(\boldsymbol{\theta}, \bar{\boldsymbol{\Omega}}) &= h(\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}), \bar{\boldsymbol{\Omega}}) + \frac{1}{2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}))^\top \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t}{T} + \lambda \widetilde{\mathbf{K}} \right) (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})) \\ &\geq h(\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}), \bar{\boldsymbol{\Omega}}) + \frac{1}{2}\underline{\rho}(\mathbf{L}_T)\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\|_{\text{F}}^2, \end{aligned} \tag{65}$$

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 $C_1 > 0$:

$$\begin{aligned} &\inf_{\bar{\boldsymbol{\Omega}} \in \mathbb{F}_{\boldsymbol{\Omega}}: \|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\text{F}} \leq C_1} h(\boldsymbol{\theta}, \bar{\boldsymbol{\Omega}}) \\ &\geq \inf_{\bar{\boldsymbol{\Omega}} \in \mathbb{F}_{\boldsymbol{\Omega}}: \|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\text{F}} \leq C_1} \left\{ h(\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}), \bar{\boldsymbol{\Omega}}) + \frac{1}{4}\underline{\rho}(\mathbf{L}) \cdot \|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\|_{\text{F}}^2 \right\}, \end{aligned} \tag{66}$$

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:

$$\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\|_{\text{F}} \geq \frac{c_T}{\sqrt{T}} - \|\boldsymbol{\theta}^* - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\|_{\text{F}}, \tag{67}$$

thus $\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\|_{\text{F}} \gtrsim O_P(c_T'/\sqrt{T})$ for some sequence $c_T' \to \infty$ since $\|\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) - \boldsymbol{\theta}^*\|_{\text{F}} = O_P(T^{-1/2})$. By the Taylor expansion in (65), 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 $\|\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) - \boldsymbol{\theta}^*\|_{\text{F}} = O_P(T^{-1/2})$. Combining this result together with the order of $\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}})\|_{\text{F}}$, we have the following hold according to (66):

$$\text{P}\left(\inf_{\sqrt{T}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\text{F}} \geq c_T} \inf_{\bar{\boldsymbol{\Omega}} \in \mathbb{F}_{\boldsymbol{\Omega}}: \|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\text{F}} \leq C_1} h(\boldsymbol{\theta}, \bar{\boldsymbol{\Omega}}) > \inf_{\bar{\boldsymbol{\Omega}} \in \mathbb{F}_{\boldsymbol{\Omega}}: \|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\text{F}} \leq C_1} h(\boldsymbol{\theta}^*, \bar{\boldsymbol{\Omega}}) \right) \to 1. \tag{68}$$

The result in (68) 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 $\widehat{\mathbf{B}}_p$.

■

E.3 Proof of Lemma 16

Proof Based on the definition of $\mathbf{W}$ in equation (53), we have

$$\begin{aligned}
\frac{\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}}{T} &= \widehat{\boldsymbol{\Sigma}}_{\mathbf{x},\mathbf{x}} \otimes \mathbf{I}_S - \left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \otimes \mathbf{K}\right)\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{z}} \otimes \mathbf{K} + \lambda \mathbf{I}_{SD}\right)^{-1}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}} \otimes \mathbf{I}_S\right) \\
&= \left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},\mathbf{x}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1} \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}}\right) \otimes \mathbf{I}_S \\
&+ \left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}} \otimes \mathbf{I}_S\right)^\top \left[\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{z}}^2 \otimes \lambda^{-1}\mathbf{K} + \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{z}} \otimes \mathbf{I}_S\right]^{-1}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}} \otimes \mathbf{I}_S\right), \qquad (69)
\end{aligned}$$

where the second term in (69) is positive semi-definite since both $\underline{\rho}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{z}})$ 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 $\underline{\rho}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},\mathbf{x}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1} \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}})$. For simplicity, we will use $\mathbf{A}, \mathbf{B}, \mathbf{C}$ to denote $\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{x}}^*, \boldsymbol{\Sigma}_{\mathbf{z},\mathbf{x}}^*, (\boldsymbol{\Sigma}_{\mathbf{z},\mathbf{z}}^*)^{-1}$, and $\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}$ to denote $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},\mathbf{x}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1}$, respectively. We will use $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}^*$ to denote the estimated and true covariance matrix of $[\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top$. It is evident that $\|\mathbf{A}\|_s \leq \|\boldsymbol{\Sigma}^*\|_s$ and $\|\mathbf{B}\|_s \leq \|\boldsymbol{\Sigma}^*\|_s$, since both $\mathbf{A}$ and $\mathbf{B}$ are blocks of $\boldsymbol{\Sigma}^*$ and can thus be represented as $\mathbf{E}_1^\top \boldsymbol{\Sigma}^* \mathbf{E}_2$ 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 $S \log S/T \to 0$, $\underline{\rho}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x},\mathbf{x}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1} \widehat{\boldsymbol{\Sigma}}_{\mathbf{z},\mathbf{x}}) \xrightarrow{p} \underline{\rho}(\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{x}}^* - \left(\boldsymbol{\Sigma}_{\mathbf{z},\mathbf{x}}^*\right)^\top \left(\boldsymbol{\Sigma}_{\mathbf{z},\mathbf{z}}^*\right)^{-1} \boldsymbol{\Sigma}_{\mathbf{z},\mathbf{x}}^*)$. For brevity, we omit the subscript s for the spectral norm notation and simply use $\|\cdot\|$ in this proof.

To start with, we have:

$$\begin{aligned}
&\|\widehat{\mathbf{A}} - \widehat{\mathbf{B}}^\top \widehat{\mathbf{C}} \widehat{\mathbf{B}} - (\mathbf{A} - \mathbf{B}^\top \mathbf{C} \mathbf{B})\| \\
&\leq \|\widehat{\mathbf{A}} - \mathbf{A}\| + \|\widehat{\mathbf{B}}^\top \widehat{\mathbf{C}} \widehat{\mathbf{B}} - \mathbf{B}^\top \widehat{\mathbf{C}} \mathbf{B}\| + \|\mathbf{B}^\top \widehat{\mathbf{C}} \mathbf{B} - \mathbf{B}^\top \mathbf{C} \mathbf{B}\| \\
&\leq \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| + \|(\widehat{\mathbf{B}} - \mathbf{B})^\top \widehat{\mathbf{C}} \widehat{\mathbf{B}}\| + \|\mathbf{B}^\top \mathbf{C} (\widehat{\mathbf{B}} - \mathbf{B})\| + \|\mathbf{B}^\top (\widehat{\mathbf{C}} - \mathbf{C}) \widehat{\mathbf{B}}\| \\
&\leq \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| + \|\widehat{\mathbf{B}} - \mathbf{B}\| \cdot \left(\|\widehat{\mathbf{C}}\| \cdot \|\widehat{\mathbf{B}}\| + \|\mathbf{C}\| \cdot \|\mathbf{B}\| \right) \\
&+ \|\mathbf{B}\| \cdot \|\widehat{\mathbf{B}}\| \cdot \|\widehat{\mathbf{C}} - \mathbf{C}\|. \qquad (70)
\end{aligned}$$

Based on Corollary 21, 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 (70), with probability approaching 1, we have:

$$\begin{aligned}
\|\widehat{\mathbf{A}} - \widehat{\mathbf{B}}^\top \widehat{\mathbf{C}} \widehat{\mathbf{B}} - (\mathbf{A} - \mathbf{B}^\top \mathbf{C} \mathbf{B})\| \leq (1 + 7\|\mathbf{C}\| \cdot \|\mathbf{\Sigma}^*\|) \cdot \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\| \\
+ 3\|\mathbf{\Sigma}^*\|^2 \cdot \|\widehat{\mathbf{C}} - \mathbf{C}\|. \qquad (71)
\end{aligned}$$

The upper bound in (71) 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:

$$\underline{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{x},\mathbf{x}} - \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1} \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}) \geq \frac{1}{2} \underline{\rho} \left(\mathbf{\Sigma}_{\mathbf{x},\mathbf{x}}^* - \left(\mathbf{\Sigma}_{\mathbf{z},\mathbf{x}}^* \right)^\top \left(\mathbf{\Sigma}_{\mathbf{z},\mathbf{z}}^* \right)^{-1} \mathbf{\Sigma}_{\mathbf{z},\mathbf{x}}^* \right) = \frac{c_{0,S}}{2}. \qquad (72)$$

This completes the proof.

■

E.4 Proof of Lemma 17

Proof By the definition of $\mathbf{W}$ in (53), we have:

$$\begin{aligned}
\operatorname{tr}(\mathbf{I} - \mathbf{W}) &= \operatorname{tr}\left[\left(\widehat{\mathbf{\Sigma}}_{\mathbf{z}} \otimes \mathbf{K} + \lambda \mathbf{I}_{SD} \right)^{-1} \left(\widehat{\mathbf{\Sigma}}_{\mathbf{z}} \otimes \mathbf{K} \right) \right] \\
&= \sum_{s=1}^{S} \sum_{d=1}^{D} \frac{\rho_d(\widehat{\mathbf{\Sigma}}_{\mathbf{z}}) \rho_s(\mathbf{K})}{\lambda + \rho_d(\widehat{\mathbf{\Sigma}}_{\mathbf{z}}) \rho_s(\mathbf{K})} \leq D \cdot \sum_{s=1}^{S} \frac{1}{1 + \lambda \bar{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{z}})^{-1} \rho_s(\mathbf{K})^{-1}}. \qquad (73)
\end{aligned}$$

Using Lemma 20, 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 11 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 (73) can be bounded as:

$$D \cdot \sum_{s=1}^{S} \frac{1}{1 + \lambda \bar{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{z}})^{-1} \rho_s(\mathbf{K})^{-1}} \leq D \cdot \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{1}{1 + c_{\mathbf{z}} \lambda \rho_i(\mathbf{K}_1)^{-1} \rho_j(\mathbf{K}_2)^{-1}}, \tag{74}$$

where $c_{\mathbf{z}} = 1/2\bar{\rho}(\mathbf{\Sigma}^*_{\mathbf{z}})$. As $M, N \to \infty$, based on Assumption 10, 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:

$$\begin{aligned} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{1}{1 + c_2 \lambda (ij)^{r_0}/S} &\leq \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{1}{1 + c_{\mathbf{z}} \lambda \rho_i(\mathbf{K}_1)^{-1} \rho_j(\mathbf{K}_2)^{-1}} \\ &\leq \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{1}{1 + c_1 \lambda (ij)^{r_0}/S}, \end{aligned} \tag{75}$$

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 (75), we first notice that for any constant $c > 0$, one has:

$$\sum_{i=1}^{M \wedge N} \frac{1}{1 + c\lambda i^{2r_0}/S} \leq \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{1}{1 + c\lambda (ij)^{r_0}/S} \leq 2(M \vee N) \sum_{i=1}^{M \vee N} \frac{1}{1 + c\lambda i^{2r_0}/S}. \tag{76}$$

To approximate the sum in (76), 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 12, 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:

$$\lim_{S \to \infty} \sum_{i=1}^{M \vee N} \frac{1}{1 + c\lambda i^{2r_0}/S} = \int_0^C \frac{1}{1 + x^{2r_0}} \mathrm{d}x \cdot \lim_{S \to \infty} (S/c\lambda)^{1/2r_0} = O(\gamma_S^{-1/2r_0}). \tag{77}$$

Combining (73), (74), (75) and (77), 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 (75). 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 (77) 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:

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

■

E.5 Proof of Lemma 18

Proof Given any fixed $\mathbf{W}$ and $t > 0$, let $K = \sqrt{8/3}\sigma$ and $c > 0$ be some constant, then we have:

$$\mathrm{P}\left[\left|\boldsymbol{\mathcal{E}}^\top \mathbf{W}\boldsymbol{\mathcal{E}} - \sigma^2 \mathrm{tr}\,(\mathbf{W})\right| > t \middle| \mathbf{W}\right] \leq 2\exp\left[-c\min\left(\frac{t^2}{K^4\|\mathbf{W}\|_{\mathrm{F}}^2}, \frac{t}{K^2\|\mathbf{W}\|_s}\right)\right], \tag{78}$$

which holds by the Hanson-Wright inequality (Rudelson and Vershynin, 2013). Letting $t = \sigma^2\mathrm{tr}\,(\mathbf{W})/2$, $c_1 = 9c/256$, and using the fact that $\|\mathbf{W}\|_s \leq 1$, we have:

$$2\exp\left[-c\min\left(\frac{t^2}{K^4\|\mathbf{W}\|_{\mathrm{F}}^2}, \frac{t}{K^2\|\mathbf{W}\|_s}\right)\right] \leq 2\exp\left[-c_1\mathrm{tr}\,(\mathbf{W})\right]. \tag{79}$$

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

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

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

$$\mathrm{P}\left[\mathrm{tr}\left(\mathbf{W}\right) \geq \frac{SD\lambda}{\lambda + \bar{c}}\right] \to 1, \quad \text{as } T \to \infty, \tag{80}$$

where $\bar{c} = 2\bar{\rho}(\mathbf{\Sigma}_{\mathbf{z}}^*)\bar{\rho}(\mathbf{K})$. 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 $\{\left|\boldsymbol{\mathcal{E}}^\top \mathbf{W} \boldsymbol{\mathcal{E}} - \sigma^2 \mathrm{tr}\left(\mathbf{W}\right)\right| > \sigma^2 \mathrm{tr}\left(\mathbf{W}\right)/2\}$ as follows:

$$\begin{aligned}
&\mathrm{P}\left[\left|\boldsymbol{\mathcal{E}}^\top \mathbf{W} \boldsymbol{\mathcal{E}} - \sigma^2 \mathrm{tr}\left(\mathbf{W}\right)\right| > \sigma^2 \mathrm{tr}\left(\mathbf{W}\right)/2\right] \\
&\leq \mathrm{E}\left[2\exp\left[-c\mathrm{tr}\left(\mathbf{W}\right)\right]\right] \\
&\leq 2\left\{1 \cdot \mathrm{P}\left(\mathrm{tr}\left(W\right) < \frac{SD\lambda}{\lambda + \bar{c}}\right) + \exp\left[-c\frac{SD\lambda}{\lambda + \bar{c}}\right] \cdot \mathrm{P}\left(\mathrm{tr}\left(W\right) \geq \frac{SD\lambda}{\lambda + \bar{c}}\right)\right\} \to 0.
\end{aligned} \tag{81}$$

This indicates that $\boldsymbol{\mathcal{E}}^\top \mathbf{W} \boldsymbol{\mathcal{E}} \asymp \mathrm{tr}\left(\mathbf{W}\right)$.

The proof of $\boldsymbol{\mathcal{E}}^\top \left(\mathbf{I} - \mathbf{W}\right)^2 \boldsymbol{\mathcal{E}} \asymp \mathrm{tr}\left(\left(\mathbf{I} - \mathbf{W}\right)^2\right)$ is similar to the proof above. In the first step, similar to (78) and (79), we have the following tail probability bound:

$$\begin{aligned}
&\mathrm{P}\left[\left|\boldsymbol{\mathcal{E}}^\top \left(\mathbf{I} - \mathbf{W}\right)^2 \boldsymbol{\mathcal{E}} - \sigma^2 \mathrm{tr}\left(\left(\mathbf{I} - \mathbf{W}\right)^2\right)\right| > \frac{\sigma^2 \mathrm{tr}\left(\left(\mathbf{I} - \mathbf{W}\right)^2\right)}{2} \middle| \mathbf{W}\right] \\
&\leq 2\exp\left\{-c\mathrm{tr}\left(\left(\mathbf{I} - \mathbf{W}\right)^2\right)\right\}.
\end{aligned} \tag{82}$$

We can actually establish the unboundedness of $\mathrm{tr}\left(\left(\mathbf{I} - \mathbf{W}\right)^2\right)$ by following the same idea as the proof for Lemma 17, where we have:

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

with probability approaching 1 and c is some constant. Therefore, we have $\mathrm{tr}\left(\left(\mathbf{I} - \mathbf{W}\right)^2\right) \gtrsim O_P(\gamma_S^{-1/2r_0})$. The rest of the proof follows the idea of (81) and we omit the details here. ■

E.6 Proof of Lemma 19

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. ■

F Additional Details on Simulation Experiments

We generate the simulated dataset according to the MARAC(P, Q) model specified by (1) and (3). We simulate the autoregressive coefficients $\mathbf{A}_p, \mathbf{B}_p$ such that they satisfy the stationarity condition specified in Theorem 6 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 5, we plot the simulated $\mathbf{A}_1, \mathbf{B}_1, \mathbf{\Sigma}_r, \mathbf{\Sigma}_c$ when $(M, N) = (20, 20)$.

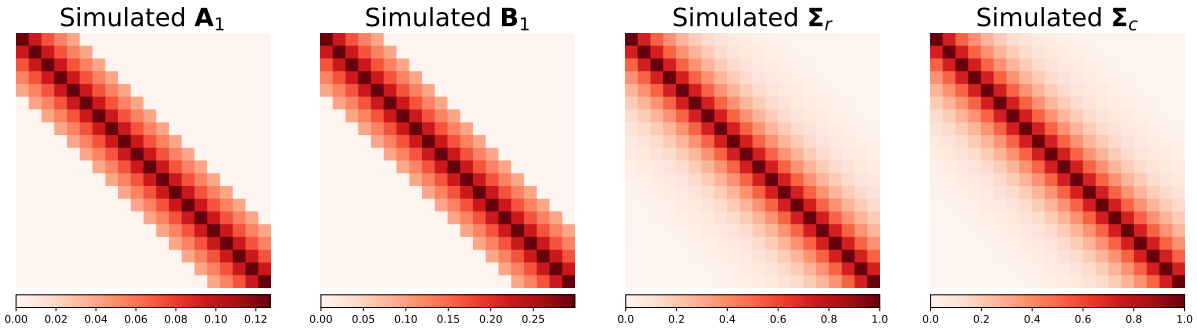

Figure 5: Visualization of the simulated $\mathbf{A}_1, \mathbf{B}_1, \mathbf{\Sigma}_r, \mathbf{\Sigma}_c$ with $M = N = 20$.

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:

$$k_\eta(s_1, s_2) = \left(\frac{1}{4\pi} + \frac{\eta}{12\pi}\right) - \frac{\eta}{8\pi}\sqrt{\frac{1 - \langle s_1, s_2\rangle}{2}}, \quad s_1, s_2 \in \mathbb{S}^2, \tag{83}$$

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 at $\rho_i(\mathbf{K}) \approx i^{-r}$ with $r \in [1.3, 1.5]$.

We randomly sample g_1, g_2, g_3 from Gaussian processes with covariance kernel being the Lebedev kernel in (83). Finally, we simulate the vector time series $\mathbf{z}_t$ using a VAR(1) process. In Figure 6, we visualize the simulated functional parameters as well as the vector time series from one random draw.

In Figure 7, 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.

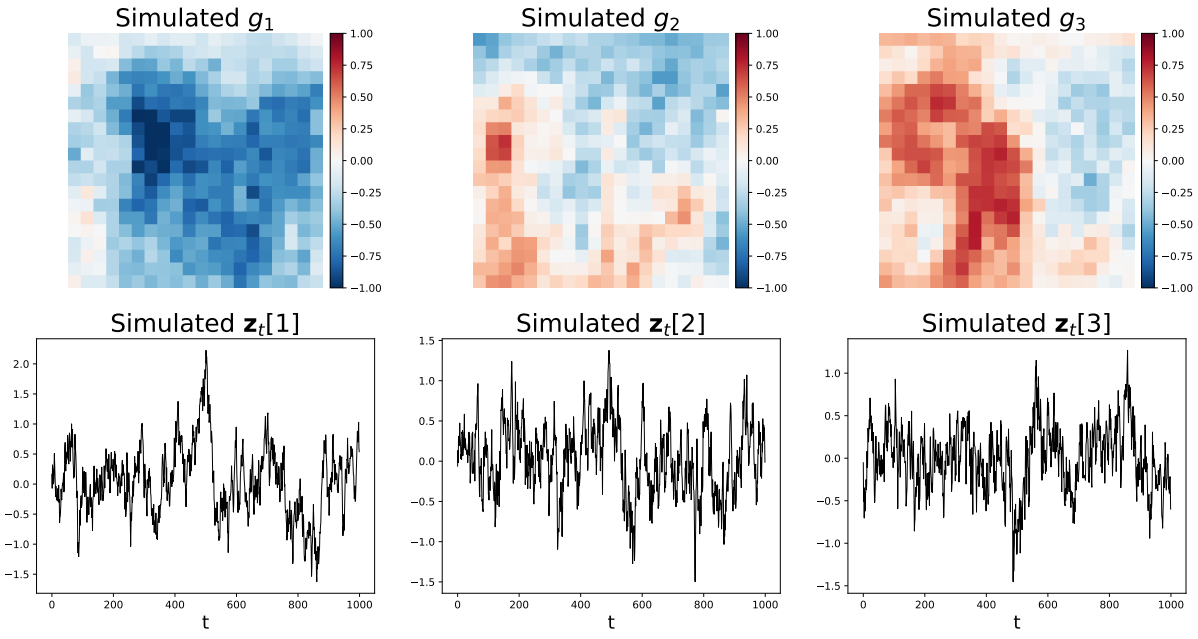

Figure 6: Simulated functional parameters g_1, g_2, g_3 evaluated on a 20×20 spatial grid (top row) and the corresponding auxiliary vector time series (bottom row).

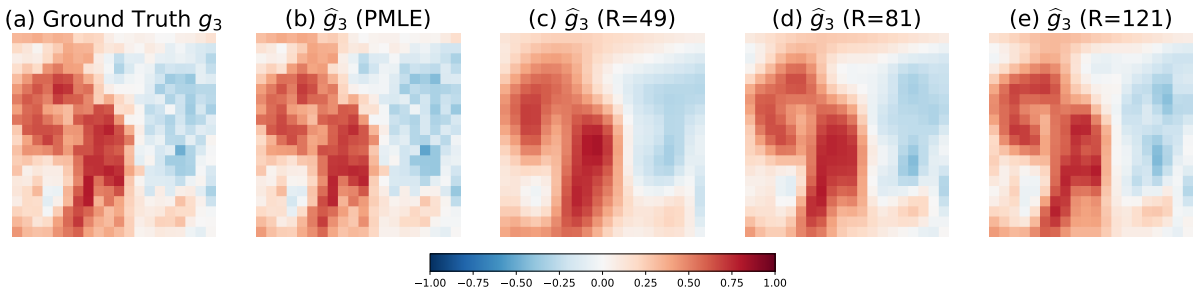

Figure 7: Ground truth g_3 (panel (a)) against the penalized MLE estimator $\widehat{g}_3$ (panel (b)) and the truncated penalized MLE estimator $\widehat{g}_3$ using $R \in \{49, 81, 121\}$ basis functions. $M = 20$.