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

February 6, 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 time series of matrices with observations distributed on a fixed 2-D spatial grid, i.e., the spatio-temporal data, 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: Matrix Autoregressive Model, Auxiliary Covariates, Reproducing Kernel Hilbert Space (RKHS), Tensor Data Model, Spatio-Temporal Forecast

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. In this paper, we focus on the matrix-valued data whose 2-D grid contains spatial/geographical information of the individual observations, i.e., the spatio-temporal data. As a concrete example, we visualize the global Total Electron Content (TEC) distribution in Figure 1 from the field of geophysics. 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 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 2.5°-latitude-by-5°-longitude spatial grid. Our statistical challenge is to make forecasts of the future TEC matrices with the historical, high-dimensional TEC matrix time series.

Auxiliary vector time series covariates bring up an additional challenge. In the last panel 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 Earth's global TEC distribution. These auxiliary covariates carry additional information related to the matrix time series data dynamics. To see this, Figure 1(A) and 1(B) have similar auxiliary covariates and TEC distributions, but 1(C) has a dramatic decrease of the Sym-H index and a much higher TEC near the equator. The key distinction between the two time series is that, compared with vector time series data, matrix time series data has more spatial flavor.

Adding the auxiliary covariates benefits the forecasting purposes and enables domain

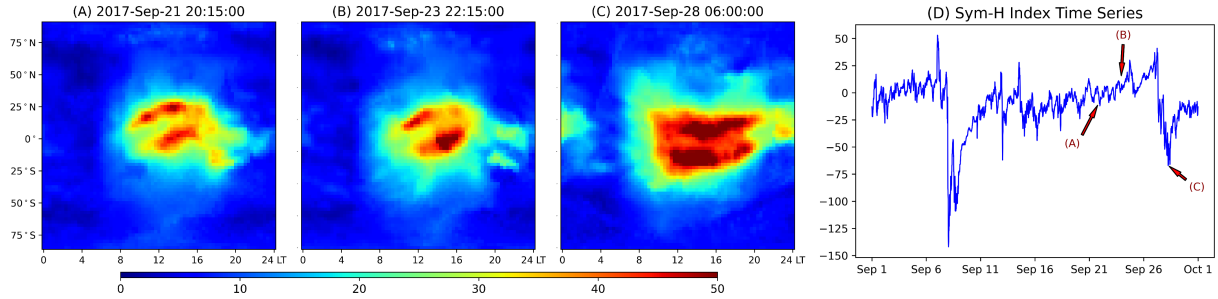

Figure 1: An example of matrix time series with auxiliary vector time series. Panels (A)-(C) show the global Total Electron Content (TEC) distribution at three timestamps on the latitude-local-time grid (source: the IGS TEC database (Hernández-Pajares et al., 2009)). Panel (D) plots the auxiliary Sym-H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

scientists to understand the interplay between different data modalities and how the information from the non-spatial data disseminates among the spatial locations. Therefore, a statistical methodology that could benefit these modeling contexts has high demands.

To state our modeling context more formally, we denote the matrix times series as $\mathbf{X}_t \in \mathbb{R}^{M \times N}$ and the auxiliary vector time series as $\mathbf{z}_t \in \mathbb{R}^D$. Our goal is to predict $\mathbf{X}_t$ given $\mathbf{X}_{t'}$ and $\mathbf{z}_{t'}$ for $t' < t$. In the remainder of this section, we briefly review the related literature and outline our unique contribution as compared to the existing approaches.

The most straightforward approach is to vectorize all the matrices as MN-dimensional vectors and then run the Vector Autoregression (VAR) (Stock and Watson, 2001) with exogenous covariates. However, vectorizing matrix data leads to the loss of structural information of the matrices and also requires a significant amount of parameters given the high-dimensionality of the data. Specifically, such a VAR model requires estimating $\mathcal{O}(MN \cdot (MN + D))$ parameters, which then requires a significant amount of training data and computation, making such a naive approach undesirable.

The VAR approach is essentially conducting elementwise least-square regression on each element of $\mathbf{X}_t$, and each regression model takes the vectorized matrix predictors and the auxiliary vector predictors as inputs. *Scalar-on-tensor regression* (Zhou et al., 2013; Guhaniyogi et al., 2017; Li et al., 2018; Papadogeorgou et al., 2021) is tackling a similar problem without vectorizing the matrix/tensor predictors but assume low-rank structure over the regression coefficient to reduce the total amount of parameters and thus the computational burden. 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 run scalar-on-tensor regression model still requires significant number of parameters and more importantly, it fails to take the structural information of the response into account.

The regression model that accommodates the matrix structure of the response is the flipped version of scalar-on-tensor regression called *tensor-on-scalar regression* (Rabusseau and Kadri, 2016; Sun and Li, 2017; Li and Zhang, 2017; Guha and Guhaniyogi, 2021). Typically, low-order scalar/vector predictors are mapped to high-order matrix/tensor responses via taking the tensor-vector product with a high-order tensor coefficient that has a low-rank structure following either the CANDECOMP/PARAFAC (CP) (Sun and Li, 2017) or Tucker decomposition (Li and Zhang, 2017), potentially with additional sparsity structure on the coefficients when dealing with ultra-high dimensionality. As compared to these works, our model requires both matrix responses and matrix predictors in a single framework, and as we will demonstrate in more detail, we take a very different approach for modeling the tensor coefficient associated with the auxiliary vector predictors, by imposing a locally-smooth instead of low-rank structure that leads to drastically different theoretical results and estimating algorithm.

The statistical framework that can incorporate matrix/tensor as both predictors and response is called the *tensor-on-tensor regression* (Lock, 2018; Liu et al., 2020). The most relevant literature in this category are the *matrix/tensor autoregression* models (Chen et al., 2021; Li and Xiao, 2021; Hsu et al., 2021; Wang et al., 2024) since we are dealing with time series data. In these works, the matrix/tensor predictors are mapped to matrix/tensor responses via *multi-linear transformations* that greatly reduce the amount of parameters when modeling such high-dimensional structural data. Our work extends this framework to allow for auxiliary low-order vector time series predictors to enter the model by borrowing the insight from the *tensor-on-scalar regression* models. Consequently, our method incorporates predictors of non-uniform modes (matrix as 2-mode tensor and vector as 1-mode tensor) in a unified framework. As we have demonstrated earlier in Figure 1, merging the information of tensor data with varying modes has very concrete motivation in real applications.

In our modeling context, where we attempt to map the non-spatial vector predictors to the spatial matrix response, our key assumption is that the vector predictor has similar effect on neighboring locations in the response. This translates into the assumption that the tensor coefficient associated with the vector predictor is spatially-smooth. There are two main techniques for building the smoothness into tensor coefficients. The first technique is to estimate the tensor coefficient with total-variation (TV) regularization (Wang et al., 2017; Shen et al., 2022; Sun et al., 2023). TV-regularization creates coefficient maps that are piecewise smooth and has sharp edges, enabling the regression model to select sub-regions with significant coefficients. However, the non-convexity of the TV penalty requires proximal or augmented Lagrangian method for solving the optimization problem, making the asymptotic analysis of the resulting estimator difficult. The second technique, which

turns out to be much simpler, is the kernel method (Kang et al., 2018). As seen in Section 2, we assume that the tensor coefficients are discrete evaluations of functional parameters from a Reproducing Kernel Hilbert Space (RKHS). We facilitate the estimation of the functional parameters with 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 matrix autoregressive setting.

To encapsulate, our paper has two major contributions. Firstly, we build a unified matrix autoregression framework for spatio-temporal data that allows for the presence of lower-order scalar/vector time series covariates. Such a framework has strong application motivation where domain scientists are curious of pooling the information of spatial and non-spatial data for predictions and inference. The framework also bridges regression methodologies with tensor predictors and responses of varying 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 RKHS functional norm penalty. We establish the joint asymptotics of the autoregressive coefficients and the functional parameters under fixed and high matrix dimensionality regimes and propose efficient exact and approximate alternating minimization algorithm for estimation and validate it with extensive simulations and real applications.

The remainder of the paper is organized as follows. We outline our model formally in

Section 2 and provide model interpretations and comparisons in sufficient details. 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 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 described in Section 6 and make conclusions in Section 7. Technical proofs and additional details of the simulation and real data applications are deferred to Appendices.

2 Model

2.1 Notation

We adopt the following notations throughout the article. Tensor (with at least three modes) is denoted by calligraphic bold-face letters (e.g. $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{G}}$). We will use uppercase bold-face letters (e.g. $\mathbf{X}, \mathbf{G}$) for matrix (two-mode tensor) and lowercase bold-face letters (e.g. $\mathbf{x}, \mathbf{z}, \mathbf{e}_i$ for vector (one-mode tensor) 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 reserve the subscript t inside the square bracket to index time. Any fibers and slices of tensor are subscript-ed with colons such as $[\boldsymbol{\mathcal{G}}]_{ij:}$ and $[\boldsymbol{\mathcal{G}}]_{::d}$ and thus any row and column of a matrix can be denoted as $[\mathbf{X}_t]_{i:}$ and $[\mathbf{X}_t]_{:j}$. If the slices of tensor/matrix is based on the last mode such as $[\boldsymbol{\mathcal{G}}]_{::d}$ and $[\mathbf{X}_t]_{:j}$, we will often omit the colons and write them as $[\boldsymbol{\mathcal{G}}]_d$ and $[\mathbf{X}_t]_j$ for brevity. For any two K-mode tensors $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}$, we define their inner product as: $\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle = \sum_{i_1,\ldots,i_K} [\boldsymbol{\mathcal{X}}]_{i_1,\ldots,i_K} \cdot [\boldsymbol{\mathcal{Y}}]_{i_1,\ldots,i_K}$, and we use $\|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}}$

to denote the Frobenius norm of a tensor and obviously one has $\|\boldsymbol{\mathcal{X}}\|_{\mathrm{F}} = \sqrt{\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \rangle}$.

For an arbitrary tensor $\boldsymbol{\mathcal{X}}$ of size $d_1 \times \cdots \times d_k \times \cdots \times d_K$ and a matrix $\mathbf{U}_k$ of size $r_k \times d_k$, the *tensor k-mode product* $\boldsymbol{\mathcal{X}} \times_k \mathbf{U}_k$ is a tensor of size $d_1 \times \cdots \times r_k \times \cdots \times d_K$ with $[\boldsymbol{\mathcal{X}} \times_k \mathbf{U}_k]_{i_1 \ldots j \ldots j_K} = \sum_{i_k} [\boldsymbol{\mathcal{X}}]_{i_1 \ldots i_k \ldots j_K} \cdot [\mathbf{U}_k]_{j i_k}$. Following Li and Zhang (2017), the *tensor k-mode 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 $d_1 \times \cdots \times d_K$ where $[\boldsymbol{\mathcal{G}} \bar{\times} \mathbf{z}]_{i_1 \ldots i_K} = \sum_{i_{K+1}} [\boldsymbol{\mathcal{G}}]_{i_1 \ldots i_K i_{K+1}} \cdot [\mathbf{z}]_{i_{K+1}}$. We use $\mathbf{vec}\,(\cdot)$ to denote tensor/matrix vectorization where the first index changes the fastest, so $\mathbf{vec}\,(\mathbf{X})$ is simply stacking the columns of $\mathbf{X}$ into a long vector. For tensor $\boldsymbol{\mathcal{X}}$ of size $d_1 \times \cdots \times d_K$, we use $\mathbf{X}_{(k)}$ to denote its k-mode matricization, where $\mathbf{X}_{(k)}$ is of size $d_k \times \prod_{m \neq k} d_m$ and each row is a vectorized fiber of the tensor along its k^{th} mode. 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. More details about these tensor/matrix operations can be found in Kolda and Bader (2009).

In our specific modeling context where we consider the spatio-temporal data, we observe a sequence of matrix spatial data $\mathbf{X}_1, \ldots, \mathbf{X}_T, \ldots$ of size $M \times N$ each. Without loss of generality, we assume that all $S = MN$ spatial locations are evenly-spaced grid points on a 2-D $M \times N$ grid within the domain $\bar{\mathbb{S}} \coloneqq [0,1] \times [0,1]$. 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 data is denoted as s_{ij}. Oftentimes we will 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 will use square brackets around integers to denote index set, i.e. $[N] = \{1, 2, \ldots, N\}$. Thus more compactly, we have $\mathbb{S} = \{(\frac{i}{M}, \frac{j}{N}) | i \in [M], j \in [N]\}$. We will use lowercase letters such as $f(\cdot), g(\cdot)$ to denote functions. In particular, we will use $k(\cdot,\cdot) : \bar{\mathbb{S}} \times \bar{\mathbb{S}} \mapsto \mathbb{R}$ to denote a

spatial kernel function and the corresponding Reproducing Kernel Hilbert Space (RKHS) as $\mathbb{H}_k$.

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 time series 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 in abbreviation, 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 auxiliary vector predictor, and $\mathbf{E}_t$ are an i.i.d. noise terms 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).

To connect $\mathbf{X}_t$, which is a spatial data, to $\mathbf{z}_{t-q}$, which is a non-spatial data, we use a linear map with coefficient $\boldsymbol{\mathcal{G}}_q$. Any particular element $[\boldsymbol{\mathcal{G}}_q]_{ijd}$ measures the predictive effect of the d^{th} auxiliary covariate at lag-q for the $(i, j)^{\text{th}}$ element of $\mathbf{X}_t$. Element-wisely for $\mathbf{X}_t$, 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 covariates enter the model via an elementwise linear model.

Alternatively, if one stacks $\mathbf{X}_1, \ldots, \mathbf{X}_T$ into a tensor $\boldsymbol{\mathcal{X}}_T \in \mathbb{R}^{M\times N\times T}$, $\mathbf{z}_1, \ldots, \mathbf{z}_T$ into a matrix $\mathbf{Z}_T \in \mathbb{R}^{D\times T}$, and $\mathbf{E}_1, \ldots, \mathbf{E}_T$ into an error tensor $\boldsymbol{\mathcal{E}}_T$, one can interpret the MARAC

model in (1) as a tensor decomposition where:

$$\boldsymbol{\mathcal{X}}_T = \sum_{p=1}^{P} \boldsymbol{\mathcal{X}}_{T-p} \times_1 \mathbf{A}_p \times_2 \mathbf{B}_p + \sum_{q=1}^{Q} \boldsymbol{\mathcal{G}}_q \times_3 \mathbf{Z}_{T-q}^{\top} + \boldsymbol{\mathcal{E}}_T, \tag{3}$$

where the definition of $\boldsymbol{\mathcal{X}}_{T-p}, \mathbf{Z}_{T-q}$ requires extending the indices of $\mathbf{X}_t$ and $\mathbf{z}_t$ to negative integers, but the confusion should be minimal here. From (3), we can see that the autoregressive term is simply conducting a *multi-linear transformation* on the lagged tensor data. Using multi-linear transformation greatly reduces the total amount of parameters (Chen et al., 2021; Li and Xiao, 2021) in that each lagged predictor previously requires M^2N^2 parameters but now only requires $M^2 + N^2$ parameters in $\mathbf{A}_p, \mathbf{B}_p$. Both of the autoregressive term and the auxiliary covariate term appears to be a tensor factor model (Wang et al., 2019; Chen et al., 2022) except that the temporal factors are fixed to be $\boldsymbol{\mathcal{X}}_{T-p}, \mathbf{Z}_{t-q}$.

As we have briefly mentioned in Section 1, we assume that $\boldsymbol{\mathcal{G}}_q$ has a spatially-smooth structure. To be more specific, we assume that $\forall d \in [D]$, $[\boldsymbol{\mathcal{G}}_q]_{ijd}$ and $[\boldsymbol{\mathcal{G}}_q]_{uvd}$ are similar if $(i, j), (u, v)$ are close neighbors in $\mathbb{S}$. To formally characterize this, 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 $g_{q,d}$ 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 $g_{q,d}$ and thus the tensor coefficient $\boldsymbol{\mathcal{G}}_q$.

In order to distinguish our model from the existing approaches such as matrix autoregression and kriging for spatio-temporal data, we compare them through the lens of the assumptions made on the noise process $\mathbf{E}_t$. In the kriging framework (Cressie, 1986), typically $\text{Cov}(\mathbf{vec}\,(\mathbf{E}_t)\,, \mathbf{vec}\,(\mathbf{E}_{t'}))$ is characterized by a spatio-temporal kernel that captures the dependencies among spatial and temporal neighbors. In Chen et al. (2021), the authors do not consider the local spatial dependencies of $\mathbf{E}_t$ and further assumes that $\mathbf{E}_t$ are i.i.d. The covariance of $\mathbf{vec}\,(\mathbf{E}_t)$, denoted as $\boldsymbol{\Sigma}$, is assumed to have a separable Kronecker-product

structure:

$$\mathbf{vec}\left(\mathbf{E}_t\right) \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r\right), \quad t \in [T] \tag{4}$$

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 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 and accounting for the grid geometry of the spatial observations. However, it neglects the local spatial correlations of the noises. In Hsu et al. (2021), the matrix autoregression framework in Chen et al. (2021) is generalized to adapt to spatial data via fixed-rank co-kriging (FRC) (Cressie and Johannesson, 2008). Specifically, $\mathbf{\Sigma} = \mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r + \mathbf{F}\mathbf{M}\mathbf{F}^\top$, 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 framework assumes that noises are more correlated if either they are spatially close or they share the same row/column index and is thus more flexible for matrix spatial data.

Different from these previous works, our framework leverages the presence of the auxiliary covariates. Similar to Chen et al. (2021), we also assume that $\mathbf{E}_t$ follows (4). 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:

$$\mathrm{Cov}(\mathbf{vec}\left(\mathbf{E}_t\right), \mathbf{vec}\left(\mathbf{E}_{t'}\right)) = \delta_{tt'} \cdot \mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r + \mathbf{F}\mathbf{M}\mathbf{F}^\top$$

$$\mathbf{F} = [(\boldsymbol{\mathcal{G}}_1)_{(3)}^\top : \cdots : (\boldsymbol{\mathcal{G}}_Q)_{(3)}^\top], \quad \mathbf{M} = [\mathrm{Cov}(\mathbf{z}_{t-q_1}, \mathbf{z}_{t'-q_2})]_{q_1,q_2\in[Q]}$$

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. Our model is flexible enough to account for dependencies from all sources including time dependency via the auto-correlated auxiliary covariates, grid dependency via the Kronecker-product structured error covariance $\mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r$ and the local dependency via the spatially-smooth tensor coefficients $\boldsymbol{\mathcal{G}}$. We summarize the comparisons in Table 1.

Combining (1) and (4) yields the complete MARAC(P, Q) model. The collection of

Method	Local Dependency	Grid Dependency	Time Dependency
Kriging (Cressie, 1986)	✓	✗	✓
MAR (Chen et al., 2021)	✗	✓	✗
FRC (Hsu et al., 2021)	✓	✓	✗
MARAC (Our Method)	✓	✓	✓

Table 1: Comparison of the flexibility of the error covariances $\mathrm{Cov}(\mathbf{vec}\left(\mathbf{E}_t\right), \mathbf{vec}\left(\mathbf{E}_{t'}\right))$.

model parameters $\mathbf{\Theta} = \{\mathbf{A}_1, \ldots, \mathbf{A}_P, \mathbf{B}_1, \ldots, \mathbf{B}_P, \boldsymbol{\mathcal{G}}_1, \ldots, \boldsymbol{\mathcal{G}}_Q, \mathbf{\Sigma}_r, \mathbf{\Sigma}_c\}$ include three categories of parameters: the autoregressive parameters $\{\mathbf{A}_p, \mathbf{B}_p\}_{p=1}^P$; the auxiliary covariate parameters $\{\boldsymbol{\mathcal{G}}_q\}_{q=1}^Q$; and the covariance components $\mathbf{\Sigma}_r, \mathbf{\Sigma}_c$. In Section 3 we will introduce an alternating minimization algorithm to estimate $\mathbf{\Theta}$.

Before concluding this section, we want to finalize the model discussion by vectorizing both sides of MARAC(P, Q) in (1), which yields:

$$\mathbf{x}_t = \sum_{p=1}^{P} \left(\mathbf{B}_p \otimes \mathbf{A}_p\right) \mathbf{x}_{t-p} + \sum_{q=1}^{Q} \mathbf{G}_q^\top \mathbf{z}_{t-q} + \mathbf{e}_t, \tag{5}$$

where $\mathbf{x}_t = \mathbf{vec}\left(\mathbf{X}_t\right), \mathbf{e}_t = \mathbf{vec}\left(\mathbf{E}_t\right)$, and recall that $\mathbf{G}_q = (\boldsymbol{\mathcal{G}}_q)_{(3)}$. The Kronecker product form of the autoregressive coefficient (also known as the transition matrix in VAR model) as well as the error covariance greatly reduces the number of parameters and introduce low-rank structure over the high-dimensional parameter space. The spatially-smooth assumption over rows of $\mathbf{G}_q$, on the other hand, introduce smoothness over the high-dimensional parameter space. This *low-rank plus smooth* parameter structure simplifies the computation for high-dimensional matrix predictors and preserves the interpretability for relatively low-dimensional auxiliary covariates for the spatio-temporal forecasting problem.

An alternative formulation for $\boldsymbol{\mathcal{G}}_q$ would be a low-rank Tucker decomposition form (Li and Zhang, 2017). We choose locally-smooth over low-rank structure with the purpose

of explicitly modeling the spatial smoothness of the coefficients and avoiding tuning the multi-linear rank of $\boldsymbol{\mathcal{G}}_q$ during the model selection step. We leave the low-rank model for future research and focus on the RKHS framework for the current paper.

3 Estimating Algorithm

In this section, we discuss parameter estimations for the MARAC(P, Q) model in (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 information criterion (IC) used for selecting the lag hyper-parameters and will validate the consistency of different ICs numerically in Section 5.

3.1 Penalized Maximum Likelihood Estimation (MLE)

To estimate the MARAC(P, Q) model parameters $\boldsymbol{\Theta}$, we propose a penalized maximum likelihood estimation (MLE) approach. Following the Gaussianity assumption in (4), we can write the negative log-likelihood with a squared RKHS functional norm penalty as:

$$\mathfrak{L}_\lambda(\boldsymbol{\Theta}) = -\frac{1}{T}\sum_{t\in[T]} \ell\left(\mathbf{X}_t|\mathbf{X}_{t-1},\ldots,\mathbf{X}_{t-P},\mathbf{z}_{t-1},\ldots,\mathbf{z}_{t-Q};\boldsymbol{\Theta}\right) + \frac{\lambda}{2}\sum_{q\in[Q]}\sum_{d\in[D]} \|g_{q,d}\|^2_{\mathbb{H}_k}, \quad (6)$$

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

$$\ell\left(\mathbf{X}_t|\mathbf{X}_{t-1},\ldots,\mathbf{X}_{t-P},\mathbf{z}_{t-1},\ldots,\mathbf{z}_{t-Q};\boldsymbol{\Theta}\right) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_c\otimes\boldsymbol{\Sigma}_r| - \frac{1}{2}\mathbf{r}_t^\top\left(\boldsymbol{\Sigma}_c^{-1}\otimes\boldsymbol{\Sigma}_r^{-1}\right)\mathbf{r}_t, \quad (7)$$

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. To estimate the parameters in $\boldsymbol{\Theta}$, one needs to solve a constrained minimization problem:

$$\min_{\boldsymbol{\Theta}}\left\{\mathfrak{L}_\lambda(\boldsymbol{\Theta}), \quad \text{s.t. } g_{q,d}(s_{ij}) = [\boldsymbol{\mathcal{G}}_q]_{ijd}, \forall s_{ij}\in\mathbb{S}\right\}. \quad (8)$$

We now define the functional norm penalty in (6) 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):

$$k(s_{ij}, s_{uv}) = \sum_{r=1}^{\infty} \lambda_r \psi_r(s_{ij}) \psi_r(s_{uv}), \quad s_{ij}, s_{uv} \in [0,1]^2, \tag{9}$$

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 from the RKHS $\mathbb{H}_k$ endowed with kernel $k(\cdot,\cdot)$ is defined according to van Zanten and van der Vaart (2008):

$$\|g\|_{\mathbb{H}_k} = \sqrt{\sum_{r=1}^{\infty} \frac{\beta_r^2}{\lambda_r}}, \quad \forall g \in \mathbb{H}_k, \qquad \text{where } g(\cdot) = \sum_{r=1}^{\infty} \beta_r \psi_r(\cdot). \tag{10}$$

Given any $\lambda > 0$ in (6), the generalized representer theorem (Schölkopf et al., 2001) suggests that the solution of the functional parameters, denoted as $\{\widetilde{g}_{q,d}(\cdot), \forall q, d\}$, of the minimization problem (8), with all other parameters held fixed, is a linear combination of the representers $\{k(\cdot, s), \forall s \in \mathbb{S}\}$ plus a linear combination of the basis functions $\{\phi_1, \ldots, \phi_J\}$ of the null space of $\mathbb{H}_k$, i.e.,

$$\widetilde{g}_{q,d}(\cdot) = \sum_{s \in \mathbb{S}} \gamma_s k(\cdot, s) + \sum_{j=1}^{J} \alpha_j \phi_j(\cdot), \quad \|\phi_j\|_{\mathbb{H}_k} = 0, \tag{11}$$

where we omit the subscript (q,d) for the coefficient γ_s, α_j for brevity but they are 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 (11), the minimization problem in (8) 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* (8) *is equivalent to the*

following unconstrained kernel ridge regression problem:

$$\min_{\mathbf{\Theta}} \left\{ \frac{1}{2} \log |\mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r| + \frac{1}{2T} \sum_{t \in [T]} \mathbf{r}_t^\top \left(\mathbf{\Sigma}_c^{-1} \otimes \mathbf{\Sigma}_r^{-1} \right) \mathbf{r}_t + \frac{\lambda}{2} \sum_{q \in [Q]} tr\left(\mathbf{\Gamma}_q^\top \mathbf{K} \mathbf{\Gamma}_q \right) \right\}, \tag{12}$$

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* $[\mathbf{K}]_{u_1 u_2} = k(s_{i_1 j_1}, s_{i_2 j_2}), \forall s_{ijl} \in \mathbb{S}, u_l = i_l + (j_l - 1)M, l = 1, 2$ *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.*

We give the proof in Appendix A.1. 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 (12) with an alternating minimization algorithm (Attouch et al., 2013) where we update one parameter at a time while keeping the others fixed at their current values in the algorithm. We update the parameters following the order of: $\mathbf{A}_1 \to \mathbf{B}_1 \to \cdots \to \mathbf{A}_P \to \mathbf{B}_P \to \mathbf{\Gamma}_1 \to \cdots \to \mathbf{\Gamma}_Q \to \mathbf{\Sigma}_r \to \mathbf{\Sigma}_c \to \mathbf{A}_1 \to \cdots$. The alternating minimization algorithm is chosen here due to 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 (12) with respect to $\mathbf{A}_p$ at the $(l+1)^{\text{th}}$ iteration, it suffices to solve the following least-square problem:

$$\min_{\mathbf{A}_p} \left\{ \sum_{t \in [T]} \text{tr} \left(\left[\widetilde{\mathbf{X}}_{t,-p} - \mathbf{A}_p \mathbf{X}_{t-p} \left(\mathbf{B}_p^{(l)} \right)^\top \right]^\top \mathbf{\Sigma}_r^{-1} \left[\widetilde{\mathbf{X}}_{t,-p} - \mathbf{A}_p \mathbf{X}_{t-p} \left(\mathbf{B}_p^{(l)} \right)^\top \right] \mathbf{\Sigma}_c^{-1} \right) \right\}, \tag{13}$$

where $\widetilde{\mathbf{X}}_{t,-p}$ is the partial residual of $\mathbf{X}_t$, leaving out the lag-p autoregressive predictor:

$$\widetilde{\mathbf{X}}_{t,-p} = \mathbf{X}_t - \sum_{p' < p} \mathbf{A}_{p'}^{(l+1)} \mathbf{X}_{t-p'} \left(\mathbf{B}_{p'}^{(l+1)} \right)^\top - \sum_{p' > p} \mathbf{A}_{p'}^{(l)} \mathbf{X}_{t-p'} \left(\mathbf{B}_{p'}^{(l)} \right)^\top - \sum_{q \in [Q]} \mathbf{K} \mathbf{\Gamma}_q^{(l)} \mathbf{z}_{t-q}. \tag{14}$$

To simplify the notation, we define $\mathbf{\Phi}(\mathbf{A}_t, \mathbf{B}_t, \mathbf{\Sigma}) = \sum_t \mathbf{A}_t^\top \mathbf{\Sigma}^{-1} \mathbf{B}_t$, where $\mathbf{\Sigma}, \mathbf{A}_t, \mathbf{B}_t$ are arbitrary matrices/vectors with conformal matrix sizes. Oftentimes, we will be using $\mathbf{\Phi}(\mathbf{A}_t, \mathbf{\Sigma})$ for short if $\mathbf{A}_t = \mathbf{B}_t$. Solving (13) yields the following updating formula for $\mathbf{A}_p^{(l+1)}$:

$$\mathbf{A}_p^{(l+1)} \leftarrow \mathbf{\Phi}\left(\widetilde{\mathbf{X}}_{t,-p}^\top, \mathbf{B}_p^{(l)} \mathbf{X}_{t-p}^\top, \mathbf{\Sigma}_c^{(l)}\right) \mathbf{\Phi}\left(\mathbf{B}_p^{(l)} \mathbf{X}_{t-p}^\top, \mathbf{\Sigma}_c^{(l)}\right)^{-1} \tag{15}$$

Similar result can be obtained for updating $\mathbf{B}_p^{(l+1)}$:

$$\mathbf{B}_p^{(l+1)} \leftarrow \mathbf{\Phi}\left(\widetilde{\mathbf{X}}_{t,-p}, \mathbf{A}_p^{(l+1)} \mathbf{X}_{t-p}, \mathbf{\Sigma}_r^{(l)}\right) \mathbf{\Phi}\left(\mathbf{A}_p^{(l+1)} \mathbf{X}_{t-p}, \mathbf{\Sigma}_r^{(l)}\right)^{-1} \tag{16}$$

For updating $\mathbf{\Gamma}_q$, we can rewrite the optimization problem in (12) by keeping only the terms with respect to $\mathbf{\Gamma}_q$, or its vectorized version $\boldsymbol{\gamma}_q = \mathbf{vec}\left(\mathbf{\Gamma}_q\right)$, as:

$$\min_{\boldsymbol{\gamma}_q} \left\{ \frac{1}{2T} \mathbf{\Phi}\left(\widetilde{\mathbf{x}}_{t,-q} - \left(\mathbf{z}_{t-q}^\top \otimes \mathbf{K}\right) \boldsymbol{\gamma}_q, \mathbf{\Sigma}_c^{(l)} \otimes \mathbf{\Sigma}_r^{(l)}\right) + \frac{\lambda}{2} \boldsymbol{\gamma}_q^\top \left(\mathbf{I}_D \otimes \mathbf{K}\right) \boldsymbol{\gamma}_q \right\}, \tag{17}$$

where $\widetilde{\mathbf{x}}_{t,-q}$ is the partial residual of $\mathbf{vec}\left(\mathbf{X}_t\right)$ by leaving out the lag-q auxiliary covariates. We omit the definition of $\widetilde{\mathbf{x}}_{t,-q}$ here but it can be similarly written as (14). Solving the kernel ridge regression in (17) leads to the updating rule for $\boldsymbol{\gamma}_q^{(l+1)}$:

$$\boldsymbol{\gamma}_q^{(l+1)} \leftarrow \left[\left(\sum_{t\in[T]} \mathbf{z}_{t-q} \mathbf{z}_{t-q}^\top\right) \otimes \mathbf{K} + \lambda T \left(\mathbf{I}_D \otimes \mathbf{\Sigma}_c^{(l)} \otimes \mathbf{\Sigma}_r^{(l)}\right)\right]^{-1} \left[\sum_{t\in[T]} \left(\mathbf{z}_{t-q} \otimes \widetilde{\mathbf{x}}_{t,-q}\right)\right]. \tag{18}$$

The current algorithm that relies on kernel ridge regression requires inverting a square matrix of dimension $MND \times MND$, which might be computationally intensive when the dimensionality of the matrix or vector predictor is high. Later in Section 3.2, we propose an approximation to (18) to speed up the computation under high dimensionality.

The updating rule of $\mathbf{\Sigma}_r^{(l+1)}$ and $\mathbf{\Sigma}_c^{(l+1)}$ can be easily derived by taking their derivative in (12) and setting it to zero. So we state the result here directly:

$$\mathbf{\Sigma}_r^{(l+1)} \leftarrow \frac{1}{NT} \mathbf{\Phi}\left(\widetilde{\mathbf{X}}_t^\top, \mathbf{\Sigma}_c^{(l)}\right) \tag{19}$$

$$\mathbf{\Sigma}_c^{(l+1)} \leftarrow \frac{1}{MT} \mathbf{\Phi}\left(\widetilde{\mathbf{X}}_t, \mathbf{\Sigma}_r^{(l+1)}\right) \tag{20}$$

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

The algorithm cycles through (15), (16), (18), (19) and (20) and terminates when $\mathbf{B}_p^{(l)} \otimes \mathbf{A}_p^{(l)}$, $\boldsymbol{\mathcal{G}}_q^{(l)}$, $\boldsymbol{\Sigma}_c^{(l)} \otimes \boldsymbol{\Sigma}_r^{(l)}$ has their relative changes between iterations fall under a pre-specified threshold. We make two additional remarks on the algorithm:

Remark 2 *(Model 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 constant* $c \neq 0$ *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 check 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* $\|\mathbf{B}_p^{(l+1)} \otimes \mathbf{A}_p^{(l+1)} - \mathbf{B}_p^{(l)} \otimes \mathbf{A}_p^{(l)}\|_{\mathrm{F}}$ *is given by:*

$$\|\mathbf{B}_p^{(l+1)} - \mathbf{B}_p^{(l)}\|_{\mathrm{F}} \cdot \|\mathbf{A}_p^{(l+1)}\|_{\mathrm{F}} + \|\mathbf{B}_p^{(l)}\|_{\mathrm{F}} \cdot \|\mathbf{A}_p^{(l+1)} - \mathbf{A}_p^{(l)}\|_{\mathrm{F}}, \tag{21}$$

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

We summarize the alternating minimization algorithm below in Algorithm 1.

3.2 Computationally Efficient Penalized MLE with Kernel Truncation

The iterative algorithm in Section 3.1 requires inverting an $MND \times MND$ matrix in (18) 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 further 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

Algorithm 1 Alternating Minimization Algorithm for MARAC(P, Q) Model Estimation

Input: Matrix time series $\mathbf{X}_1, \ldots, \mathbf{X}_T$; Vector time series $\mathbf{z}_1, \ldots, \mathbf{z}_T$; Spatial kernel function $k(\cdot, \cdot)$; Penalty tuning parameter λ; Max iteration L; Convergence threshold $\tau = 10^{-4}$.

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

Compute the kernel gram matrix $\mathbf{K}$, set $l \leftarrow 0, \delta \leftarrow 1$.

while $l \leq L$ and $\delta \geq \tau$ **do**

$\delta \leftarrow 1$.

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

Update $\mathbf{A}_p^{(l+1)}$ by (15).

Update $\mathbf{B}_p^{(l+1)}$ by (16).

$\delta \leftarrow \max\Big(\delta, \|\mathbf{B}_p^{(l+1)} \otimes \mathbf{A}_p^{(l+1)} - \mathbf{B}_p^{(l)} \otimes \mathbf{A}_p^{(l)}\|_{\mathrm{F}}/MN\Big)$, compute by (21).

end for

for $q = 1, \ldots, Q$ **do**

Update $\mathbf{\Gamma}_q^{(l+1)}$ by (18).

$\delta \leftarrow \max\Big(\delta, \|\mathbf{\Gamma}_q^{(l+1)} - \mathbf{\Gamma}_q^{(l)}\|_{\mathrm{F}}/\sqrt{MND}\Big)$.

end for

Update $\mathbf{\Sigma}_r^{(l+1)}$ by (19).

Update $\mathbf{\Sigma}_c^{(l+1)}$ by (20).

$\delta \leftarrow \max\Big(\delta, \|\mathbf{\Sigma}_c^{(l+1)} \otimes \mathbf{\Sigma}_r^{(l+1)} - \mathbf{\Sigma}_c^{(l)} \otimes \mathbf{\Sigma}_r^{(l)}\|_{\mathrm{F}}/MN\Big)$, compute similarly by (21).

$l \leftarrow l + 1$.

end while

For each $p \in [P]$, let $c_p \leftarrow \mathrm{sign}[\mathrm{tr}(\mathbf{A}_p^{(l)})]/\|\mathbf{A}_p^{(l)}\|_{\mathrm{F}}$, then $\widehat{\mathbf{A}}_p \leftarrow c_p \cdot \mathbf{A}_p^{(l)}, \widehat{\mathbf{B}}_p \leftarrow (c_p)^{-1} \cdot \mathbf{B}_p^{(l)}$.

For each $q \in [Q]$, let $[\widehat{\boldsymbol{\mathcal{G}}}_q]_{ijd} \leftarrow \langle [\mathbf{K}]_{u:}^{\top}, [\mathbf{\Gamma}_q^{(l)}]_{:d} \rangle$, where $u = i + (j-1)M$.

Output: $\widehat{\mathbf{A}}_1, \ldots, \widehat{\mathbf{A}}_P, \widehat{\mathbf{B}}_1, \ldots, \widehat{\mathbf{B}}_P, \widehat{\boldsymbol{\mathcal{G}}}_1, \ldots, \widehat{\boldsymbol{\mathcal{G}}}_Q, \widehat{\mathbf{\Sigma}}_r, \widehat{\mathbf{\Sigma}}_c$.

following the order of $\boldsymbol{\gamma}_{q,1} \to \cdots \to \boldsymbol{\gamma}_{q,D}$. Such a procedure only requires inverting a matrix of size $MN \times MN$, which could still be high-dimensional, but potentially requires a larger number of iterations.

To circumvent the issue of inverting high-dimensional 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 τ of the rows and columns of the matrix and only use the representers at the subsampled "knots" as the basis functions. This is equivalent to selecting a subset of columns of $\mathbf{K}$ as $\mathbf{K}_R$, and $R = \tau^2 MN$. In this subsection, we consider an alternative approach by truncating the Mercer decomposition in (9) instead. A similar technique can be found in Kang et al. (2018).

Given the eigen-decomposition of $k(\cdot,\cdot)$ in (9), 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 estimation. The choice of R depends on the decaying rate of the eigenvalue sequence $\{\lambda_r\}_{r=1}^{\infty}$ (thus the smoothness of the underlying functional parameters), and thus the choice of the kernel. Our simulation result shows that the estimation and prediction errors of the model shrink monotonically as $R \to \infty$. Therefore, R can be chosen based on the computational resources available.

It is beneficial to think about the kernel truncation from the perspective of the functional norm penalty (10). When λ_r is small, a non-zero coefficient for the eigen-function $\psi_r(\cdot)$ would lead to a high functional norm penalty. Typically, these eigen-functions are less smooth as compared to those associated with larger eigenvalues. As a result, the kernel truncation procedure speeds up the computation at the cost of getting a smoother estimator for the functional parameters, as we will demonstrate empirically in Appendix C.

Given the kernel truncation, any functional parameter $g_{q,d}$ is now approximated by: $g_{q,d}(\cdot) \approx \sum_{r\in[R]}[\boldsymbol{\theta}_{q,d}]_r\psi_r(\cdot)$, where $\boldsymbol{\theta}_{q,d}$ contains the linear approximation parameters. We can estimate $\boldsymbol{\Theta}_q = [\boldsymbol{\theta}_{q,1} : \cdots : \boldsymbol{\theta}_{q,D}] \in \mathbb{R}^{R\times D}$ by translating the kernel ridge regression problem in (12) to the following ridge regression problem:

$$\min_{\boldsymbol{\Theta}} \left\{ \frac{1}{2}\log|\boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r| + \frac{1}{2T}\sum_{t\in[T]} \mathbf{r}_t^\top \left(\boldsymbol{\Sigma}_c^{-1} \otimes \boldsymbol{\Sigma}_r^{-1}\right)\mathbf{r}_t + \frac{\lambda}{2}\mathrm{tr}\left(\boldsymbol{\Theta}_q^\top \boldsymbol{\Lambda}_R^{-1}\boldsymbol{\Theta}_q\right) \right\}, \qquad (22)$$

where $\mathbf{r}_t = \mathbf{x}_t - \sum_p(\mathbf{B}_p \otimes \mathbf{A}_p)\mathbf{x}_{t-p} - \sum_q \mathbf{K}_R\boldsymbol{\Theta}_q\mathbf{z}_{t-q}$ is the vectorized residual, $\mathbf{K}_R$ satisfies $[\mathbf{K}_R]_{ur} = \psi_r(s_{ij}), u = i + (j-1)M$, and $\boldsymbol{\Lambda}_r = \mathrm{diag}(\lambda_1, \ldots, \lambda_R)$, with λ_r being the r^{th} largest eigenvalue of the Mercer decomposition of $k(\cdot,\cdot)$. Solving (22) with respect to $\boldsymbol{\theta}_q = \mathbf{vec}\left(\boldsymbol{\Theta}_q\right)$ is thus reduced to a ridge regression problem with the solver given by

$$\boldsymbol{\theta}_q^{(l+1)} \leftarrow \left[\boldsymbol{\Phi}\left(\mathbf{z}_{t-q}^\top \otimes \mathbf{K}_R, \boldsymbol{\Sigma}^{(l)}\right) + \lambda T\left(\mathbf{I}_D \otimes \boldsymbol{\Lambda}_R^{-1}\right)\right]^{-1} \boldsymbol{\Phi}\left(\mathbf{z}_{t-q}^\top \otimes \mathbf{K}_R, \widetilde{\mathbf{x}}_{t,-q}, \boldsymbol{\Sigma}^{(l)}\right), \qquad (23)$$

where $\boldsymbol{\Sigma}^{(l)} = \boldsymbol{\Sigma}_c^{(l)} \otimes \boldsymbol{\Sigma}_r^{(l)}$. As compared to (18) that requires inverting a matrix of size $MND \times MND$, the step in (23) only requires inverting a matrix of size $RD \times RD$. The estimation procedure for the other parameters remains the same as the non-truncated case in Section 3.1 and thus we omit the detailed descriptions here.

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 based on procedures such as cross-validation. The choice of P and Q requires a more formal model selection criterion. In this section, we jointly select P and Q by using 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 validate their empirical consistency via simulation experiments in Section 5.

Let $\widehat{\mathbf{\Theta}}$ 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 MARAC(P, Q) model, we can define the AIC and the BIC as follows,

$$\text{AIC}(P, Q, \lambda) = -2 \sum_{t \in [T]} \ell(\mathbf{X}_t | \mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-P}, \mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-Q}, \widehat{\mathbf{\Theta}}) + 2 \cdot \mathbf{df}_{P,Q,\lambda}, \tag{24}$$

$$\text{BIC}(P, Q, \lambda) = -2 \sum_{t \in [T]} \ell(\mathbf{X}_t | \mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-P}, \mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-Q}, \widehat{\mathbf{\Theta}}) + \log(T) \cdot \mathbf{df}_{P,Q,\lambda}. \tag{25}$$

To calculate $\mathbf{df}_{P,Q,\lambda}$, we decompose it into the sum of three components: 1) for 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 covariance structure $\widehat{\mathbf{\Sigma}}_r, \widehat{\mathbf{\Sigma}}_c$, 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 (18), we define the sum of their degrees of freedom as:

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

where $\widetilde{\mathbf{K}} = \left(\frac{1}{T} \sum_{t \in [T]} \mathbf{z}_{t-q} \mathbf{z}_{t-q}^\top \right) \otimes \mathbf{K}$. It is evident that 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. Based on numerical studies in Section 5, we find that BIC is a consistent lag selection criterion for the MARAC model.

4 Theoretical Analysis

This section presents major theoretical results of the MARAC model. First, we point out the condition under which the matrix and vector time series are *jointly stationary*. Under this stationarity condition, we then establish the consistency and asymptotic normality of the penalized MLE estimators under *fixed* matrix dimensionality (i.e., M, N are fixed) as $T \to \infty$. Finally, we relax the constraint that M, N are fixed and allow for $M, N \to \infty$ as

$T \to \infty$ and then derive the convergence rate of the penalized MLE estimator and also the optimal order of the functional norm penalty tuning parameter λ with respect to M, N, T.

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.

4.1 Stationarity Condition

To facilitate the theoretical analysis for the MARAC model, we make an additional assumption for the vector time series $\mathbf{z}_t$, which is not required in numerical studies.

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

$$\mathbf{z}_t = \sum_{\widetilde{q}=1}^{\widetilde{Q}} \mathbf{C}_{\widetilde{q}} \mathbf{z}_{t-\widetilde{q}} + \boldsymbol{\nu}_t, \tag{26}$$

where $\mathbf{C}_{\widetilde{q}} \in \mathbb{R}^{D \times D}$ is the lag-$\widetilde{q}$ transition matrix and $\{\boldsymbol{\nu}_t\}_{t=-\infty}^{\infty}$ is a serially-independent noise process with bounded fourth-order moment. Also, $\{\boldsymbol{\nu}_t\}_{t=-\infty}^{\infty}$ is independent of $\{\mathbf{E}_t\}_{t=-\infty}^{\infty}$.

Remark 5 *With assumption 4, we can combine the MARAC model for $\mathbf{X}_t$ in (5) and the VAR process for $\mathbf{z}_t$ in (26) as a single VAR(L) model, with $L = \max(P, Q, \widetilde{Q})$:*

$$\begin{bmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{bmatrix} = \sum_{l=1}^{L} \begin{bmatrix} (\mathbf{B}_l \otimes \mathbf{A}_l) \odot \mathbf{1}_{\{l \le P\}} & \mathbf{G}_l^{\top} \odot \mathbf{1}_{\{l \le Q\}} \\ \mathbf{O}_{D \times S} & \mathbf{C}_l \odot \mathbf{1}_{\{l \le \widetilde{Q}\}} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{t-l} \\ \mathbf{z}_{t-l} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_t \\ \boldsymbol{\nu}_t \end{bmatrix}, \tag{27}$$

with $\odot$ being elementwise 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 Equation(27) shows, $\mathbf{z}_{t'}$ can help forecast $\mathbf{x}_t, t > t'$ but not the opposite, indicating that $\mathbf{z}_t$ is exogenous for $\mathbf{x}_t$.

Given the joint vector autoregressive model in (27), 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* $|y| \leq 1$*, we have*

$$\det\left[\mathbf{I}_S - \sum_{p=1}^{P} (\mathbf{B}_p \otimes \mathbf{A}_p)\, y^p\right] \neq 0, \quad \det\left[\mathbf{I}_D - \sum_{\widetilde{q}=1}^{\widetilde{Q}} \mathbf{C}_{\widetilde{q}} y^{\widetilde{q}}\right] \neq 0. \tag{28}$$

As a special case where $P = \widetilde{Q} = 1$*, we require that* $\rho(\mathbf{A}_1) \cdot \rho(\mathbf{B}_1) < 1$ *and* $\rho(\mathbf{C}_1) < 1$ *for MARAC to be jointly stationary, where* $\rho(\cdot)$ *is the spectral radius of a square matrix.*

The proof of Theorem 6 largely follows from the stationarity of the joint autoregressive model in (27) and is relegated to Appendix A.2.

It can be seen from Theorem 6 that the stationarity of the matrix and vector time series relies on the stationarity of the autoregressive components of the MARAC(P, Q) and VAR$(\widetilde{Q})$ models, respectively. The tensor coefficients $\boldsymbol{\mathcal{G}}_1, \ldots, \boldsymbol{\mathcal{G}}_Q$ do not enter either characteristic polynomials since we assume that $\mathbf{z}_t$ follows a self-evolving process that is independent of $\mathbf{X}_t$, i.e. $\mathbf{z}_t$ is exogenous with respect to $\mathbf{X}_t$. The MARAC model can be extended to the joint autoregressive process (27) and the assumption on the vector time series can be removed so that one can predict $\mathbf{X}_t$ and $\mathbf{z}_t$ jointly using their historical values, but here we stick to the simpler case for ease of the presentation of our theoretical analysis.

4.2 Finite Spatial-Dimension Asymptotics

In this subsection, we establish the consistency and asymptotic normality of the MARAC model estimators $\widehat{\boldsymbol{\Theta}} = \{\widehat{\mathbf{A}}_1, \widehat{\mathbf{B}}_1, \ldots, \widehat{\mathbf{A}}_P, \widehat{\mathbf{B}}_P, \widehat{\boldsymbol{\mathcal{G}}}_1, \ldots, \widehat{\boldsymbol{\mathcal{G}}}_Q, \widehat{\boldsymbol{\Sigma}}_r, \widehat{\boldsymbol{\Sigma}}_c\}$ under the scenario that M, N being *finite*. 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 $\widehat{\boldsymbol{\mathcal{G}}}_q$. Later in Section 4.3, we will discuss the *double* asymptotics where both $S, T \to \infty$.

For the remainder of the paper, we denote the true model coefficient with an asterisk superscript, e.g., $\mathbf{A}_1^*, \mathbf{B}_1^*, \boldsymbol{\mathcal{G}}_1^*, \boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_c^* \otimes \boldsymbol{\Sigma}_r^*$. We make the following assumption on $\mathbf{K}$:

Assumption 7 *The minimum eigenvalue of* $\mathbf{K}$ *satisfies:* $\underline{\rho}(\mathbf{K}) = \underline{c} > 0$.

As a result of assumption 7, every $\boldsymbol{\mathcal{G}}_q^*$ has a unique kernel decomposition: $\mathbf{vec}\left(\boldsymbol{\mathcal{G}}_q^*\right) = (\mathbf{I}_D \otimes \mathbf{K})\boldsymbol{\gamma}_q^*$. Our asymptotic results for $\widehat{\boldsymbol{\mathcal{G}}}_q$ will be presented via the asymptotics for $\widehat{\boldsymbol{\gamma}}_q$. We first establish the consistency of the covariance matrix estimator $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}_c} \otimes \widehat{\boldsymbol{\Sigma}_r}$ in Proposition 8.

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

The proof of the proposition is relegated to Appendix A.3. Given this result, we can further establish the joint 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. error process* $\{\mathbf{E}_t\}_{t=-\infty}^{\infty}$ *following* (4) *and assumption 4, 7 and the stationarity condition in Theorem 6 hold and* $\lambda = o(T^{-1/2})$. *Additionally, we assume that* $\underline{\rho}(\boldsymbol{\Gamma}_0) = \underline{c}' > 0$ *where* $\boldsymbol{\Gamma}_0 = \mathrm{Cov}([\mathbf{x}_t^\top \mathbf{z}_t^\top]^\top)$.

Then with fixed S *and known* P, Q, *the MLEs of the MARAC model,* $\{\widehat{\mathbf{A}}_p, \widehat{\mathbf{B}}_p, \widehat{\boldsymbol{\gamma}}_q\}, p \in [P], q \in [Q]$ *are jointly asymptotically normal:*

$$\sqrt{T}\begin{bmatrix} \mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^*\right) \\ \mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^*\right) \\ \mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{R}}} - \boldsymbol{\mathcal{R}}^*\right) \end{bmatrix} \xrightarrow{d.} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Xi}\right), \tag{29}$$

where $\widehat{\boldsymbol{\mathcal{A}}}, \widehat{\boldsymbol{\mathcal{B}}}, \widehat{\boldsymbol{\mathcal{R}}}$ *are defined as* $[\widehat{\boldsymbol{\mathcal{A}}}]_{::p} = \widehat{\mathbf{A}}_p, [\widehat{\boldsymbol{\mathcal{B}}}]_{::p} = \widehat{\mathbf{B}}_p^\top$, $[\widehat{\boldsymbol{\mathcal{R}}}]_{:dq} = \widehat{\boldsymbol{\gamma}}_{q,d}$ *and* $\boldsymbol{\mathcal{A}}^*, \boldsymbol{\mathcal{B}}^*, \boldsymbol{\mathcal{R}}^*$ *are the corresponding true coefficients. The asymptotic covariance matrix* $\boldsymbol{\Xi}$ *is defined as*

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

and $\mathbf{W}_t$ *is defined as*

$$\mathbf{W}_t = \left[\left[\mathbf{B}_1^* \mathbf{X}_{t-1}^\top \cdots \mathbf{B}_P^* \mathbf{X}_{t-P}^\top\right] \otimes \mathbf{I}_M; \mathbf{I}_N \otimes \left[\mathbf{A}_1^* \mathbf{X}_{t-1} \cdots \mathbf{A}_P^* \mathbf{X}_{t-P}\right]; [\mathbf{z}_{t-1}^\top, \ldots \mathbf{z}_{t-Q}^\top] \otimes \mathbf{K}\right],$$

and $\mathbf{H} = \mathrm{E}[\mathbf{W}_t^\top (\boldsymbol{\Sigma}^*)^{-1} \mathbf{W}_t] + \boldsymbol{\eta}\boldsymbol{\eta}^\top$ *where* $\boldsymbol{\eta} = [\mathbf{vec}\left(\mathbf{A}_1^*\right)^\top \cdots \mathbf{vec}\left(\mathbf{A}_P^*\right)^\top \mathbf{0}^\top]^\top$.

We relegate the proof to Appendix A.4. The established asymptotic distribution (29) has a convergence rate of $\sqrt{T}$, which nests the result of MAR in Chen et al. (2021) as a special case. Under a fixed dimensionality S, the kernel ridge penalty should vanish to zero to eliminate the bias and the functional parameters $g_{q,d} \in \mathbb{H}_k$ are estimated only at fixed locations, so it is fully parametric and the rate of convergence is the same as the autoregressive coefficients.

4.3 High Spatial and Temporal 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 both S and T go to infinity. To simplify the proof, we assume that the covariance matrix $\boldsymbol{\Sigma}^*$ is a diagonal matrix $\boldsymbol{\Sigma}^* = \sigma^2 \mathbf{I}$ with σ being known. The result presented in Theorem 12 below, however, can be generalized to arbitrary *known* $\boldsymbol{\Sigma}^*$ with additional assumptions over the eigen-spectrum of $\boldsymbol{\Sigma}^*$. If $\boldsymbol{\Sigma}^*$ is unknown and needs to be estimated, this will significantly complicate the theoretical analysis; therefore, we leave it for future work.

To establish the high-dimensional convergence rate of the MARAC estimators, we make the following 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* $k(s_{ij}, s_{uv})$ *can be decomposed into the product of a row kernel* $k_1(s_i, s_u)$ *and a column kernel* $k_2(s_j, s_v)$. *Both* k_1, k_2 *have their eigenvalues decaying at a polynomial rate* $\lambda_j(k_1) \asymp j^{-r_0}, \lambda_j(k_2) \asymp j^{-r_0}$, $r_0 > 1/2$.

Assumption 11 elicits a simple eigen-spectrum characterization of the spatial kernel $k(\cdot, \cdot)$, whose eigenvalue can be written as $\lambda_i(k_1)\lambda_j(k_2)$, and thus $\lambda_{ij}(k) \asymp (ij)^{-r_0}$. Also, the gram matrix $\mathbf{K}$ is separable, i.e. $\mathbf{K} = \mathbf{K}_2 \otimes \mathbf{K}_1$ and $\rho_{ij}(\mathbf{K}) = \rho_i(\mathbf{K}_1)\rho_j(\mathbf{K}_2)$, where $\mathbf{K}_1, \mathbf{K}_2$ are the gram matrix for kernel k_1, k_2, respectively.

Under assumption 10, we further have $\rho_i(\mathbf{K}_1) \to M\lambda_i(k_1)$ and $\rho_j(\mathbf{K}_2) \to N\lambda_j(k_2)$ and thus $\rho_{ij}(\mathbf{K}) \to S\lambda_{ij}(k)$, as $M, N \to \infty$. 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 construction for easier theoretical derivations.

In assumption 11, we assume the kernel separability to accommodate the grid structure of the spatial locations. We do not constrain r_0 to be an integer but just a parameter that characterizes the smoothness of the functional parameters. Examples of kernels k_1, k_2 that satisfy assumption 11 could be the reproducing kernel for the Sobolev space $\mathcal{W}_2^{r_0}([0, 1])$. See Cui et al. (2018) for more discussions. Now we 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* $\mathbf{\Sigma}^* = \sigma^2\mathbf{I}_S$

known. Then as $S, T \to \infty$ (D is fixed) and $S \log S/T \to 0$, and under the additional assumptions that:

1. $\min(M, N) \asymp \max(M, N) \asymp \sqrt{S}$;
2. $\gamma_S := \lambda/S \to 0$, $\gamma_S \cdot S^{r_0} \to \infty$ *or* $\gamma_S \cdot S^{r_0} \to C_1$ *as* $S \to \infty$, *where* $C_1 > 0$ *is a constant;*
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 blocks of the covariance matrix of* $[\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top$ *and* $c_{0,S}$ *is a constant only related to* S;
4. $\bar{\rho}(\mathbf{K})$ *is bounded for any* S *and* $\mathbf{K}$ *is positive definite,*

then we have:

$$AR_{err} = \frac{1}{\sqrt{P}S} \sqrt{\sum_{p=1}^{P} \left\| \widehat{\mathbf{B}}_p \otimes \widehat{\mathbf{A}}_p - \mathbf{B}_p^* \otimes \mathbf{A}_p^* \right\|_{\mathrm{F}}^2} \leq O_P\left(\sqrt{\frac{C_g \cdot \gamma_S}{c_{0,S} \cdot S}}\right) + O_P\left(\sqrt{\frac{D}{c_{0,S} \cdot TS}}\right), \quad (30)$$

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

Given any $\mathbf{z}_t$ *time series generated by a VAR*$(\widetilde{Q})$ *process, we further have:*

$$AC_{err} \leq O_P\left(\frac{\sqrt{\gamma_S^{-1/2r_0}}}{\sqrt{T}\sqrt[4]{S}}\right) + O_P(\sqrt{\gamma_S}) + O_P\left(\frac{1}{\sqrt{S}}\right) + O_P\left(\frac{\sqrt{\gamma_S^{-1}}}{\sqrt{TS}}\right), \quad (31)$$

where $AC_{err} = \sqrt{(TS)^{-1} \sum_{t=1}^{T} \left\| \sum_{q=1}^{Q} \left(\widehat{\boldsymbol{\mathcal{G}}}_q - \boldsymbol{\mathcal{G}}_q^* \right) \bar{\times} \mathbf{z}_{t-q} \right\|_{\mathrm{F}}^2}$. *$AR_{err}$ and AC_{err} are the elementwise RMSE for the autoregressive parameters and auxiliary covariates predictions, respectively.*

The proof is relegated to Appendix A.5. Here in Theorem 12, (30) gives the error bound of the autoregressive coefficients and (31) gives the error bound of the prediction made by the auxiliary time series, which contains the functional parameter estimators. As a special case for (30), when $\gamma_S = 0$ and S is fixed, the convergence rate for the autoregressive coefficients is $O_P(T^{-1/2})$, which coincides with the result in Theorem 9.

Remark 13 (Optimal Choice of λ and Phase Transition) *According to our proof in Appendix A.5, the error bound* (31) *can be decomposed as:*

$$\underbrace{O_P\left(\frac{1}{\sqrt{T}\sqrt[4]{S}}\sqrt{\gamma_S^{-1/2r_0}}\right)+O_P(\sqrt{\gamma_S})}_{nonparametric\ error}+\underbrace{O_P\left(\sqrt{C_g\gamma_S+c_0C_gS^{-1}+DT^{-1}+c_0D(TS\gamma_S)^{-1}}\right)}_{autoregressive\ error}$$

where the autoregressive error stems from the inaccurate estimation of $\mathbf{B}_p^* \otimes \mathbf{A}_p^*$. *The nonparametric error, if there is no autoregressive error, is similar to 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* $\lambda \asymp n^{-2r_0/(2r_0+1)}$. *Here, if no autoregressive error, the optimal tuning parameter* γ_S *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 error bound due to assumption 11 where the eigenvalue of* $\mathbf{K}$, *i.e.* $\rho_i(\mathbf{K})$, *decays 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* $S \asymp T^c$, *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* $\log_T S$, *as illustrated by Figure 2.*

The Figure 2 suggests that the higher the matrix dimensionality, the smaller the optimal γ_S is and thus less functional norm penalty is needed. This is an intuitive result that when we have more locations available, we have a denser set of observations and thus less smoothing is needed.

It is interesting to point out that the auxiliary covariates prediction error bound, i.e. AC_{err}, is at the order of $1/\sqrt{S}$ except when the matrix dimensionality is high and the

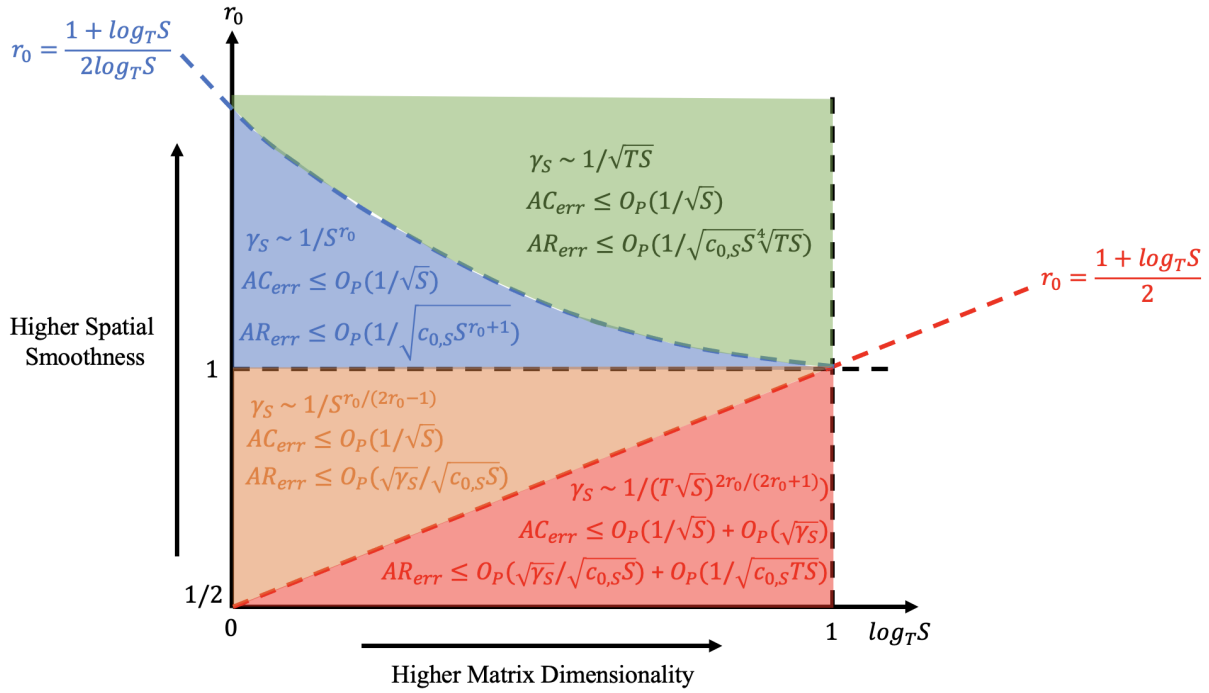

Figure 2: Phase transition diagram for the optimal tuning parameter $\gamma_S = \lambda/S$ and the corresponding error bound for the autoregressive coefficient (30) and functional coefficient (31) (abbreviated as AR_{err} and AC_{err}) with respect to spatial smoothness r_0 and matrix dimensionality $\log_T S$. Essentially, we require that $S \log S/T \to 0$ but here we plot the right border at $S \asymp T$ for visualization purpose.

functional parameter is not very smooth (the red region in Figure 2). In this special case, the estimation of the functional parameter is close to fitting a large number of individual linear regression models for each pixel, which cannot take advantage of the information from the spatial neighborhood and thus leading to higher prediction error.

In (30), 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 on the spatial covariance of $\mathbf{X}_t$. If $c_{0,S} \geq c_0 > 0$ for some universal constant c_0, we summarize the convergence rate result of MARAC estimators in Table 2. Unfortunately, in practice, it is

common to have $c_{0,S} \to 0$ as $S \to \infty$, which make the autoregressive coefficient converge at a slower rate. Note that the constant does not affect the convergence rate of the functional parameter in (31). We leave the constant $c_{0,S}$ here in (30) to give a general result and leave the characterization of $c_{0,S}$ under specific assumptions to future work.

Condition ($\log_T S = c$)	Optimal γ_S	AR Error	AC Error
$c \geq \frac{1}{2r_0-1}, r_0 \geq 1$	$O((TS)^{-\frac{1}{2}})$	$O_P(T^{-\frac{1}{4}}S^{-\frac{3}{4}})$	$O_P(S^{-\frac{1}{2}})$
$c < \frac{1}{2r_0-1}, r_0 \geq 1$	$O(S^{-r_0})$	$O_P(S^{-\frac{r_0+1}{2}})$	$O_P(S^{-\frac{1}{2}})$
$c \geq 2r_0 - 1, \frac{1}{2} < r_0 < 1$	$O(S^{-r_0(2r_0-1)})$	$O_P(S^{-\frac{r_0(2r_0-1)+1}{2}})$	$O_P(S^{-\frac{1}{2}})$
$c < 2r_0 - 1, \frac{1}{2} < r_0 < 1$	$(T\sqrt{S})^{-\frac{2r_0}{2r_0+1}}$	$O_P((TS)^{-\frac{1}{2}})+$ $O_P((T\sqrt{S})^{-\frac{r_0}{2r_0+1}}S^{-\frac{1}{2}})$	$O_P(S^{-\frac{1}{2}})+$ $O_P((T\sqrt{S})^{-\frac{r_0}{2r_0+1}})$

Table 2: Summary of optimal tuning parameter γ_S and estimators convergence rate following (30) and (31), under the assumption that $c_{0,S} \geq c_0 > 0, \forall S$ and $S \asymp T^c$ for some constant c. Only under the last regime will we achieve an optimal tuning order of γ_S that is close to the classic nonparametric optimal rate at $(TS)^{-2r_0/(2r_0+1)}$. The reason is because the error bound of the nonparametric component is dominated by the estimation error of the autoregressive component.

5 Simulation Experiments

5.1 Convergence Rate of the Estimator

In this section, we validate the consistency and convergence rate of the MARAC estimators output by Algorithm 1 as S and T grow. 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 $\mathbf{\Sigma}_r^*, \mathbf{\Sigma}_c^*$ to have symmetric

banded structures. To simulate g_1, g_2, g_3 (we drop the lag subscript for brevity) from the RKHS $\mathbb{H}_k$, we choose $k(\cdot,\cdot)$ as the Lebedev kernel (Kennedy et al., 2013) and generate g_1, g_2, g_3 randomly from Gaussian processes with the Lebedev kernel. Finally, we simulate the auxiliary vector time series $\mathbf{z}_t \in \mathbb{R}^3$ from a VAR(1) process. See Appendix C for additional details about the detailed simulation setups and visualizations.

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 $\boldsymbol{\Theta} \in \{\mathbf{B}_1 \otimes \mathbf{A}_1, \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r, \boldsymbol{\mathcal{G}}_1, \boldsymbol{\mathcal{G}}_2, \boldsymbol{\mathcal{G}}_3\}$ where we report the average RMSE for $\boldsymbol{\mathcal{G}}_1, \boldsymbol{\mathcal{G}}_2, \boldsymbol{\mathcal{G}}_3$ for brevity. The dataset is configured with $M \in \{5, 10, 20, 40\}$ and $N = M$. For each M, we train the MARAC model 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 test set. We simulate a very long time series and choose the train set starting from the first frame and the validation set right after the train set. The test set is always fixed as the last $5,000$ frames of the time series. All results are reported with 20 repeated experiments in Figure 3.

The result in Figure 3 shows that all model estimators are consistent and the convergence rate, under a fixed matrix 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 matrix dimensionality S increases, the RMSE for $\widehat{\mathbf{B}}_1 \otimes \widehat{\mathbf{A}}_1$ becomes even smaller, echoing the result in (30) and Table 2. The RMSE of the nonparametric estimators $\widehat{g}_1, \widehat{g}_2, \widehat{g}_3$ also decays at a rate of $1/\sqrt{T}$, also echoing the result in Theorem 9. The RMSE of the covariance matrix estimator $\widehat{\boldsymbol{\Sigma}_c} \otimes \widehat{\boldsymbol{\Sigma}_r}$ is also consistent, confirming the conclusion from 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

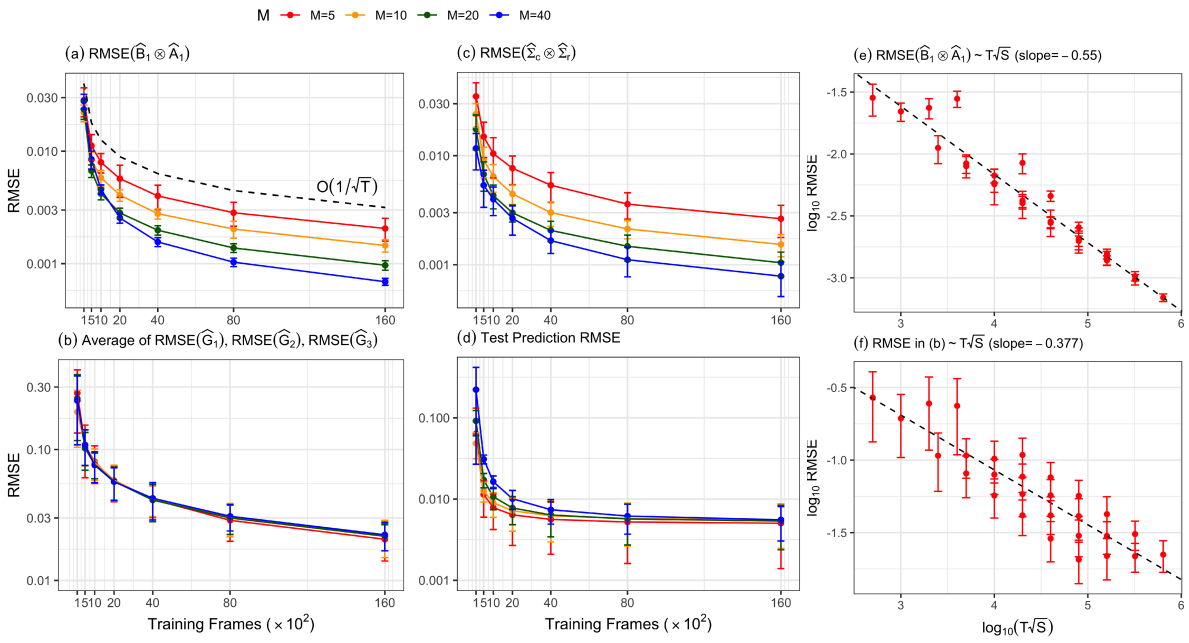

Figure 3: (a)-(d) MARAC estimator RMSE and test set prediction RMSE under various training length T (x-axis) and matrix dimensionality S (colored lines). Test prediction RMSE is subtracted by 1 when plotting for better visualization, where 1 is the noise level of the simulated data. (e) and (f) are RMSE of the autoregressive parameters and auxiliary covariates parameters under different $T\sqrt{S}$, plotted in $\log_{10} - \log_{10}$ scale with a fitted linear regression line.

convergence rate theoretically.

In this simulation, we fix the noise level of $\{\mathbf{vec}\,(\mathbf{E}_t)\}_{t=1}^{T}$ to be unity. Thus the optimal prediction RMSE should be unity if there is no overfitting. 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 (e) and (f) of Figure 3. 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 result in the last row of Table 2. We do acknowledge that the error bound in (30) and (31) are not necessarily the tightest, so the empirical convergence rate in (e) and (f) does not align perfectly with the theory, but the theory does provide a close prediction.

All the results reported in Figure 3 are based on the PMLE framework in Algorithm 1, namely 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 Appendix C.

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 criterion 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 3, 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 repeated experiments.

From Table 3, we find that AIC tends to select the model with more autoregressive lags

	$T = 1 \times 10^3$	$T = 2 \times 10^3$	$T = 4 \times 10^3$	$T = 8 \times 10^3$
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 3: Probability that AIC and BIC select the correct P (first number), Q (second number) and (P, Q) (third number) from 100 repeated experiments.

but BIC performs consistently better under large sample sizes. This coincides with the findings in Hsu et al. (2021). In practice, when the sample size is not large, we recommend using a validation set for choosing P, Q, λ; however, BIC can be a good alternative.

5.3 Comparison with Alternative Methods

We compare our proposed method against other competing methods for the matrix autoregression task. We simulate the matrix time series $\mathbf{X}_t$ from an 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 test set prediction RMSE. Each simulation scenario is repeated 20 times.

Under each P, Q, M, N specification, we consider the following five competing methods for the modeling fitting 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):

$$\mathbf{X}_t - \sum_{p=1}^{P} \widehat{\mathbf{A}}_p \mathbf{X}_{t-p} \widehat{\mathbf{B}}_p^\top = \sum_{q=1}^{Q} \boldsymbol{\mathcal{G}}_q \bar{\times} \mathbf{z}_{t-q} + \mathbf{E}_t, \mathbf{vec}\,(\mathbf{E}_t) \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2 \mathbf{I}), \tag{32}$$

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 with only one iteration.

4. Pixel-wise autoregression (Pixel-AR): for each $1 \le i \le M, 1 \le j \le N$ we fit

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

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, as well as the 2.5% and 97.5% quantile of the prediction RMSE obtained from the 20 repeated runs, are plotted in Figure 4. Overall, our MARAC model outperforms the other competing methods under varying matrix dimensionality and lag parameters. We make two additional remarks. First, when the matrix size is small (e.g., 5×5), the vector autoregression model (VARX) performs almost as good as the MARAC model and is better than other methods. However, the performance of the VARX model quickly gets worse as the matrix becomes bigger, 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. Second, 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 bringing up prediction performances.

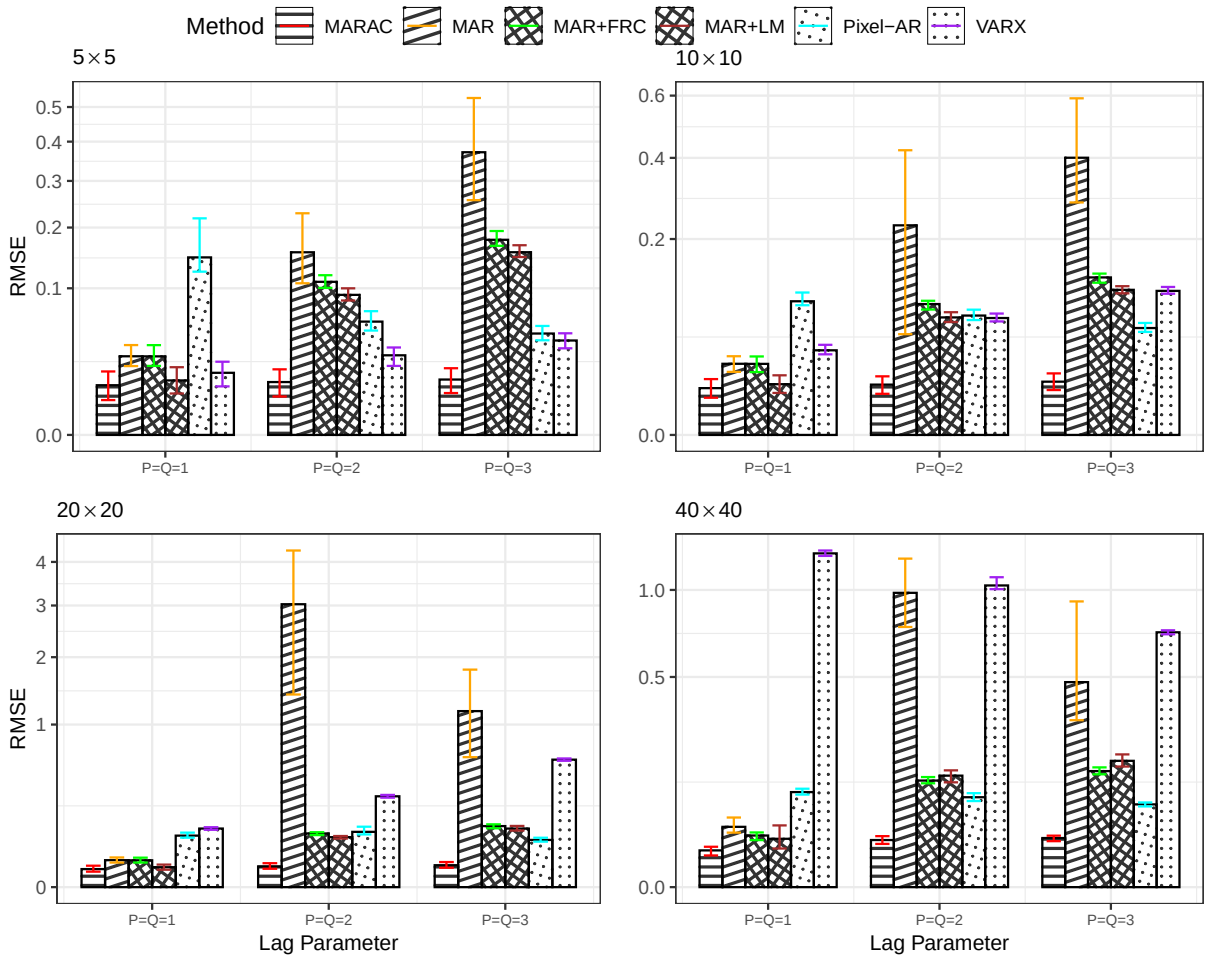

Figure 4: Test set prediction RMSE comparison across six competing methods on the matrix autoregression task. Four panels correspond to four different matrix dimensionality (labeled on the top-left corner of each panel). Test prediction RMSE is subtracted by 1 for better visualization, where 1 is the noise level of the simulated data. Error bar shows the 2.5% and 97.5% quantile of the 20 repeated runs.

6 Application to Global Total Electron Content Forecast

For real data application, we consider the problem of predicting the global total electron content (TEC) distribution. 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 download the data for September 2017, and the whole month of data form a matrix time series with $T = 2880$ and $M = 71, N = 73$. 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 follows:

$$\text{TEC}_{t+h} = \sum_{p=1}^{P} \mathbf{A}_p \text{TEC}_{t-p} \mathbf{B}_p^\top + \sum_{q=1}^{Q} \boldsymbol{\mathcal{G}}_q \bar{\times} [\text{IMF Bz}_{t-q}, \text{Sym-H}_{t-q}, \text{F10.7}_{t-q}]^\top + \mathbf{E}_t, \qquad (33)$$

where h is the forecast latency time. We consider the forecasting scenario where $h \in [1, 24]$ which corresponds to making forecasts from 15 minutes up to 6 hours ahead. At each latency time, we fit our MARAC(P, Q) model following (33) with $1 \leq P, Q \leq 3$. We fit the MARAC model with kernel truncation approximation using $R = 121$ basis functions from the truncation of the Lebedev kernel; see Appendix C for details of the kernel. As a

comparison, we also fit the MAR model with $1 \leq P \leq 3$ and followed by the MAR+LM model with $1 \leq P, Q \leq 3$, see the definition of MAR+LM model in (32). As a benchmark, we consider using TEC_{t-1} to predict TEC_{t+h} and name this naive approach 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 $\mathbf{\Sigma}_r, \mathbf{\Sigma}_c$ are diagonal when fitting all models. We zero-meaned all sets of data using the time-average of the matrix and vector time series of the training set.

In Figure 5(A), we report the pixel-wise prediction RMSE on the testing set. From the result, it appears 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 on improving the prediction performance as compared to the MAR model. This coincides with the intuition that the disturbances from the solar wind to Earth's ionosphere will affect the global TEC distribution, but with a delay in time up to several hours. The additional prediction gain from incorporating the auxiliary covariates vanish 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 5(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: they provide a single patch in the middle. The MARAC model, however, is able to

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 value, which complements the prediction made by the autoregressive component.

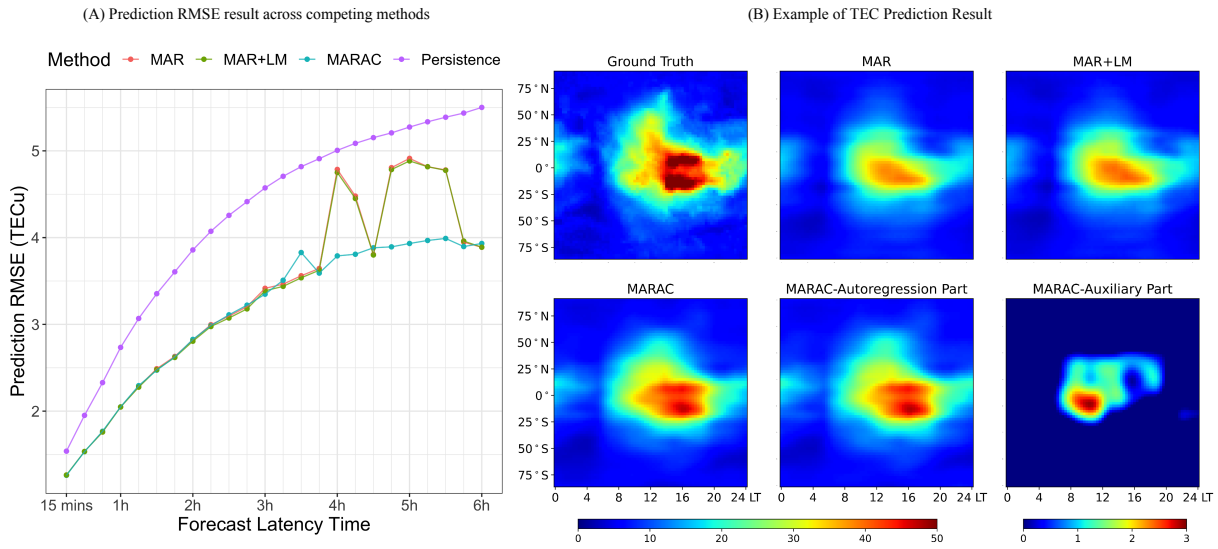

Figure 5: IGS TEC prediction results. Panel (A) shows the testing set prediction RMSE across four competing methods under 24 different latency times. Panel (B) shows an example of the predicted TEC at 10:45:00 UT, 2017-Sep-28, under the 4-hour latency time scenario. Note that the "MARAC-Auxiliary Part" plot has a different colorbar underneath it as compared to the other plots.

7 Summary

In this paper, we propose a new methodology for spatio-temporal matrix autoregressive model with exogenous vector covariates called the MARAC. 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 diverging 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 contains 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 setting for spatial data. We leave the discussions for these setups to future research.

Acknowledgments and Disclosure of Funding

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.

Rong Chen, Dan Yang, and Cun-Hui Zhang. Factor Models for High-Dimensional Tensor Time Series. *Journal of the American Statistical Association*, 117(537):94–116, 2022.

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.

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.

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.

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.

A Proof of Theoretical Results in Section 4

Throughout the whole section, 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. And $a \vee b, a \wedge b$ denotes the maximum and minimum of a and b, respectively.

We put the proofs of all main propositions and theorems in this section and delay the statements and proofs of technical lemmas that assist the proofs here in Appendix B.

A.1 Proof of Proposition 1

Proof For each function $g_{q,d}(\cdot)$, 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 $\phi_j(\cdot) = 0, \forall j$.

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

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

for any $h_{q,d}$. Consequently, the global minimizer for the constrained optimization problem (8) 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 $\mathbf{vec}\,([\boldsymbol{\mathcal{G}}]_{::d}) = \mathbf{K}\gamma_{q,d}$. The remaining results can be easily derived with simple linear algebra. ■

A.2 Proof of Theorem 6

Proof Under assumption 4 that the vector time series $\mathbf{z}_t$ follows a VAR($\widetilde{Q}$) process, we can derive that the matrix time series $\mathbf{X}_t$ and the vector time series $\mathbf{z}_t$ jointly follows a VAR($\max(P, Q, \widetilde{Q})$) process, i.e.,

$$\begin{bmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{bmatrix} = \sum_{l=1}^{\max(P,Q,\widetilde{Q})} \begin{bmatrix} (\mathbf{B}_l \otimes \mathbf{A}_l) \odot \mathbf{1}_{\{l \le P\}} & \mathbf{G}_l^\top \odot \mathbf{1}_{\{l \le Q\}} \\ \mathbf{O}_{D \times S} & \mathbf{C}_l \odot \mathbf{1}_{\{l \le \widetilde{Q}\}} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{t-l} \\ \mathbf{z}_{t-l} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_t \\ \boldsymbol{\nu}_t \end{bmatrix}. \tag{34}$$

Let $W = \max(P, Q, \widetilde{Q})$ and $\mathbf{y}_t = [\mathbf{x}_t^\top \mathbf{z}_t^\top]$. Denote the transition matrix in (34) 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 process (34) as a VAR(1) process as:

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-W+1} \end{bmatrix} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{J}_2 & \cdots & \mathbf{J}_{W-1} & \mathbf{J}_W \\ \mathbf{I}_{S+D} & \mathbf{O}_{S+D} & \cdots & \cdots & \mathbf{O}_{S+D} \\ \mathbf{O}_{S+D} & \mathbf{I}_{S+D} & \mathbf{O}_{S+D} & \cdots & \mathbf{O}_{S+D} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{O}_{S+D} & \mathbf{O}_{S+D} & \cdots & \mathbf{I}_{S+D} & \mathbf{O}_{S+D} \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-W} \end{bmatrix} + \begin{bmatrix} \mathbf{u}_t \\ \mathbf{0}_{S+D} \\ \vdots \\ \mathbf{0}_{S+D} \end{bmatrix}, \tag{35}$$

where we use $\mathbf{O}_{S+D}$ to denote a square matrix of zeros with 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 $|\lambda| \geq 1$, where $\mathbf{J}$ is the transition matrix in (35). The determinant, $\det(\lambda\mathbf{I} - \mathbf{J})$, can be simplified by column operations, resulting in

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

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

A.3 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 combination. For the vectorized MARAC$(1,1)$ model (5), we can equivalently write it as:

$$\mathbf{x}_t = \mathbf{y}_t\boldsymbol{\theta} + \mathbf{e}_t, \tag{36}$$

where $\mathbf{y}_t = [\mathbf{x}_{t-1}^\top \otimes \mathbf{I}_S; \mathbf{z}_{t-1}^\top \otimes \mathbf{K}]$ and $\boldsymbol{\theta} = [\mathbf{vec}\,(\mathbf{B}_1 \otimes \mathbf{A}_1)^\top, \boldsymbol{\gamma}_1^\top]^\top$. Using $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ to denote the precision matrix for $\mathbf{e}_t$, we can rewrite the penalized likelihood in (12) for $(\boldsymbol{\theta}, \boldsymbol{\Omega})$ as:

$$h(\boldsymbol{\theta}, \boldsymbol{\Omega}) = -\frac{1}{2}\log|\boldsymbol{\Omega}| + \frac{1}{2}\mathrm{tr}\,(\boldsymbol{\Omega}\mathbf{S}(\boldsymbol{\theta})) + \frac{\lambda}{2}\boldsymbol{\theta}^\top\widetilde{\mathbf{K}}\boldsymbol{\theta}, \tag{37}$$

where $\mathbf{S}(\boldsymbol{\theta}) = T^{-1}\sum_{t=1}^{T}(\mathbf{x}_t - \mathbf{y}_t\boldsymbol{\theta})(\mathbf{x}_t - \mathbf{y}_t\boldsymbol{\theta})^\top$, $\widetilde{\mathbf{K}} = \mathbf{C} \otimes \mathbf{K}$, and $\mathbf{C}$ is an $(S+D)\times(S+D)$ square matrix with all entries being 0 except the lower right $D \times D$ block being $\mathbf{I}_D$. We use $\boldsymbol{\theta}^*, \boldsymbol{\Omega}^*$ to denote the ground truth of $\boldsymbol{\theta}, \boldsymbol{\Omega}$, which lie in the space denoted by $\mathbb{F}_{\boldsymbol{\theta}}, \mathbb{F}_{\boldsymbol{\Omega}}$, respectively. 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 scalar $c > 0$:

$$\mathrm{P}\left(\inf_{\|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\mathrm{F}} \geq c} \inf_{\bar{\boldsymbol{\theta}}} h(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\Omega}}) \leq h(\boldsymbol{\theta}^*, \boldsymbol{\Omega}^*)\right) \to 0, \text{ as } T \to \infty. \tag{38}$$

This is because if (38) is established, then as $T \to \infty$ we have:

$$\mathrm{P}\left(\inf_{\|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\mathrm{F}} \geq c} \inf_{\bar{\boldsymbol{\theta}} \in \mathbb{F}_{\boldsymbol{\theta}}} h(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\Omega}}) \geq \inf_{\|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\mathrm{F}} \geq c} \inf_{\bar{\boldsymbol{\theta}}} h(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\Omega}}) > h(\boldsymbol{\theta}^*, \boldsymbol{\Omega}^*) \geq h(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Omega}})\right) \to 1,$$

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 (38), 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:

$$\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) = \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t}{T} + \lambda \widetilde{\mathbf{K}} \right)^{-1} \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{x}_t}{T} \right), \tag{39}$$

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:

$$\widetilde{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) = (\mathbf{I} - \lambda \widetilde{\mathbf{K}}) \boldsymbol{\theta}^* + \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t}{T} + \lambda \widetilde{\mathbf{K}} \right)^{-1} \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{e}_t}{T} \right), \tag{40}$$

and the first term converges to $\boldsymbol{\theta}^*$ since $\lambda = o(1)$. The second term in (40) converges to 0 since $\bar{\boldsymbol{\Omega}}^{1/2} \mathbf{y}_t$ is covariance stationary and independent of $\bar{\boldsymbol{\Omega}}^{1/2} \mathbf{e}_t$.

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

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

To prove (38), it suffices to show that:

$$\mathrm{P} \left(\inf_{\|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\mathrm{F}} \geq c} \ell(\bar{\boldsymbol{\Omega}}) \leq \ell(\boldsymbol{\Omega}^*) \right) \to 0, \text{ as } T \to \infty, \tag{41}$$

since $\ell(\boldsymbol{\Omega}^*) \leq h(\boldsymbol{\theta}^*, \boldsymbol{\Omega}^*)$. Given that $\widetilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^*$, we can rewrite $\ell(\bar{\boldsymbol{\Omega}})$ as:

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

By Theorem 4 of Chen et al. (2021), for any $\bar{\boldsymbol{\Omega}}$ with $\|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\mathrm{F}} \geq c$:

$$\mathrm{P} \left[\widetilde{\ell}(\bar{\boldsymbol{\Omega}}) \geq \widetilde{\ell}(\widetilde{\boldsymbol{\Omega}}) + \frac{c^2}{32} \underline{\rho}((\boldsymbol{\Omega}^*)^{-1})^2 \right] \to 1, \text{ as } T \to \infty,$$

where $\widetilde{\boldsymbol{\Omega}} = \arg\min_{\boldsymbol{\Omega}} \widetilde{\ell}(\boldsymbol{\Omega})$, or simply $\widetilde{\boldsymbol{\Omega}} = (\sum_t \mathbf{e}_t \mathbf{x}_t^\top / T)^{-1}$ which satisfies $\widetilde{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}^*$ and thus $\widetilde{\ell}(\widetilde{\boldsymbol{\Omega}}) \xrightarrow{p} \widetilde{\ell}(\boldsymbol{\Omega}^*)$. Using the fact that $\ell(\bar{\boldsymbol{\Omega}}) = \widetilde{\ell}(\bar{\boldsymbol{\Omega}}) + o_P(1)$, eventually we have:

$$\mathrm{P} \left(\inf_{\|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_{\mathrm{F}} \geq c} \ell(\bar{\boldsymbol{\Omega}}) \geq \ell(\boldsymbol{\Omega}^*) + \frac{c^2}{32} \underline{\rho}((\boldsymbol{\Omega}^*)^{-1})^2 \right) \to 1, \text{ as } T \to \infty,$$

which proves (41) and thus completes the proof. ■

A.4 Proof of Theorem 9

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

Lemma 14 *Under the same assumption as Theorem 9, all model estimators for MARAC are consistent and satisfy:*

$$\widehat{\mathbf{A}}_p = \mathbf{A}_p^* + O_P(T^{-\frac{1}{2}}), \quad \widehat{\mathbf{B}}_p = \mathbf{B}_p^* + O_P(T^{-\frac{1}{2}}), \quad \widehat{\boldsymbol{\gamma}}_q = \boldsymbol{\gamma}_q^* + O_P(T^{-\frac{1}{2}}),$$

for $p \in [P], q \in [Q]$, where the convergence in probability is elementwise.

We relegate the proof of Lemma 14 to Appendix B.1. Now we are ready to present the proof of Theorem 9.

Proof For simplicity, we fix P, Q as 1 but the proving technique can be generalized to arbitrary P, Q. We revisit the updating rule for $\mathbf{A}_p^{(l+1)}$ in (15). By plugging in the data-generating model for $\mathbf{X}_t$ according to MARAC$(1,1)$ model, we can transform (15) into:

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

where $\Delta\mathbf{M} = \widehat{\mathbf{M}} - \mathbf{M}^*$, $\mathbf{M}$ is an arbitrary matrix. By using the result in Proposition 8 and Lemma 14, one can simplify the equation above by left multiplying $\widehat{\boldsymbol{\Sigma}}_r^{-1}$ and vectorize both sides to obtain:

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

Similar transformations can be applied to (16) and (18), 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 into:

$$\left(\frac{1}{T} \sum_{t \in [T]} \mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t \right) \mathbf{vec}\left(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \right) = \frac{1}{T} \sum_{t \in [T]} \mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{vec}\left(\mathbf{E}_t \right) + o_P(T^{-1/2}), \quad (42)$$

where $\mathbf{vec}\left(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \right) = [\mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}^* \right)^\top, \mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^* \right)^\top, \mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{R}}} - \boldsymbol{\mathcal{R}}^* \right)^\top]^\top$.

In (42), we first establish that:

$$(1/T) \sum_{t \in [T]} \mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t \xrightarrow{p.} \mathrm{E}\left[\mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t \right]. \quad (43)$$

By the assumption that $\mathbf{X}_t$ and $\mathbf{z}_t$ are jointly stationary, we have $(1/T) \sum_{t \in [T]} \widetilde{\mathbf{x}}_t \widetilde{\mathbf{x}}_t^\top \xrightarrow{p.} \mathrm{E}[\widetilde{\mathbf{x}}_t \widetilde{\mathbf{x}}_t^\top]$ using lemma 21, where $\widetilde{\mathbf{x}}_t = [\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top$. Then since each element of $\mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t$ is a linear combination of terms in $\widetilde{\mathbf{x}}_t \widetilde{\mathbf{x}}_t^\top$, it is straightforward that (43) holds elementwise.

For the term on the right hand side of (42), 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 $\mathbf{z}_{t-1}$. We will find the limiting distribution of $(1/T) \sum_{t \in [T]} \boldsymbol{\eta}_t$ via the central limit theorem (CLT) of MDS (proposition 7.9 of Hamilton (2020)). To verify the regularity condition of the CLT of MDS, firstly, we verify that $(1/T) \sum_{t=1}^T \boldsymbol{\eta}_t \boldsymbol{\eta}_t^\top \xrightarrow{p.} \mathrm{E}[\boldsymbol{\eta}_t \boldsymbol{\eta}_t^\top]$, which can be proved via a two-step procedure. One first fixes all $\mathbf{W}_t$, then $(1/T) \sum_{t=1}^T \boldsymbol{\eta}_t \boldsymbol{\eta}_t^\top \xrightarrow{p.} (1/T) \sum_{t \in [T]} \mathbf{W}_t^\top (\mathbf{\Sigma}^*)^{-1} \mathbf{W}_t$ and finishes the proof using (43). The other regularity condition to verify is that the fourth moment of $\boldsymbol{\eta}_t$ is bounded. Since we assume that $\mathbf{x}_t$ and $\mathbf{z}_t$ are jointly stationary and the error term for the vector autoregressive process (27) is Gaussian, it is evident that the distribution of $\widetilde{\mathbf{x}}_t = [\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top$ is Gaussian and any fourth moment of $\boldsymbol{\eta}_t$ is a linear combination of fourth moment of $\widetilde{\mathbf{x}}_t$, which is thus bounded.

By the CLT of MDS and (43), we have:

$$\mathrm{E}\left[\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t\right]\mathbf{vec}\left(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\right)\xrightarrow{d.}\mathcal{N}(\mathbf{0},\mathrm{E}\left[\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t\right]). \tag{44}$$

However, the matrix $\mathbf{L}=\mathrm{E}\left[\mathbf{W}_t^\top(\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t\right]$ is not a full-rank matrix, because $\mathbf{L}\boldsymbol{\mu}=\mathbf{0}$, where $\boldsymbol{\mu}=[\mathbf{vec}\left(\boldsymbol{\mathcal{A}}^*\right)^\top,-\mathbf{vec}\left(\boldsymbol{\mathcal{B}}^*\right)^\top,\mathbf{0}^\top]^\top$. Given the identifiability constraint that $\|\mathbf{A}_1^*\|_{\mathrm{F}}=\|\widehat{\mathbf{A}}_1\|_{\mathrm{F}}=1$ and the rate of convergence of $\widehat{\mathbf{A}}_1$ being $O_P(T^{-1/2})$, we have $\langle\mathbf{A}_1^*,\widehat{\mathbf{A}}_1-\mathbf{A}_1^*\rangle=o_P(T^{-1/2})$. Now, by setting $\boldsymbol{\eta}=[\mathbf{vec}\left(\mathbf{A}_1^*\right)^\top\mathbf{0}^\top]^\top$, we have:

$$\boldsymbol{\eta}^\top\mathbf{vec}\left(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\right)=o_P(T^{-1/2})\xrightarrow{p.}0. \tag{45}$$

Combining (44) and (45) and using the Slutsky's theorem, we have $\mathbf{Hvec}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*)\xrightarrow{d.}\mathcal{N}(\mathbf{0},\mathbf{L})$, where $\mathbf{H}=\mathbf{L}+\boldsymbol{\eta}\boldsymbol{\eta}^\top$. This completes the proof. A similar technique can be found in the proof of Theorem 4 of Chen et al. (2021). ■

A.5 Proof of Theorem 12

Proof In this proof, we will fix P,Q as 1 again for the ease of notations but the technical details can be readily 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:

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

and we introduce the following additional notations:

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

and we will drop the subscript T later 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 evaluate $g_1^*, \ldots, g_D^*$ on a matrix grid, i.e. $\mathbf{K}\boldsymbol{\gamma}_d^*$ is a discrete evaluation of g_d^* on the matrix grid. Let $\mathbb{F}_{\boldsymbol{\phi}} = \{\mathbf{vec}\,(\mathbf{B} \otimes \mathbf{A}) \,|\, \|\mathbf{A}\|_{\mathrm{F}} = \mathrm{sign}(\mathrm{tr}\,(\mathbf{A})) = 1, \mathbf{A} \in \mathbb{R}^{M\times M}, \mathbf{B} \in \mathbb{R}^{N\times N}\}$. Recall that we use $S = MN$ to denote the matrix dimensionality. With these new notations, the MARAC estimator is obtained via solving the following penalized least square problem:

$$\min_{\boldsymbol{\phi}\in\mathbb{F}_{\boldsymbol{\phi}}, \boldsymbol{\gamma}\in\mathbb{R}^{SD}} \mathfrak{L}_\lambda(\boldsymbol{\phi}, \boldsymbol{\gamma}) \coloneqq \left\{ \frac{1}{2T}\|\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\phi} - (\widetilde{\mathbf{z}} \otimes \mathbf{K})\,\boldsymbol{\gamma}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\boldsymbol{\gamma}^\top (\mathbf{I}_D \otimes \mathbf{K})\,\boldsymbol{\gamma} \right\}. \tag{46}$$

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

$$\widehat{\boldsymbol{\gamma}}(\boldsymbol{\phi}) = T^{-1}\left[\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}} \otimes \mathbf{K} + \lambda \cdot \mathbf{I}_{SD}\right]^{-1} \left(\widetilde{\mathbf{z}}^\top \otimes \mathbf{I}_S\right) \left(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\phi}\right). \tag{47}$$

Simply plugging (47) into (46) yields the profile likelihood for $\boldsymbol{\phi}$:

$$\ell_\lambda(\boldsymbol{\phi}) = \mathfrak{L}_\lambda(\boldsymbol{\phi}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\phi})) = \frac{1}{2T}\left(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\phi}\right)^\top \mathbf{W} \left(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\phi}\right), \tag{48}$$

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

$$\mathbf{W} = \left\{ \mathbf{I} - \frac{(\widetilde{\mathbf{z}} \otimes \mathbf{K})\left[\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}} \otimes \mathbf{K} + \lambda \cdot \mathbf{I}_{SD}\right]^{-1} \left(\widetilde{\mathbf{z}}^\top \otimes \mathbf{I}_S\right)}{T} \right\} = \left(\mathbf{I} + \frac{\widetilde{\mathbf{z}}\widetilde{\mathbf{z}}^\top}{\lambda T} \otimes \mathbf{K}\right)^{-1}, \tag{49}$$

and the second equality in (49) is by the Woodbury matrix identity. To improve the clarity and organization of the proof, we break down the results into two separate claims and prove them subsequently. Our first claim is about $\widehat{\boldsymbol{\phi}}$:

Proposition 15 *Under the assumptions of Theorem 12, we have:*

$$\left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\right)^\top \left(\frac{\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}}{T}\right) \left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\right) \leq O_P(C_g\lambda) + O_P(SD/T), \tag{50}$$

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:*

$$\underline{\rho}\left(\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}/T\right) \geq \frac{c_{0,S}}{2} > 0, \tag{51}$$

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

The proof of proposition 15 and lemma 16 are relegated to Appendix A.6 and B.3, respectively. Combining proposition (15) and lemma 16, we have:

$$\frac{1}{S}\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\|_{\mathrm{F}} \leq O_P(\sqrt{\frac{C_g \gamma_S}{c_{0,S} S}}) + O_P(\sqrt{\frac{D}{c_{0,S} T S}}). \tag{52}$$

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

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

and we will bound terms J_1, J_2, J_3 respectively.

For J_1, we have the following lemma:

Lemma 17 *Given the definition of* $\mathbf{W}$ *in* (49) *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* $tr(\mathbf{W}) \leq SD$*.*

We leave the proof to Appendix B.2. Notice that $J_1^2 = \sigma^2 \cdot O_P(\mathrm{E}[\mathrm{tr}\left((\mathbf{I} - \mathbf{W})^2\right)]) \leq \sigma^2 \cdot O_P(\mathrm{E}[\mathrm{tr}\left((\mathbf{I} - \mathbf{W})\right)])$, thus we have $J_1 \leq O_P(S^{1/4}\gamma_S^{-1/4r_0})$.

For J_2, we have the following bound:

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

To bound $\|\mathbf{W}^{-1/2}\|_s$, we can take advantage of the simpler form of $\mathbf{W}$ using the Woodbury matrix identity in (49) and have:

$$\|\mathbf{W}^{-1/2}\|_s = \sqrt{\bar{\rho}(\mathbf{W}^{-1})} = \sqrt{\bar{\rho}\left(\mathbf{I}+(\widetilde{\mathbf{z}}\widetilde{\mathbf{z}}^\top/\lambda T)\otimes\mathbf{K}\right)}$$
$$\leq \sqrt{1+\lambda^{-1}\bar{\rho}(\mathbf{K})\bar{\rho}(\widetilde{\mathbf{z}}\widetilde{\mathbf{z}}^\top/T)} \leq \sqrt{1+\lambda^{-1}\bar{\rho}(\mathbf{K})\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}})}.$$

In lemma 21, which we state later in Appendix B, we have shown that for N-dimensional stationary vector autoregressive process, the covariance estimator is consistent in spectral norm as long as $N\log N/T \to 0$. Therefore, since $\mathbf{z}_t$ follows a stationary VAR($\widetilde{Q}$) process and its dimensionality is fixed, we have $\|\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}-\boldsymbol{\Sigma}^*_{\mathbf{z}}\|_s \xrightarrow{p} 0$ and thus with high probability, 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 $\boldsymbol{\Sigma}^*_{\mathbf{z}}$ and $\bar{\rho}(\mathbf{K})$. Together with the result in proposition 15, we can bound J_2 as:

$$J_2 \leq O_P(\sqrt{C_g\lambda T + c_0C_gT + SD + c_0D\gamma_S^{-1}}). \tag{53}$$

Finally, for J_3, the bound of I_1 in (55) can be used directly here since $J_3 \leq \bar{\rho}(\mathbf{W}^{1/2})\sqrt{I_1} \leq \sqrt{I_1}$ and thus $J_3 \leq O_P(\sqrt{\lambda T})$. Combining all the results together we have:

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

■

A.6 Proof of Proposition 15

Proof The MARAC estimator is the global minimizer of $\ell_\lambda(\boldsymbol{\phi})$ within $\mathbb{F}_\phi$ and thus $\ell_\lambda(\widehat{\boldsymbol{\phi}}) \leq \ell_\lambda(\boldsymbol{\phi}^*)$, which is equivalent to:

$$\frac{1}{2}\left(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*\right)^\top\left(\frac{\widetilde{\mathbf{X}}^\top\mathbf{W}\widetilde{\mathbf{X}}}{T}\right)\left(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*\right)-\frac{1}{T}\left[\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^*+\boldsymbol{\mathcal{E}}\right]^\top\mathbf{W}\widetilde{\mathbf{X}}\left(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*\right)\leq 0.$$

If we let $\boldsymbol{\delta}=\mathbf{W}^{1/2}\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*)/\sqrt{T}$ and $\boldsymbol{\omega}=\mathbf{W}^{1/2}\left[\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^*+\boldsymbol{\mathcal{E}}\right]/\sqrt{T}$, then (54) can be written as $\boldsymbol{\delta}^\top\boldsymbol{\delta}\leq 2\boldsymbol{\delta}^\top\boldsymbol{\omega}$, and we can upper bound $\boldsymbol{\delta}^\top\boldsymbol{\delta}$, the quantity of our interest, 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 $\|\boldsymbol{\omega}\|_{\mathrm{F}}^2$:

$$\begin{aligned}\|\boldsymbol{\delta}\|_{\mathrm{F}}^2\leq 4\|\boldsymbol{\omega}\|_{\mathrm{F}}^2&=\frac{4}{T}\left[\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^*+\boldsymbol{\mathcal{E}}\right]^\top\mathbf{W}\left[\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^*+\boldsymbol{\mathcal{E}}\right]\\&\leq\frac{8}{T}\left[\underbrace{\|\mathbf{W}^{1/2}\left(\widetilde{\mathbf{z}}\otimes\mathbf{K}\right)\boldsymbol{\gamma}^*\|_{\mathrm{F}}^2}_{I_1}+\underbrace{\|\mathbf{W}^{1/2}\boldsymbol{\mathcal{E}}\|_{\mathrm{F}}^2}_{I_2}\right].\end{aligned}\tag{54}$$

For I_1, it can be bounded as:

$$\begin{aligned}I_1&=(\lambda T)[(\mathbf{I}_D\otimes\mathbf{K})\boldsymbol{\gamma}^*]^\top\{\mathbf{I}_{SD}-\left(\lambda^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\otimes\mathbf{K}+\mathbf{I}_{SD}\right)^{-1}\}\boldsymbol{\gamma}^*\\&=(\lambda T)\left(\sum_{d=1}^D\|g_d^*\|_{\mathbb{H}_k}^2\right)-(\lambda^2T)\left(\boldsymbol{\gamma}^*\right)^\top\left[\left(\mathbf{I}_D\otimes\mathbf{K}\right)\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\otimes\mathbf{K}+\lambda\mathbf{I}_{SD}\right)^{-1}\right]\boldsymbol{\gamma}^*\leq C_g\lambda T,\end{aligned}\tag{55}$$

where $C_g=\sum_{d=1}^D\|g_d^*\|_{\mathbb{H}_k}^2$ is a constant relates to the smoothness of the functional parameters. The last inequality of (55) follows from the fact that the matrix in the term led by λ^2T is positive semi-definite. To see why, first note that:

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

Then, we have the following lemma:

Lemma 18 *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 Appendix B.5. Let $\mathbf{M} = \widehat{\mathbf{\Sigma}}_{\mathbf{z}} \otimes \mathbf{I}_S + \lambda^{-1} \widehat{\mathbf{\Sigma}}_{\mathbf{z}}^2 \otimes \mathbf{K}$ and $\mathbf{N} = \widehat{\mathbf{\Sigma}}_{\mathbf{z}} \otimes \mathbf{I}_S$, then both $\mathbf{M}$ and $\mathbf{N}$ are positive definite and $\mathbf{M} - \mathbf{N}$ is positive semi-definite. By Lemma 18, we have $\mathbf{N}^{-1} - \mathbf{M}^{-1}$ being positive semi-definite and thus (55) holds.

To bound I_2, we first prove the following lemma:

Lemma 19 *Suppose that* $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)$ *is a* d*-dimensional Gaussian random vector with* $\xi_1, \ldots, \xi_d \overset{i.i.d.}{\sim} N(0, \sigma^2)$*, and* $\mathbf{A} \in \mathbb{R}^{d \times d}$ *is a positive semi-definite matrix with* $\|\mathbf{A}\|_s \leq 1, tr(\mathbf{A}) \to \infty$ *as* $d \to \infty$*, then:*

$$\boldsymbol{\xi}^\top \mathbf{A} \boldsymbol{\xi} = O_P(\mathrm{E}[\boldsymbol{\xi}^\top \mathbf{A} \boldsymbol{\xi}]), \quad \textit{and} \quad \boldsymbol{\xi}^\top (\mathbf{I}_d - \mathbf{A}) \boldsymbol{\xi} = O_P(\mathrm{E}[\boldsymbol{\xi}^\top (\mathbf{I}_d - \mathbf{A}) \boldsymbol{\xi}]).$$

We leave the corresponding proof to Appendix B.4. For this lemma, we make one additional remark:

Remark 20 *In Lemma 19, if* $\mathbf{A}$ *is also random and independent of* $\boldsymbol{\xi}$ *and has* $tr(\mathbf{A}) = O_P(c_d)$ *for some sequence* $c_d \to \infty$ *as* $d \to \infty$*, then results in Lemma 19 can be stated as* $\boldsymbol{\xi}^\top \mathbf{A} \boldsymbol{\xi} = O_P(c_d), \boldsymbol{\xi}^\top (\mathbf{I}_d - \mathbf{A}) \boldsymbol{\xi} = O_P(d - c_d)$.

Using lemma 17 and the assumption that $\gamma_S \to 0$, we have $\mathrm{tr}\,(\mathbf{I} - \mathbf{W}) \xrightarrow{p} \infty$. By remark 20, we can bound I_2 as $I_2 = O_P(\mathrm{E}[\boldsymbol{\mathcal{E}}^\top \mathbf{W} \boldsymbol{\mathcal{E}}]) = O_P(\mathrm{tr}\,(\mathbf{W})) = O_P(SD)$. Combining the bounds for I_1 and I_2, we have:

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

which completes the proof. ■

B Statement and Proof of Technical Lemmas

The first lemma extends proposition 6 of Li and Xiao (2021) and covers a general result of the consistency of covariance estimator of stationary vector autoregressive processes:

Lemma 21 *Let* $\mathbf{x}_t \in \mathbb{R}^N$ *be generated by a zero-meaned stationary AR(p) process:* $\mathbf{x}_t = \sum_{l=1}^{p} \mathbf{\Phi}_p \mathbf{x}_{t-p} + \boldsymbol{\xi}_t$*, where* $\boldsymbol{\xi}_t$ *have i.i.d. sub-Gaussian entries. With* $\widehat{\mathbf{\Sigma}} = (1/T)\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t^\top$ *and* $\mathbf{\Sigma} = \mathrm{E}[\widehat{\mathbf{\Sigma}}]$*, we have:*

$$\mathrm{E}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_s \leq C\left(\sqrt{\frac{pN\log pN}{T}} + \frac{pN\log pN}{T}\right)\|\mathbf{\Sigma}_{\mathbf{y}}\|_s, \tag{56}$$

where $\mathbf{\Sigma}_{\mathbf{y}} = \mathrm{E}[\mathbf{y}_t\mathbf{y}_t^\top]$*, and* $\mathbf{y}_t = [\mathbf{x}_t^\top, \ldots, \mathbf{x}_{t-p+1}^\top]^\top \in \mathbb{R}^{pN}$ *and* C *is an absolute constant.*

Proof By proposition 6 of Li and Xiao (2021), we have (56) hold for AR(1) process. For AR(p) process, we first convert it to an AR(1) process via: $\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{u}_t$, where $\mathbf{u}_t = [\boldsymbol{\xi}_t^\top, \mathbf{0}^\top]^\top$. Therefore, we have (56) hold for $\mathbf{\Sigma}_{\mathbf{y}} = \mathrm{E}[\mathbf{y}_t\mathbf{y}_t^\top]$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}$. Notice that $\mathbf{\Sigma} = \mathrm{E}[\mathbf{x}_t\mathbf{x}_t^\top]$ is the top-left block of $\mathbf{\Sigma}_{\mathbf{y}}$, thus $\mathbf{\Sigma} = \mathbf{E}\mathbf{\Sigma}_{\mathbf{y}}\mathbf{E}^\top$ with $\mathbf{E} = [\mathbf{I}; \mathbf{O}]$. Eventually, we have:

$$\mathrm{E}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_s = \mathrm{E}\|\mathbf{E}(\widehat{\mathbf{\Sigma}}_{\mathbf{y}} - \mathbf{\Sigma}_{\mathbf{y}})\mathbf{E}^\top\|_s \leq \mathrm{E}\|\widehat{\mathbf{\Sigma}}_{\mathbf{y}} - \mathbf{\Sigma}_{\mathbf{y}}\|_s,$$

and thus for AR(p) process $\mathbf{x}_t$, we have (56) hold. ■

We have the following corollary of Lemma 21 under the MARAC model:

Corollary 22 *Let* $\mathbf{X}_t$ *be generated by a MARAC*(P, Q) *model following* (1) *and* $\mathbf{z}_t$ *satisfies assumption 4 and* $\mathbf{X}_t$ *and* $\mathbf{z}_t$ *are jointly stationary in the sense of Theorem 6. Further assume that the noise process* $\mathbf{E}_t$ *follows* (4)*. Then for* $\widetilde{\mathbf{x}}_t = [\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top$ *and* $\mathbf{\Sigma} = \mathrm{E}[\widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^\top]$*, we have:*

$$\mathrm{E}\|(1/T)\sum_{t=1}^{T}\widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^\top - \mathbf{\Sigma}\|_s \leq C\left(\sqrt{\frac{L(S+D)\log(S+D)}{T}} + \frac{L(S+D)\log(S+D)}{T}\right)\|\mathbf{\Sigma}_{\mathbf{y}}\|_s, \tag{57}$$

where $L = \max(P, Q, \widetilde{Q})$ *and* $\mathbf{\Sigma}_{\mathbf{y}} = \mathrm{E}[\mathbf{y}_t \mathbf{y}_t^\top]$ *with* $\mathbf{y}_t = [\widetilde{\mathbf{x}}_t^\top, \ldots, \widetilde{\mathbf{x}}_{t-L+1}^\top]^\top$. *The result implies that as long as* $S \log S/T \to 0$, *the auto-covariance estimator for* $\widetilde{\mathbf{x}}_t$ *is consistent.*

B.1 Proof of Lemma 14

Proof Without loss of generality, we fix P, Q as 1 and use the same notation as (36) in Appendix A.3, so the MARAC model can be written as $\mathbf{x}_t = \mathbf{y}_t \boldsymbol{\theta}^* + \mathbf{e}_t$. Correspondingly, the penalized log-likelihood $h(\boldsymbol{\theta}, \mathbf{\Omega})$ is specified by (37) and given any $\bar{\mathbf{\Omega}}$, we have $\widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}) = \arg\min_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \bar{\mathbf{\Omega}})$ as specified by (39). Given the decomposition of $\widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}})$ in (40), we have:

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

where the first term is $o(T^{-1/2})$ since $\lambda = o(T^{-1/2})$ and the second term is $O_P(T^{-1/2})$ since:

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

due to the joint covariance stationarity assumption of $\mathbf{X}_t$ and $\mathbf{z}_t$ and $\mathbf{y}_t^\top \bar{\mathbf{\Omega}} \mathbf{e}_t$ is a martingale difference sequence such that $\sum_t \mathbf{y}_t^\top \bar{\mathbf{\Omega}} \mathbf{e}_t / T = O_P(T^{-1/2})$ (see proposition 7.9 of Hamilton (2020) for the central limit theorem of martingale difference sequence). It can thus be concluded that $\widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}) = \boldsymbol{\theta}^* + O_P(T^{-1/2})$.

Fix $\mathbf{\Omega} = \bar{\mathbf{\Omega}}$, we can decompose $h(\boldsymbol{\theta}, \bar{\mathbf{\Omega}})$ as follows:

$$\begin{aligned} h(\boldsymbol{\theta}, \bar{\mathbf{\Omega}}) &= h(\widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}), \bar{\mathbf{\Omega}}) + \frac{1}{2} (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}))^\top \left(\frac{\sum_t \mathbf{y}_t^\top \bar{\mathbf{\Omega}} \mathbf{y}_t}{T} + \lambda \widetilde{\mathbf{K}} \right) (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}})) \\ &\geq h(\widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}), \bar{\mathbf{\Omega}}) + \frac{1}{2} \mathrm{tr} \left\{ \left[(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}})) \otimes \mathbf{I}_p \right] \bar{\mathbf{\Omega}} \left[(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}))^\top \otimes \mathbf{I}_p \right] \widehat{\mathbf{\Gamma}} \right\} \\ &\geq h(\widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}), \bar{\mathbf{\Omega}}) + \frac{p}{2} \underline{\rho}(\bar{\mathbf{\Omega}}) \underline{\rho}(\widehat{\mathbf{\Gamma}}) \| \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}) \|_{\mathrm{F}}^2, \end{aligned} \tag{58}$$

where $\widehat{\mathbf{\Gamma}} = T^{-1} \sum_t \mathbf{vec}(\mathbf{y}_t) \mathbf{vec}(\mathbf{y}_t)^\top$ and $p = S(S + D)$. The term with $\lambda \widetilde{\mathbf{K}}$ is dropped in the first inequality since $\widetilde{\mathbf{K}}$ is positive semi-definite.

To further obtain a lower bound for (58), we first bound $\underline{\rho}(\bar{\mathbf{\Omega}})$. Set $C_1 = \underline{\rho}(\mathbf{\Omega}^*)/2$, then for any $\bar{\mathbf{\Omega}}$ such that $\|\bar{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{\mathrm{F}} \leq C_1$ we have:

$$|\underline{\rho}(\bar{\mathbf{\Omega}}) - \underline{\rho}(\mathbf{\Omega}^*)| \leq \|\bar{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_s \leq \|\bar{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{\mathrm{F}} \leq \underline{\rho}(\mathbf{\Omega}^*)/2,$$

and thus $\underline{\rho}(\bar{\mathbf{\Omega}}) \geq \underline{\rho}(\mathbf{\Omega}^*)/2$.

Next, we bound $\underline{\rho}(\widehat{\mathbf{\Gamma}})$. We can rewrite $\widehat{\mathbf{\Gamma}}$ as $\mathbf{C}(\widehat{\mathbf{\Gamma}}_0 \otimes \mathbf{I}_S)\mathbf{C}$, with $\mathbf{C} = \mathrm{diag}(\mathbf{I}_{S^2}, \mathbf{I}_D \otimes \mathbf{K})$[1] and $\widehat{\mathbf{\Gamma}}_0 = T^{-1}\sum_t [\mathbf{x}_{t-1}^\top \mathbf{z}_{t-1}^\top]^\top [\mathbf{x}_{t-1}^\top \mathbf{z}_{t-1}^\top]$, which is a consistent estimator of its expectation $\mathbf{\Gamma}_0$ by Corollary 22. Using these facts, we have the following bound:

$$\left|\underline{\rho}(\widehat{\mathbf{\Gamma}}) - \underline{\rho}(\mathbf{C}(\mathbf{\Gamma}_0 \otimes \mathbf{I}_S)\mathbf{C})\right| \leq \|\mathbf{C}(\widehat{\mathbf{\Gamma}}_0 \otimes \mathbf{I}_S)\mathbf{C} - \mathbf{C}(\mathbf{\Gamma}_0 \otimes \mathbf{I}_S)\mathbf{C}\|_s \leq \|\mathbf{C}\|_s^2 \|\widehat{\mathbf{\Gamma}}_0 - \mathbf{\Gamma}_0\|_s, \quad (59)$$

and by (57), we can properly choose a T that is sufficiently large such that (59) is upper bounded by $\frac{1}{2}\underline{\rho}(\mathbf{\Gamma}_0)\underline{\rho}(\mathbf{C})^2$ with high probability. Consequently, we have:

$$\underline{\rho}(\widehat{\mathbf{\Gamma}}) \geq \underline{\rho}(\mathbf{C}(\mathbf{\Gamma}_0 \otimes \mathbf{I}_S)\mathbf{C}) - \frac{1}{2}\underline{\rho}(\mathbf{\Gamma}_0)\underline{\rho}(\mathbf{C})^2 \geq \frac{1}{2}\underline{\rho}(\mathbf{\Gamma}_0)\underline{\rho}(\mathbf{C})^2 \quad (60)$$

holds with high probability and given the assumption that $\underline{\rho}(\mathbf{K}) = \underline{c} > 0$, we have $\underline{\rho}(\mathbf{C}) > 0$. With the lower bounds on $\underline{\rho}(\bar{\mathbf{\Omega}})$ and $\underline{\rho}(\widehat{\mathbf{\Gamma}})$, we can claim that with high probability:

$$\inf_{\bar{\mathbf{\Omega}} \in \mathbb{F}_{\mathbf{\Omega}}: \|\bar{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{\mathrm{F}} \leq C_1} h(\boldsymbol{\theta}, \bar{\mathbf{\Omega}}) \geq \inf_{\bar{\mathbf{\Omega}} \in \mathbb{F}_{\mathbf{\Omega}}: \|\bar{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_{\mathrm{F}} \leq C_1} \left\{ h(\widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}), \bar{\mathbf{\Omega}}) + C_2 \cdot \|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}})\|_{\mathrm{F}}^2 \right\}, \quad (61)$$

where $C_2 = \frac{p}{8}\underline{\rho}(\mathbf{\Omega}^*)\underline{\rho}(\mathbf{\Gamma}_0)\underline{\rho}(\mathbf{C})^2$.

Previously, we have shown that $\widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}) = \boldsymbol{\theta}^* + O_P(T^{-1/2})$ for any arbitrary $\bar{\mathbf{\Omega}}$. By the Taylor expansion in (58), we can conclude that $h(\boldsymbol{\theta}^*, \bar{\mathbf{\Omega}}) = h(\widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}), \bar{\mathbf{\Omega}}) + O_P(T^{-1})$. For the second term on the RHS in (61), we consider those $\boldsymbol{\theta}$ such that $\sqrt{T}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathrm{F}} \geq c_T$ where $c_T \to \infty$ is an arbitrary sequence, then given that $\widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}}) = \boldsymbol{\theta}^* + O_P(T^{-1/2})$, we have:

$$\mathrm{P}\left(C_2 \cdot \|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}(\bar{\mathbf{\Omega}})\|_{\mathrm{F}}^2 \geq \frac{C_2}{2} \cdot \frac{c_T^2}{T} \right) \to 1, \text{ as } T \to \infty. \quad (62)$$

[1] $\mathrm{diag}(\cdot)$ concatenates a sequence of matrices into a block-diagonal matrix.

Coupling with these results, we can conclude that:

$$\mathrm{P}\left(\inf_{\sqrt{T}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_{\mathrm{F}}\geq c_T}\inf_{\bar{\boldsymbol{\Omega}}\in\mathbb{F}_{\boldsymbol{\Omega}}:\|\bar{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^*\|_{\mathrm{F}}\leq C_1} h(\boldsymbol{\theta},\bar{\boldsymbol{\Omega}}) > \inf_{\bar{\boldsymbol{\Omega}}\in\mathbb{F}_{\boldsymbol{\Omega}}:\|\bar{\boldsymbol{\Omega}}-\boldsymbol{\Omega}^*\|_{\mathrm{F}}\leq C_1} h(\boldsymbol{\theta}^*,\bar{\boldsymbol{\Omega}})\right)\to 1. \tag{63}$$

The result in (63) indicates that for any $\boldsymbol{\theta}$ that lies outside of the set $\{\boldsymbol{\theta} : \sqrt{T}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\mathrm{F}} \geq c_T\}$, the penalized log-likelihood is no smaller than a sub-optimal solution with high probability. Therefore, with high probability, 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 $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + O_P(T^{-1/2})$ 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 $T^{-1/2}$.

■

B.2 Proof of Lemma 17

Proof By the definition of $\mathbf{W}$ in (49), we have:

$$\begin{aligned}\mathrm{tr}\left(\mathbf{I}-\mathbf{W}\right) &= \mathrm{tr}\left[\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\otimes\mathbf{K}+\lambda\mathbf{I}_{SD}\right)^{-1}\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\otimes\mathbf{K}\right)\right]\\ &= \sum_{s=1}^{S}\sum_{d=1}^{D}\frac{\rho_d(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}})\rho_s(\mathbf{K})}{\lambda+\rho_d(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}})\rho_s(\mathbf{K})} \leq D\cdot\sum_{s=1}^{S}\frac{1}{1+\lambda\bar{\rho}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}})^{-1}\rho_s(\mathbf{K})^{-1}}.\end{aligned} \tag{64}$$

Using lemma 21, we can bound $\bar{\rho}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}})$ by $2\bar{\rho}(\boldsymbol{\Sigma}_{\mathbf{z}})$ with high probability. Conditioning on this 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 (64) can be bounded as:

$$D\cdot\sum_{s=1}^{S}\frac{1}{1+\lambda\bar{\rho}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}})^{-1}\rho_s(\mathbf{K})^{-1}} \leq D\cdot\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{1}{1+c_{\mathbf{z}}\lambda\rho_i(\mathbf{K}_1)^{-1}\rho_j(\mathbf{K}_2)^{-1}}, \tag{65}$$

where $c_{\mathbf{z}} = \bar{\rho}(\boldsymbol{\Sigma}_{\mathbf{z}})/2$. As $M, N\to\infty$, based on assumption 10, we have $\rho_i(\mathbf{K}_1)\to Mi^{-r_0}$ and $\rho_j(\mathbf{K}_2)\to Nj^{-r_0}$. Therefore, we can find two constants $0<c_1<c_2$, such that:

$$\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{1}{1+c_2\lambda(ij)^{r_0}/S} \leq \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{1}{1+c_{\mathbf{z}}\lambda\rho_i(\mathbf{K}_1)^{-1}\rho_j(\mathbf{K}_2)^{-1}} \leq \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{1}{1+c_1\lambda(ij)^{r_0}/S}, \tag{66}$$

where we incorporate $c_{\mathbf{z}}$ into c_1, c_2. To estimate the order of the lower and upper bound in (66), we first notice that for any constant $c > 0$:

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

Furthermore, for any $K_{M,N}$ such that $\gamma_S K_{M,N}^{2r_0} \to C_1$, with $0 < C_1 \leq \infty$, we have:

$$\begin{aligned}\sum_{i=1}^{K_{M,N}} \frac{1}{1+c\lambda i^{2r_0}/S} &= \sum_{i=1}^{K_{M,N}} \frac{1}{1+[\frac{i}{(S/c\lambda)^{1/2r_0}}]^{2r_0}} \cdot \frac{1}{(S/c\lambda)^{1/2r_0}} \cdot (S/c\lambda)^{1/2r_0} \\ &\asymp (S/c\lambda)^{1/2r_0} \int_0^{C_1} \frac{1}{1+x^{2r_0}} \mathrm{d}x \asymp \gamma_S^{-1/2r_0}. \end{aligned} \tag{67}$$

Combining (64), (65), (66) and (67), we end up having $\mathrm{tr}\,(\mathbf{I}-\mathbf{W}) \lesssim O_P((M\vee N)\gamma_S^{-1/2r_0}) = O_P(\sqrt{S}\gamma_S^{-1/2r_0})$. Similarly, by lower bounding $\rho_d(\widehat{\mathbf{\Sigma}}_{\mathbf{z}})$ using $\rho_d(\mathbf{\Sigma}_{\mathbf{z}})/2$ with high probability in (64), and following similar arguments as (65), (66) and (67), we can obtain the lower bound of $\mathrm{tr}\,(\mathbf{I}-\mathbf{W})$ as $\mathrm{tr}\,(\mathbf{I}-\mathbf{W}) \geq O_P(\gamma_S^{-1/2r_0})$, which establishes the final result.

The order of $\mathrm{tr}\,(\mathbf{W})$ is trivial in that:

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

■

B.3 Proof of Lemma 16

Proof From the definition of W in equation (49), we have

$$\begin{aligned}\frac{\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}}{T} &= \widehat{\mathbf{\Sigma}}_{\mathbf{x},\mathbf{x}} \otimes \mathbf{I}_S - \left(\widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \otimes \mathbf{K}\right)\left(\widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}} \otimes \mathbf{K} + \lambda \mathbf{I}_{SD}\right)^{-1}\left(\widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}} \otimes \mathbf{I}_S\right) \\ &= \left(\widehat{\mathbf{\Sigma}}_{\mathbf{x},\mathbf{x}} - \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1} \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}\right) \otimes \mathbf{I}_S \\ &+ \left(\widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}} \otimes \mathbf{I}_S\right)^\top \left[\widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}}^2 \otimes \lambda^{-1}\mathbf{K} + \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}} \otimes \mathbf{I}_S\right]^{-1}\left(\widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}} \otimes \mathbf{I}_S\right), \end{aligned} \tag{68}$$

where the second term in (68) is positive semi-definite since $\underline{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}}), \underline{\rho}(\mathbf{K}) \geq 0$ and the whole term is symmetric. Therefore, one can lower bound $\underline{\rho}(\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}/T)$ by $\underline{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{x},\mathbf{x}} - \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1} \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}})$. For simplicity, we will use $\mathbf{A}, \mathbf{B}, \mathbf{C}$ to denote $\mathbf{\Sigma}^*_{\mathbf{x},\mathbf{x}}, \mathbf{\Sigma}^*_{\mathbf{z},\mathbf{x}}, (\mathbf{\Sigma}^*_{\mathbf{z},\mathbf{z}})^{-1}$, and $\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}$ to denote $\widehat{\mathbf{\Sigma}}_{\mathbf{x},\mathbf{x}}, \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}, \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1}$. We will use $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}^*$ to denote the estimated and true covariance matrix of $[\mathbf{x}_t^\top, \mathbf{z}_t^\top]^\top$. It is evident that $\|\mathbf{A}\|_s, \|\mathbf{B}\|_s \leq \|\mathbf{\Sigma}^*\|_s$.

The rest of the proof focuses on showing that $\underline{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{x},\mathbf{x}} - \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1} \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}})$ converges in probability to $c_{0,S} = \underline{\rho}(\mathbf{\Sigma}_{\mathbf{x},\mathbf{x}} - \mathbf{\Sigma}_{\mathbf{z},\mathbf{x}}^\top \mathbf{\Sigma}_{\mathbf{z},\mathbf{z}}^{-1} \mathbf{\Sigma}_{\mathbf{z},\mathbf{x}})$, as long as $S \log S/T \to 0$. For notational brevity, we omit the subscript s for spectral norm notation and simply use $\|\cdot\|$.

To start with, we have:

$$\begin{aligned}
\|\widehat{\mathbf{A}} - \widehat{\mathbf{B}}^\top \widehat{\mathbf{C}} \widehat{\mathbf{B}} - (\mathbf{A} - \mathbf{B}^\top \mathbf{C} \mathbf{B})\| &\leq \|\widehat{\mathbf{A}} - \mathbf{A}\| + \|\widehat{\mathbf{B}}^\top \widehat{\mathbf{C}} \widehat{\mathbf{B}} - \mathbf{B}^\top \widehat{\mathbf{C}} \mathbf{B}\| + \|\mathbf{B}^\top \widehat{\mathbf{C}} \mathbf{B} - \mathbf{B}^\top \mathbf{C} \mathbf{B}\| \\
&\leq \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\| + \|(\widehat{\mathbf{B}} - \mathbf{B})^\top \widehat{\mathbf{C}} \widehat{\mathbf{B}}\| \\
&\quad + \|\mathbf{B}^\top \mathbf{C} (\widehat{\mathbf{B}} - \mathbf{B})\| + \|\mathbf{B}^\top (\widehat{\mathbf{C}} - \mathbf{C}) \mathbf{B}\| \\
&\leq \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\| + \|\widehat{\mathbf{B}} - \mathbf{B}\| \|\widehat{\mathbf{C}}\| (\|\widehat{\mathbf{B}}\| + \|\mathbf{B}\|) + \|\mathbf{B}\|^2 \|\widehat{\mathbf{C}} - \mathbf{C}\|.
\end{aligned} \tag{69}$$

Then using lemma 21, we have that $\|\widehat{\mathbf{C}} - \mathbf{C}\| \xrightarrow{p} 0$ and $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\| \xrightarrow{p} 0$, where the latter requires $S \log S/T \to 0$. Therefore, with high probability, we have $\|\widehat{\mathbf{C}}\| \leq 2\|\mathbf{C}\|$ and $\|\widehat{\mathbf{B}} - \mathbf{B}\| \leq \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\| \leq 2\|\mathbf{\Sigma}^*\|$. Together with these results and (69), with high probability, we have:

$$\|\widehat{\mathbf{A}} - \widehat{\mathbf{B}}^\top \widehat{\mathbf{C}} \widehat{\mathbf{B}} - (\mathbf{A} - \mathbf{B}^\top \mathbf{C} \mathbf{B})\| \leq (1 + 2\|\mathbf{C}\| + 3\|\mathbf{\Sigma}^*\|) \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\| + \|\mathbf{\Sigma}^*\|^2 \|\widehat{\mathbf{C}} - \mathbf{C}\|. \tag{70}$$

The upper bound in (70) can be arbitrarily small as $S, T \to \infty$ since $\|\widehat{\mathbf{C}} - \mathbf{C}\| \xrightarrow{p} 0$ and $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\| \xrightarrow{p} 0$. Therefore, we have:

$$\underline{\rho}(\widehat{\mathbf{\Sigma}}_{\mathbf{x},\mathbf{x}} - \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}^\top \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{z}}^{-1} \widehat{\mathbf{\Sigma}}_{\mathbf{z},\mathbf{x}}) \geq \frac{1}{2} \underline{\rho}(\mathbf{\Sigma}_{\mathbf{x},\mathbf{x}} - \mathbf{\Sigma}_{\mathbf{z},\mathbf{x}}^\top \mathbf{\Sigma}_{\mathbf{z},\mathbf{z}}^{-1} \mathbf{\Sigma}_{\mathbf{z},\mathbf{x}}), \tag{71}$$

hold with high probability. This completes the proof.

■

B.4 Proof of Lemma 19

Proof We first prove that $\boldsymbol{\xi}^\top \mathbf{A}\boldsymbol{\xi} = O_P(\mathrm{tr}\,(\mathbf{A}))$. By the Hanson-Wright inequality (see Theorem 1.1. of Rudelson and Vershynin (2013)), we have:

$$\mathrm{P}\left(\left|\boldsymbol{\xi}^\top \mathbf{A}\boldsymbol{\xi} - \mathrm{E}[\boldsymbol{\xi}^\top \mathbf{A}\boldsymbol{\xi}]\right| > t\right) < 2\exp\left[-c\min\left\{\frac{t^2}{K^4\|\mathbf{A}\|_{\mathrm{F}}^2}, \frac{t}{K^2\|\mathbf{A}\|_{op}}\right\}\right], \tag{72}$$

where c is a universal constant, $K = \sqrt{8/3}\sigma$ and since all eigenvalues of $\mathbf{A}$ are bounded within $[0,1]$, we have $\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{tr}\,(\mathbf{A}^2) \leq \mathrm{tr}\,(\mathbf{A})$, and thus we can further bound the RHS of (72) as:

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

Since $\mathrm{E}[\boldsymbol{\xi}^\top \mathbf{A}\boldsymbol{\xi}] = \sigma^2\mathrm{tr}\,(\mathbf{A})$, by setting $t = \epsilon \cdot \mathrm{tr}\,(\mathbf{A})$, we have that $\mathrm{P}[|\boldsymbol{\xi}^\top \mathbf{A}\boldsymbol{\xi}/\mathrm{E}[\boldsymbol{\xi}^\top \mathbf{A}\boldsymbol{\xi}] - 1| > \epsilon] < 2\exp(-c\epsilon\mathrm{tr}\,(\mathbf{A}))$. We can establish the final statement by $\mathrm{tr}\,(\mathbf{A}) \to \infty$.

The second statement holds simply because $\boldsymbol{\xi}^\top \mathbf{A}\boldsymbol{\xi} = O_P(\mathrm{tr}\,(\mathbf{A}))$ and $\boldsymbol{\xi}^\top \boldsymbol{\xi} = O_P(d)$. ■

B.5 Proof of Lemma 18

Proof Since $\mathbf{A} - \mathbf{B}$ is positive semi-definite, we have $\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}} \gtrsim \mathbf{I}$, where $\mathbf{M} \gtrsim \mathbf{N}$ means that $\mathbf{M} - \mathbf{N}$ is positive semi-definite. Therefore, we have $\mathbf{B}^{-\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}} \gtrsim \mathbf{I}$. Notice that $\mathbf{B}^{-\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}}$ and $\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}$ have the same eigenvalues, thus $\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{-1}\mathbf{A}^{\frac{1}{2}} \gtrsim \mathbf{I}$, and thus $\mathbf{B}^{-1} \gtrsim \mathbf{A}^{-1}$. ■

C Additional Details on Simulation Experiments

We generate the simulated dataset according to the MARAC(P, Q) model specified by (1) and (4). 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 6, we plot the simulated $\mathbf{A}_1, \mathbf{B}_1, \mathbf{\Sigma}_r, \mathbf{\Sigma}_c$ when $(M, N) = (20, 20)$.

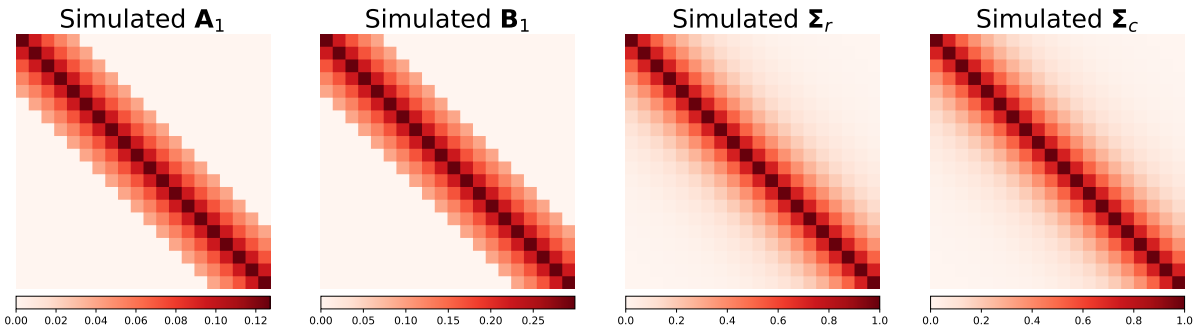

Figure 6: Visualization of the simulated $\mathbf{A}_1, \mathbf{B}_1, \mathbf{\Sigma}_r, \mathbf{\Sigma}_c$ with $M = N = 20$.

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 point 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:

$$k_\eta(s_1, s_2) = \left(\frac{1}{4\pi} + \frac{\eta}{12\pi}\right) - \frac{\eta}{8\pi}\sqrt{\frac{1 - \langle s_1, s_2\rangle}{2}}, \quad s_1, s_2 \in \mathbb{S}^2, \tag{73}$$

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 (73). Finally, we simulate the vector time series $\mathbf{z}_t$ using a VAR(1) process. In Figure 7, we visualize the simulated functional parameters as well as the vector time series from one random draw.

In Figure 8, we visualize the ground truth of g_3 and both its PMLE and truncated PMLE estimators. It is evident that the truncated PMLE 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.

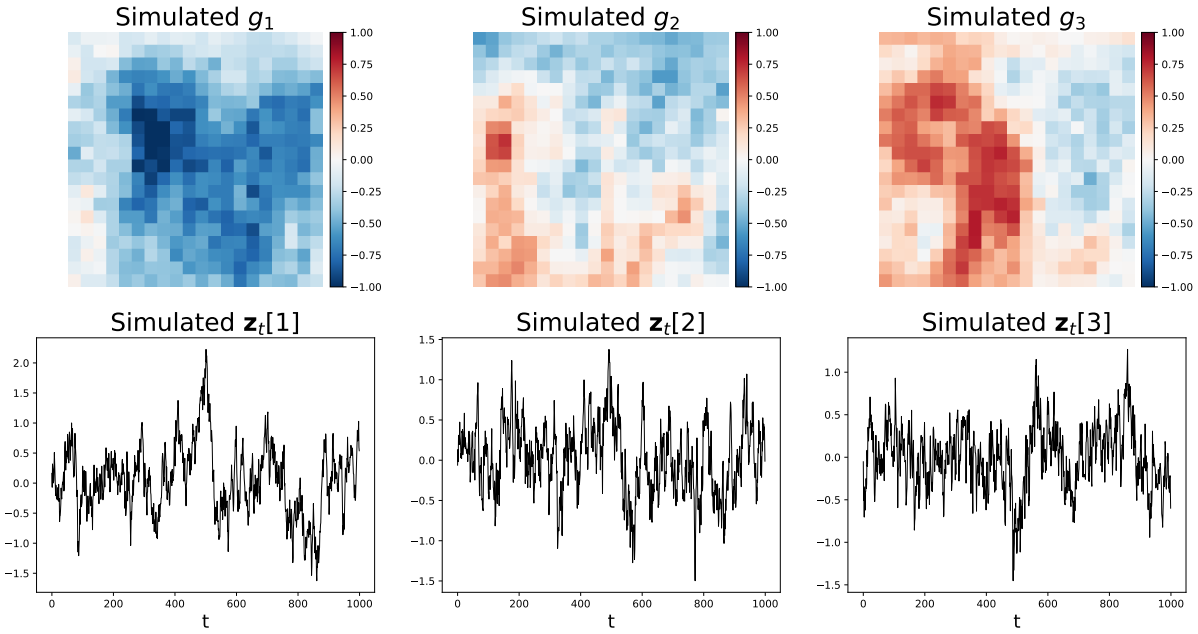

Figure 7: Simulated functional parameters g_1, g_2, g_3 evaluated on a 20×20 spatial grid (top row) and the corresponding auxiliary vector time series (bottom row).

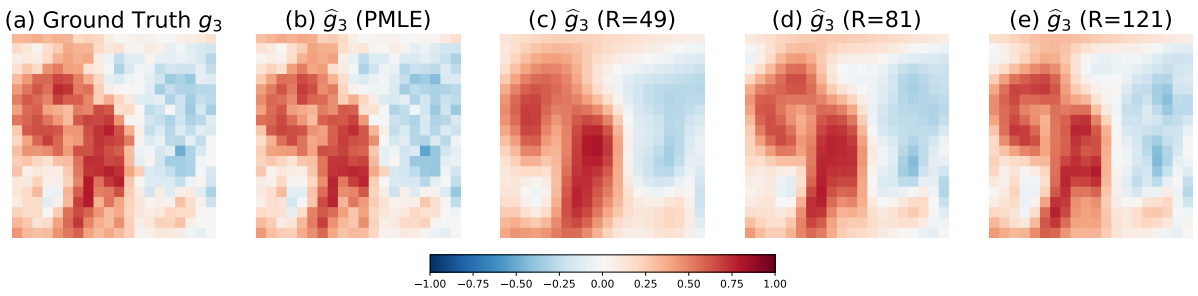

Figure 8: Ground truth g_3 (panel (a)) against the PMLE estimator $\widehat{g}_3$ (panel (b)) and the truncated PMLE estimator $\widehat{g}_3$ using $R \in \{49, 81, 121\}$ basis functions. $M = 20$.