Statistica Sinica Preprint No: SS-2025-0029	
Title	Matrix Autoregressive Model with Vector Time Series Covariates for Spatiotemporal Data
Manuscript ID	SS-2025-0029
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0029
Complete List of Authors	Hu Sun, Zuofeng Shang and Yang Chen
Corresponding Authors	Yang Chen
E-mails	ychenang@umich.edu
Notice: Accepted author version.	

Statistica Sinica

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

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

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

Abstract: We develop a new methodology for forecasting matrix-valued time series with historical matrix data and auxiliary vector time series data. We focus on a time series of matrices defined on a static 2-D spatial grid and an auxiliary time series of non-spatial vectors. The proposed model, Matrix AutoRegression with Auxiliary Covariates (MARAC), contains an autoregressive component for the historical matrix predictors and an additive component that maps the auxiliary vector predictors to a matrix response via tensor-vector product. The autoregressive component adopts a bi-linear transformation framework following Chen et al. (2021), significantly reducing the number of parameters. The auxiliary component posits that the tensor coefficient, which maps non-spatial predictors to a spatial response, contains slices of spatially smooth matrix coefficients that are discrete evaluations of smooth functions on a spatial grid from a Reproducing 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.

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

1. Introduction

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

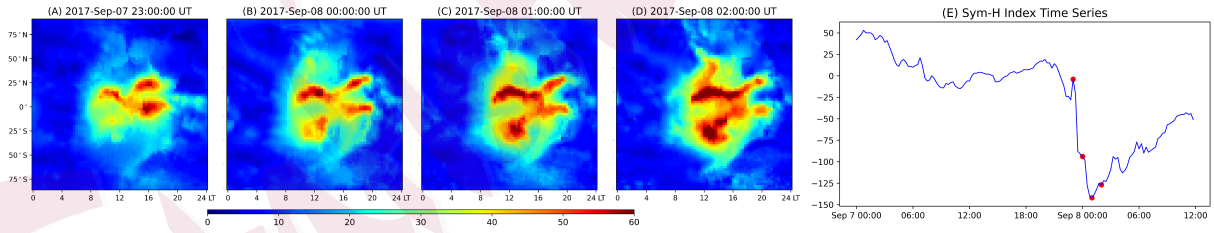

Figure 1: An example of matrix time series with auxiliary vector time series. Panels (A)-(D) show the global Total Electron Content (TEC) distribution at four timestamps, separated by 1 hour, on the latitude-local time grid (source: the IGS TEC database (Hernández-Pajares et al., 2009)). Panel (E) plots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

The matrix-valued time series, such as the TEC time series, is often associated with auxiliary vector time series that measure the same object, such as the Earth's ionosphere, from a different data modality. In panel (E) of Figure 1, we plot the global SYM-H index,

which measures the geomagnetic activity caused by solar eruptions that can finally affect the global TEC distribution. These non-spatial auxiliary covariates carry additional information related to the matrix time series dynamics, as one can tell from the sudden decrease of the Sym-H index around 00:00 UT on September 8, 2017, and the associated intensification of the global TEC near the equatorial regions.

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

A naive but straightforward prediction model is to vectorize the matrices as vectors and make predictions via Vector Autoregression (VAR) (Stock and Watson, 2001). The auxiliary vector covariates can be incorporated once concatenated with the vectorized matrix predictors. However, vectorizing matrix data leads to the loss of spatial information and also requires a significant amount of parameters, given the high dimensionality of the data. To avoid vectorizing the matrix data, scalar-on-tensor regression (Zhou et al., 2013; Guhaniyogi et al., 2017; Li et al., 2018; Papadogeorgou et al., 2021) tackles the problem by using matrix

predictors directly. However, these models are built for *scalar* responses while in our setting we are dealing with *matrix* responses. Dividing the matrix into individual scalars and fitting scalar-on-tensor regressions still requires a significant number of parameters and, more importantly, it fails to take the spatial information of the response into account.

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

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

The other distinction of our model is that our model utilizes the spatial information of the matrix response. In our model, a key assumption is that the vector predictor has similar predictive effects on neighboring locations in the matrix response. This is equivalent to saying that the tensor coefficient is spatially smooth and is typically done via adding a total-

variation (TV) penalty (Wang et al., 2017; Shen et al., 2022; Sun et al., 2023) to the unknown tensor. The TV penalty leads to piecewise smooth estimators with sharp edges and enables feature selections. However, the estimation with the TV penalty requires solving non-convex optimization problems, making the subsequent theoretical analysis difficult. Our model uses a simpler approach by assuming that the tensor coefficients are discrete evaluations of functional parameters from a Reproducing Kernel Hilbert Space (RKHS). Such a kernel method has been widely used in scalar-on-image regressions (Kang et al., 2018) where the regression coefficients of the image predictor are constrained to be spatially smooth.

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

To summarize, our paper has two major contributions. Firstly, we build a unified matrix autoregression framework for spatio-temporal data that incorporates lower-order scalar/vector time series covariates. Such a framework has strong application motivation where domain scientists are curious about integrating spatial and non-spatial data information for predictions and inference. The framework also bridges regression methodologies with tensor predictors and responses of non-uniform modes, making the theoretical investigation itself an interesting topic. Secondly, we propose to estimate coefficients of the auxiliary covariates, together with the autoregressive coefficients, in a single penalized maximum like-

lihood estimation (MLE) framework with the RKHS functional norm penalty. The RKHS framework builds spatial continuity into the regression coefficients. We establish the joint asymptotics of the autoregressive coefficients and the functional parameters under fixed/high matrix dimensionality regimes and propose an efficient alternating minimization algorithm for estimation and validate it with extensive simulations and real applications.

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

2. Model

2.1 Notation

We adopt the following notations throughout the article. We use calligraphic bold-face letters (e.g., $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{G}}$) for tensors with at least three modes, uppercase bold-face letters (e.g., $\mathbf{X}, \mathbf{G}$) for matrices, and lowercase bold-face letters (e.g., $\mathbf{x}, \mathbf{z}$) for vectors and blackboard bold-faced letters for sets (e.g., $\mathbb{R}, \mathbb{S}$). To subscript any tensor/matrix/vector, we use square brackets with subscripts such as $[\boldsymbol{\mathcal{G}}]_{ijd}, [\mathbf{z}_t]_d, [\mathbf{X}_t]_{ij}$, and we keep the subscript t inside the square bracket to index time. Any fibers and slices of tensor are subscripted with colons,

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

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

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

2.2 Matrix AutoRegression with Auxiliary Covariates (MARAC)

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

$$\mathbf{X}_t = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{X}_{t-p} \mathbf{B}_p^\top + \sum_{q=1}^{Q} \boldsymbol{\mathcal{G}}_q \bar{\times} \mathbf{z}_{t-q} + \mathbf{E}_t, \tag{2.1}$$

where $\mathbf{A}_p \in \mathbb{R}^{M\times M}, \mathbf{B}_p \in \mathbb{R}^{N\times N}$ are the autoregressive coefficients for the lag-p matrix predictor and $\boldsymbol{\mathcal{G}}_q \in \mathbb{R}^{M\times N\times D}$ is the tensor coefficient for the lag-q vector predictor, and $\mathbf{E}_t$ is a noise term whose distribution will be specified later. The lag parameters P, Q are hyperparameters of the model, and we often refer to the model (2.1) as MARAC(P, Q).

Based on model (2.1), for the $(i, j)^{\text{th}}$ element of $\mathbf{X}_t$, the MARAC(P, Q) specifies the following model:

$$[\mathbf{X}_t]_{ij} = \sum_{p=1}^{P} \left\langle [\mathbf{A}_p]_{i:}^\top [\mathbf{B}_p]_{j:}, \mathbf{X}_{t-p} \right\rangle + \sum_{q=1}^{Q} [\boldsymbol{\mathcal{G}}_q]_{ij:}^\top \mathbf{z}_{t-q} + [\mathbf{E}_t]_{ij}, \tag{2.2}$$

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 (2.2) that the auxiliary vector covariates enter the model via an elementwise linear model. The autoregressive term utilizes $\mathbf{A}_p, \mathbf{B}_p$ to transform each lag-p predictor in a bi-linear form. Using such a bi-linear transformation greatly reduces the total number of parameters of the autoregressive term from $O(M^2N^2)$ to $O((M^2 + N^2))$. When the spatial dimensions M, N are high, one can further reduce the dimensionality by assuming $\mathbf{A}_p, \mathbf{B}_p$

are low-rank matrices (Xiao et al., 2022), or the tensors with frontal slices being $\mathbf{A}_p, \mathbf{B}_p$ are low-rank (Wang et al., 2022). Apart from low-rank structure, one could also constrain the autoregressive coefficients to generate smooth predictions via restricting $\mathbf{A}_p, \mathbf{B}_p$ to be matrices with a banded structure (Guo et al., 2016; Hsu et al., 2021). However, these configurations would significantly complicate the modeling and theoretical analysis and result in additional model selection problems. Furthermore, in our satellite imaging data, the dependency structure can be arbitrary (among spatial locations). Thus, we do not want to be constrained by any assumptions. In this paper, we keep a more straightforward setup and put no constraints on $\mathbf{A}_p, \mathbf{B}_p$ while acknowledging that additional constraints can benefit the computational efficiency under high-dimensional settings. Additionally, we consider a setting where D, the dimension of the auxiliary covariates $\mathbf{z}_t$, is fixed instead of growing with M and N. This setup greatly simplifies the theoretical analysis and reflects the application scenario where one has a fixed set of auxiliary predictors but a matrix-valued data with growing spatial resolution.

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

An alternative formulation for $\boldsymbol{\mathcal{G}}_q$ would be a low-rank form (Li and Zhang, 2017). Similar low-rank assumptions can be found in matrix time series factor model (Chen and Fan, 2023; Gao and Tsay, 2025, 2023), where our vector predictors $\{\mathbf{z}_t\}_{t=-\infty}^{\infty}$ become a matrix-valued,

unknown factor time series. Typically, low-rank representations could significantly reduce the dimensionality of parameters in contexts with limited data (Zhou et al., 2013). However, even with a low-rank structure over $\boldsymbol{\mathcal{G}}_q$, the number of parameters is still at the order of $O(D(M^2+N^2))$, bounded by the autoregressive parameters, and thus cannot benefit from the low-rank form but complicates the theoretical analysis. Also, we are motivated by applications that forecast time series of spatial data with vector predictors, and we want to model the spatial continuity of the regression coefficients explicitly. Therefore, we choose an RKHS framework over the low-rank framework.

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

$$\mathbf{vec}\left(\mathbf{E}_t\right) \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r\right), \quad t \in [T] \tag{2.3}$$

where $\boldsymbol{\Sigma}_r \in \mathbb{R}^{M\times M}, \boldsymbol{\Sigma}_c \in \mathbb{R}^{N\times N}$ are the row/column covariance components. Such a Kronecker-product covariance is commonly seen in the covariance models for multi-way data (Hoff, 2011; Tsiligkaridis et al., 2013; Fosdick and Hoff, 2014; Zhou, 2014; Lyu et al., 2019; Li and Xiao, 2021) with the merit of reducing the number of parameters significantly.

Compared to existing models that only deal with either matrix or vector predictors, our model (2.1) can incorporate predictors with non-uniform modes. If one redefines $\mathbf{E}_t$ in our model as $\sum_{q=1}^{Q} \boldsymbol{\mathcal{G}}_q \bar{\times} \mathbf{z}_{t-q} + \mathbf{E}_t$, i.e., all terms except the autoregressive term, then our model

ends up specifying:

$$\mathrm{Cov}(\mathbf{vec}\left(\mathbf{E}_t\right), \mathbf{vec}\left(\mathbf{E}_{t'}\right)) = \mathbb{1}_{\{t=t'\}} \cdot \mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r + \mathbf{F}\mathbf{M}\mathbf{F}^\top$$

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

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

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

noises, nor does it consider the auxiliary covariates. Our model generalizes these previous works to allow for temporally dependent noise with both local and grid spatial dependency.

The combination of (2.1) and (2.3) specifies the complete MARAC(P, Q) model. We visualize our MARAC model in Figure 2. Vectorizing both sides of (2.1) yields the vectorized

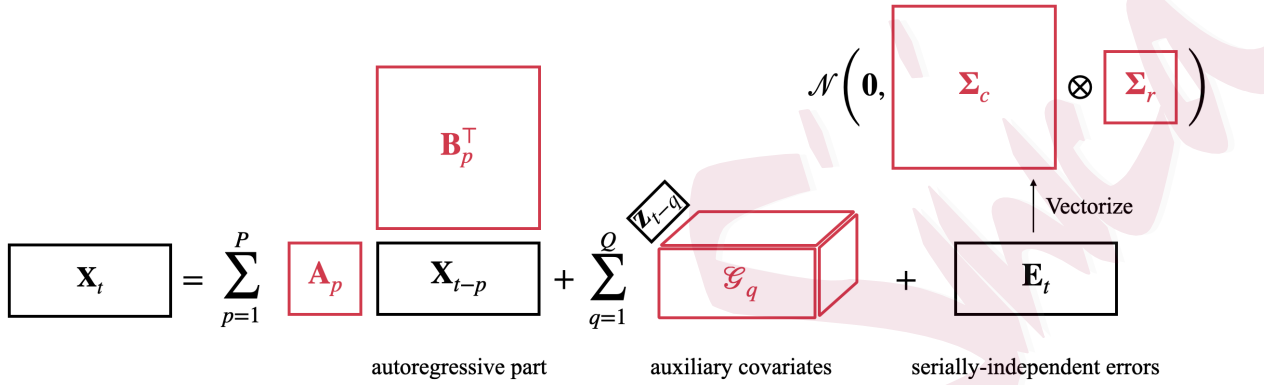

Figure 2: A schematic illustration of the MARAC model in (2.1) and (2.3). All parameters are highlighted in red.

MARAC(P, Q) model:

$$\mathbf{x}_t = \sum_{p=1}^{P} (\mathbf{B}_p \otimes \mathbf{A}_p)\, \mathbf{x}_{t-p} + \sum_{q=1}^{Q} \mathbf{G}_q^\top \mathbf{z}_{t-q} + \mathbf{e}_t, \quad \mathbf{e}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r) \tag{2.4}$$

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

3. Estimating Algorithm

This section discusses the parameter estimation for the MARAC(P, Q) model in (2.1). We propose a penalized maximum likelihood estimator (MLE) in Section 3.1 for exact parameter estimation. Then in Section 3.2, we outline the model selection criterion for selecting the lag hyperparameters whose consistency will be validated empirically in Section 5.

3.1 Penalized Maximum Likelihood Estimation (MLE)

To estimate the parameters of the MARAC(P, Q) model, which we denote collectively as $\boldsymbol{\Theta}$, we propose a penalized maximum likelihood estimation (MLE) approach. Following the distribution assumption on $\mathbf{E}_t$ in (2.3), we can write the negative log-likelihood of $\{\mathbf{X}_t\}_{t=1}^T$ with a squared RKHS functional norm penalty, after dropping the constants, as:

$$\mathfrak{L}_\lambda(\boldsymbol{\Theta}) = -\frac{1}{T}\sum_{t\in[T]} \ell\left(\mathbf{X}_t|\mathbf{X}_{t-1:P}, \mathbf{z}_{t-1:Q}; \boldsymbol{\Theta}\right) + \frac{\lambda}{2}\sum_{q\in[Q]}\sum_{d\in[D]} \|g_{q,d}\|^2_{\mathbb{H}_k}, \tag{3.5}$$

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

$$\ell\left(\mathbf{X}_t|\mathbf{X}_{t-1:P}, \mathbf{z}_{t-1:Q}; \boldsymbol{\Theta}\right) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r| - \frac{1}{2}\mathbf{r}_t^\top\left(\boldsymbol{\Sigma}_c^{-1} \otimes \boldsymbol{\Sigma}_r^{-1}\right)\mathbf{r}_t, \tag{3.6}$$

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

$$\min_{\boldsymbol{\Theta}} \mathfrak{L}_\lambda(\boldsymbol{\Theta}), \quad \text{s.t. } g_{q,d}(s_{ij}) = [\boldsymbol{\mathcal{G}}_q]_{ijd}, \text{ for all } s_{ij} \in \mathbb{S}. \tag{3.7}$$

We now explicitly define the functional norm penalty in (3.5) 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 Mercer's Theorem (Williams and Rasmussen, 2006):

$$k(s_{ij}, s_{uv}) = \sum_{r=1}^{\infty} \lambda_r \psi_r(s_{ij}) \psi_r(s_{uv}), \quad s_{ij}, s_{uv} \in [0,1]^2, \tag{3.8}$$

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

$$\|g\|_{\mathbb{H}_k} = \sqrt{\sum_{r=1}^{\infty} \frac{\beta_r^2}{\lambda_r}}, \qquad \text{where } g(\cdot) = \sum_{r=1}^{\infty} \beta_r \psi_r(\cdot), \tag{3.9}$$

following van Zanten and van der Vaart (2008).

Given any $\lambda > 0$ in (3.5), the generalized representer theorem (Schölkopf et al., 2001) suggests that the solution of the functional parameters, denoted as $\{\widetilde{g}_{q,d}\}_{q=1,d=1}^{Q,D}$, of the minimization problem (3.7), with all other parameters held fixed, is a linear combination of the representers $\{k(\cdot, s)\}_{s\in\mathbb{S}}$ plus a linear combination of the basis functions $\{\phi_1, \ldots, \phi_J\}$ of the null space of $\mathbb{H}_k$, i.e.,

$$\widetilde{g}_{q,d}(\cdot) = \sum_{s\in\mathbb{S}} \gamma_s k(\cdot, s) + \sum_{j=1}^{J} \alpha_j \phi_j(\cdot), \quad \|\phi_j\|_{\mathbb{H}_k} = 0, \tag{3.10}$$

where we omit the subscript (q,d) for the coefficient γ_s, α_j for brevity but they are essentially different for each (q,d). We assume that the null space of $\mathbb{H}_k$ contains only the zero function for the remainder of the paper. As a consequence of (3.10), the minimization problem

in (3.7) can be reduced to a finite-dimensional Kernel Ridge Regression (KRR) problem. We summarize the discussion above in Proposition 1:

Proposition 1. *If $\lambda > 0$, the constrained minimization problem in* (3.7) *is equivalent to the following unconstrained kernel ridge regression problem:*

$$\min_{\mathbf{\Theta}} \left\{ \frac{1}{2} \log |\mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r| + \frac{1}{2T} \sum_{t \in [T]} \mathbf{r}_t^\top \left(\mathbf{\Sigma}_c^{-1} \otimes \mathbf{\Sigma}_r^{-1} \right) \mathbf{r}_t + \frac{\lambda}{2} \sum_{q \in [Q]} tr\left(\mathbf{\Gamma}_q^\top \mathbf{K} \mathbf{\Gamma}_q \right) \right\}, \tag{3.11}$$

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

We provide proof in the supplemental material. After introducing the functional norm penalty, the original tensor coefficient is now converted to a linear combination of the representer functions with the relationship that $[\boldsymbol{\mathcal{G}}_q]_{ijd} = \langle [\mathbf{K}]_{u:}^\top, [\mathbf{\Gamma}_q]_{:d} \rangle$ where $u = i + (j-1)M$. For more efficient computation, one can use a set of basis functions based on the spectral decomposition of the selected kernel as an approximation: $[\boldsymbol{\mathcal{G}}_q]_{ijd} \approx \sum_{r \in R} [\boldsymbol{\theta}_{q,d}]_r \psi_r(s_{ij})$. The choice of the number of basis functions can be determined via cross-validation, and generally, we observe better results with more basis functions, given enough data. We discuss this approach in Section S7.2 of the supplemental material.

The choice of the kernel function $k(\cdot, \cdot)$ depends on the application context. For spatial data in Euclidean space, common choices include the radial-basis function (RBF) kernel or the Matérn kernel (Williams and Rasmussen, 2006, Sec. 4.2). For data distributed on a sphere, which is the context of the TEC data used in this paper, one could consider the

Lebedev kernel (Kennedy et al., 2013) or the von Mises-Fisher kernel (Banerjee et al., 2005).

We attempt to solve the minimization problem in (3.11) with an alternating minimization algorithm (Attouch et al., 2013) where we update one block of parameters at a time, keeping the others fixed, following the order of: $\mathbf{A}_1 \to \mathbf{B}_1 \to \cdots \to \mathbf{A}_P \to \mathbf{B}_P \to \mathbf{\Gamma}_1 \to \cdots \to \mathbf{\Gamma}_Q \to \mathbf{\Sigma}_r \to \mathbf{\Sigma}_c \to \mathbf{A}_1 \to \cdots$. We choose the alternating minimization algorithm for its simplicity and efficiency. Each step of the algorithm conducts exact minimization over one block of the parameters, leading to a non-increasing sequence of the objective function, which guarantees the convergence of the algorithm towards a local stationary point. We abstract away the exact updating formula for each parameter here and include them in Section S1 of the supplemental material. We conclude this session with a remark on identifiability.

Remark 1. (Identifiability Constraint) The MARAC(P, Q) model specified in (2.1) is scale-unidentifiable in that one can re-scale each pair of $(\mathbf{A}_p, \mathbf{B}_p)$ by a non-zero constant c and obtain $(c \cdot \mathbf{A}_p, c^{-1} \cdot \mathbf{B}_p)$ without changing their Kronecker product. To enforce scale identifiability, we re-scale the algorithm output for each pair of $(\mathbf{A}_p, \mathbf{B}_p)$ such that $\|\mathbf{A}_p\|_{\mathrm{F}} = 1$, $\mathrm{sign}(\mathrm{tr}\,(\mathbf{A}_p)) = 1$. The identifiability constraint is enforced before outputting the estimators.

3.2 Lag Selection

The MARAC(P, Q) model (2.1) has three hyperparameters: the autoregressive lag P, the auxiliary covariate lag Q, and the RKHS norm penalty weight λ. In practice, λ can be chosen by cross-validation, while choosing P and Q requires a more formal model selection criterion. We propose to select P and Q by using information criterion, including the Akaike Information Criterion (AIC) (Akaike, 1998) and the Bayesian Information Criterion (BIC) (Schwarz, 1978). Here, we formally define the AIC and BIC for the MARAC(P, Q)

model and empirically validate their 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 freedom* of the model. We can then define the AIC and the BIC as follows:

$$\mathrm{AIC}(P, Q, \lambda) = -2 \sum_{t \in [T]} \ell(\mathbf{X}_t | \mathbf{X}_{t-1:P}, \mathbf{z}_{t-1:Q}, \widehat{\mathbf{\Theta}}) + 2 \cdot \mathbf{df}_{P,Q,\lambda}, \tag{3.12}$$

$$\mathrm{BIC}(P, Q, \lambda) = -2 \sum_{t \in [T]} \ell(\mathbf{X}_t | \mathbf{X}_{t-1:P}, \mathbf{z}_{t-1:Q}, \widehat{\mathbf{\Theta}}) + \log(T) \cdot \mathbf{df}_{P,Q,\lambda}. \tag{3.13}$$

To calculate $\mathbf{df}_{P,Q,\lambda}$, we decompose it into the sum of three components: 1) for each pair of the autoregressive coefficient $\widehat{\mathbf{A}}_p, \widehat{\mathbf{B}}_p$, the model has $(M^2 + N^2 - 1)$ degrees of freedom; 2) for the noise covariance $\widehat{\mathbf{\Sigma}}_r, \widehat{\mathbf{\Sigma}}_c$, the model has $(M^2 + N^2)$ degrees of freedom; and 3) for the auxiliary covariate functional parameters $\widehat{g}_{q,1}, \ldots, \widehat{g}_{q,D}$, inspired by the kernel ridge regression estimator in (S1.4), we define the sum of their degrees of freedom as:

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

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

4. Theoretical Analysis

This section presents the theoretical analyses of the MARAC model. We first formulate the condition under which the matrix and vector time series are *jointly stationary*. Under

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

4.1 Stationarity Condition

To facilitate the theoretical analysis, we make another assumption for the vector time series $\mathbf{z}_t$, which significantly simplifies the presentation of our theoretical analysis.

Assumption 1. The D-dimensional auxiliary vector time series $\{\mathbf{z}_t\}_{t=-\infty}^{\infty}$ follows a stationary VAR($\widetilde{Q}$) process:

$$\mathbf{z}_t = \sum_{\widetilde{q}=1}^{\widetilde{Q}} \mathbf{C}_{\widetilde{q}} \mathbf{z}_{t-\widetilde{q}} + \boldsymbol{\nu}_t, \tag{4.14}$$

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

Given Assumption 1, we now derive the condition for $\mathbf{x}_t$ and $\mathbf{z}_t$ to be *jointly stationary*:

Theorem 1 (MARAC Stationarity Condition)**.** *Assume that Assumption 1 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* (2.1)*, 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*

$$\det\left[\mathbf{I}_S - \sum_{p=1}^{P} (\mathbf{B}_p \otimes \mathbf{A}_p)\, y^p\right] \neq 0, \quad \det\left[\mathbf{I}_D - \sum_{\widetilde{q}=1}^{\widetilde{Q}} \mathbf{C}_{\widetilde{q}} y^{\widetilde{q}}\right] \neq 0. \tag{4.15}$$

As a special case where $P = \widetilde{Q} = 1$, the stationarity condition in (4.15) is equivalent to $\bar{\rho}(\mathbf{A}_1) \cdot \bar{\rho}(\mathbf{B}_1) < 1$ and $\bar{\rho}(\mathbf{C}_1) < 1$, where $\bar{\rho}(\cdot)$ is the spectral radius of a square matrix. Based on Theorem 1, the stationarity of the matrix and vector time series relies on the stationarity of the autoregressive coefficients of the MARAC(P, Q) and VAR$(\widetilde{Q})$ models. The tensor coefficients $\boldsymbol{\mathcal{G}}_1, \ldots, \boldsymbol{\mathcal{G}}_Q$ do not affect the stationarity.

We can relax Assumption 1 to $\{\mathbf{z}_t\}_{t=-\infty}^{\infty}$ being covariance-stationary and independent of $\{\mathbf{E}_t\}_{t=-\infty}^{\infty}$ without affecting most of the theory below, just as the VARX model (Hamilton, 2020). But we decide to keep this assumption for the rest of the paper since having a joint autoregressive process for $\{\mathbf{X}_t\}_{t=-\infty}^{\infty}$ and $\{\mathbf{z}_t\}_{t=-\infty}^{\infty}$ greatly simplifies the analysis, especially under the high-dimensional regime in Section 4.3.

4.2 Finite Spatial Dimension Asymptotics

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

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

Assumption 2. The minimum eigenvalue of $\mathbf{K}$ is bounded below, i.e. $\underline{\rho}(\mathbf{K}) = \underline{c} > 0$.

As a result of Assumption 2, every $\boldsymbol{\mathcal{G}}_q^*$ has a unique kernel decomposition: $\mathbf{vec}\left(\boldsymbol{\mathcal{G}}_q^*\right) = (\mathbf{I}_D \otimes \mathbf{K})\boldsymbol{\gamma}_q^*$. Now we are ready to establish the consistency of the covariance matrix estimator $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_c \otimes \widehat{\boldsymbol{\Sigma}}_r$, which we summarize in Proposition 2.

Proposition 2 (Covariance Consistency)**.** *Assume that $\lambda \to 0$ as $T \to \infty$ and S is fixed, and Assumption 1, 2 and the stationarity condition in Theorem 1 hold, , then $\widehat{\boldsymbol{\Sigma}} \xrightarrow{p.} \boldsymbol{\Sigma}^*$.*

We can further establish the asymptotic normality of the other model estimators:

Theorem 2 (Asymptotic Normality)**.** *Assume that the matrix time series $\{\mathbf{X}_t\}_{t=-\infty}^{\infty}$ follows the MARAC(P,Q) model (2.1) with i.i.d. noise $\{\mathbf{E}_t\}_{t=-\infty}^{\infty}$ following (2.3) and Assumption 1, 2 and the stationarity condition in Theorem 1 hold and $\lambda = o(T^{-1/2})$. Additionally, assume that $\underline{\rho}(\mathrm{Var}([\mathbf{vec}\left(\mathbf{X}_t\right)^\top, \mathbf{z}_t^\top]^\top)) = \underline{c}' > 0$. Then suppose M, N are fixed and P, Q are known and denote $\mathbf{vec}\left(\mathbf{A}_p\right), \mathbf{vec}\left(\mathbf{B}_p^\top\right)$ as $\boldsymbol{\alpha}_p$ and $\boldsymbol{\beta}_p$ for any $p \in [P]$, the penalized MLE of the MARAC(P,Q) model is asymptotically normal:*

$$\sqrt{T}\begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 \otimes \widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\beta}_1^* \otimes \boldsymbol{\alpha}_1^* \\ \vdots \\ \widehat{\boldsymbol{\beta}}_P \otimes \widehat{\boldsymbol{\alpha}}_P - \boldsymbol{\beta}_P^* \otimes \boldsymbol{\alpha}_P^* \\ \mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{G}}}_1 - \boldsymbol{\mathcal{G}}_1^*\right) \\ \vdots \\ \mathbf{vec}\left(\widehat{\boldsymbol{\mathcal{G}}}_Q - \boldsymbol{\mathcal{G}}_Q^*\right) \end{bmatrix} \xrightarrow{d.} \mathcal{N}\left(\mathbf{0}, \mathbf{V}\boldsymbol{\Xi}\mathbf{V}^\top\right), \tag{4.16}$$

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

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

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

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

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

$$\mathbf{H} = \mathrm{E}\left[\mathbf{W}_t^\top (\boldsymbol{\Sigma}^*)^{-1}\mathbf{W}_t\right] + \boldsymbol{\zeta}\boldsymbol{\zeta}^\top, \quad \boldsymbol{\zeta}^\top = \left[(\boldsymbol{\alpha}_1^*)^\top, \cdots, (\boldsymbol{\alpha}_P^*)^\top, \mathbf{0}^\top\right].$$

The asymptotic distribution (4.16) indicates that all parameters are $\sqrt{T}$-consistent under fixed matrix dimensionality. Given this result, we have a corollary on testing the existence of the auxiliary covariates in the model:

Corollary 1 (Specification Test)**.** *Given the same assumption as Theorem 2, we have:*

$$T \cdot (\widehat{\mathbf{g}} - \mathbf{g}^*)^\top \boldsymbol{\Psi}^\dagger (\widehat{\mathbf{g}} - \mathbf{g}^*) \xrightarrow{d.} \chi_r^2, \tag{4.17}$$

where $\mathbf{g}^* = [\mathbf{vec}\left(\boldsymbol{\mathcal{G}}_1^*\right)^\top, \ldots, \mathbf{vec}\left(\boldsymbol{\mathcal{G}}_Q^*\right)^\top]^\top$ *and* $\widehat{\mathbf{g}}$ *is its estimator and* $\boldsymbol{\Psi}^\dagger$ *is the Moore-Penrose pseudo-inverse of* $\boldsymbol{\Psi} \coloneqq [\mathbf{O} : \mathbf{I}_{QD} \otimes \mathbf{K}]\boldsymbol{\Xi}[\mathbf{O} : \mathbf{I}_{QD} \otimes \mathbf{K}]^\top$*. Furthermore, we can prove that* $r \geq MNQD - 1$*. To test the hypothesis at significance level* α*:*

$$H_0 : \mathbf{g}^* = \mathbf{0}, \quad vs.\ H_1 : \mathbf{g}^* \neq \mathbf{0},$$

we have a test statistics $T \cdot \widehat{\mathbf{g}}^\top \mathbf{\Psi}^\dagger \widehat{\mathbf{g}}$ *with a rejection region* $\{\widehat{\mathbf{g}} | T \cdot \widehat{\mathbf{g}}^\top \mathbf{\Psi}^\dagger \widehat{\mathbf{g}} \geq \chi^2_{r,1-\alpha}\}$.

In practice, we will use plug-in estimators to estimate $\mathbf{\Psi}$ and use $\chi^2_{MNQD-1,1-\alpha}$ as the critical value. The test statistics can take a longer time to compute under large hyperparameters M, N, P, Q, D. We will discuss simulation results of this test in Section 5.1 for relatively smaller model hyperparameters, and show a real use case for data application in Section 6. We leave the problem of scaling this test to higher-dimensional contexts for future work.

4.3 High Spatial Dimension Asymptotics

The previous section presents the asymptotic normality of the MARAC estimators under a *fixed* matrix dimensionality S. In this section, we relax this assumption and establish the convergence rate of the MARAC estimators when $S, T \to \infty$. For technical reasons, we assume that the covariance of $\mathbf{vec}\,(\mathbf{E}_t)$ is known but allows for an arbitrary covariance $\mathbf{\Sigma}$. To establish the convergence rate, we make several additional assumptions.

Assumption 3. The spatial kernel function $k(\cdot, \cdot)$ can be decomposed into the product of a row kernel $k_1(\cdot, \cdot)$ and a column kernel $k_2(\cdot, \cdot)$ that satisfies $k((u, v), (s, t)) = k_1(u, s)k_2(v, t)$. Both k_1, k_2 have their eigenvalues decaying at a polynomial rate: $\lambda_j(k_1) \asymp j^{-r_0}, \lambda_j(k_2) \asymp j^{-r_0}$ with $r_0 \in (1/2, 2)$.

Assumption 4. The spatial locations of the rows and columns of $\mathbf{X}_t$ are sampled independently from a uniform distribution on $[0, 1]$.

Assumption 3 elicits a simple eigen-spectrum characterization of the spatial kernel $k(\cdot, \cdot)$, whose eigenvalue can be written as $\lambda_i(k_1)\lambda_j(k_2)$. Also, the Gram matrix $\mathbf{K}$ is separable, i.e. $\mathbf{K} = \mathbf{K}_2 \otimes \mathbf{K}_1$ and all eigenvalues of $\mathbf{K}$ have the form of $\rho_i(\mathbf{K}_1)\rho_j(\mathbf{K}_2)$, where $\mathbf{K}_1 \in$

$\mathbb{R}^{M\times M}, \mathbf{K}_2 \in \mathbb{R}^{N\times N}$ are the Gram matrix for the kernel k_1, k_2, respectively. The separability of the kernel can accommodate the grid structure of the spatial locations.

Under Assumption 4, we further have $\rho_i(\mathbf{K}_1) \to M\lambda_i(k_1)$ and $\rho_j(\mathbf{K}_2) \to N\lambda_j(k_2)$, as $M, N \to \infty$. 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 4 to non-uniform sampling, but here, we stick to this more straightforward setting. With these assumptions, we are ready to present the main result in Theorem 3.

Theorem 3 (Asymptotics for High-Dimensional MARAC)**.** *Assume that Assumptions 1, 3 and 4 hold and* $\mathbf{X}_t$ *is generated by the MARAC*(P, Q) *model* (2.1) *with* $\mathbf{vec}(\mathbf{E}_t) \overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ *and* $\mathbf{\Sigma}$ *is known. With* $S, T \to \infty$ *(*D *is fixed) and* $S \log S/T \to 0$*, and assume that:*

1. $M = O(\sqrt{S}), N = O(\sqrt{S})$*;*
2. $\gamma_S := \lambda/S \to 0$ *and* $\gamma_S \cdot S^{r_0} \to C_1$ *as* $S \to \infty$*, with* $0 < C_1 \leq \infty$*;*
3. $\underline{\rho}(\mathbf{\Sigma}^*_{\mathbf{x},\mathbf{x}} - (\mathbf{\Sigma}^*_{\mathbf{z},\mathbf{x}})^\top (\mathbf{\Sigma}^*_{\mathbf{z},\mathbf{z}})^{-1} \mathbf{\Sigma}_{\mathbf{z},\mathbf{x}}) = c_{0,S} > 0$ *as* $S, T \to \infty$*, where* $\mathbf{\Sigma}^*_{\mathbf{x},\mathbf{x}}, \mathbf{\Sigma}^*_{\mathbf{z},\mathbf{z}}, \mathbf{\Sigma}^*_{\mathbf{z},\mathbf{x}}$ *are* $\mathrm{Var}(\mathbf{x}_t)$, $\mathrm{Var}(\mathbf{z}_t)$ *and* $\mathrm{Cov}(\mathbf{z}_t, \mathbf{x}_t)$*, respectively.* $c_{0,S}$ *is a constant that only relates to* S*;*
4. *For any* S*, we have* $0 < \underline{\rho}(\mathbf{K}) < \bar{\rho}(\mathbf{K}) \leq C_0$*, where* C_0 *is a finite constant;*
5. $\bar{\rho}(\mathbf{\Sigma})/\underline{\rho}(\mathbf{\Sigma}) \leq C_1 < \infty$*, where* C_1 *is a constant, and* $\bar{\rho}(\mathbf{\Sigma}) = c_{1,S}$*,*

Then we have:

$$\frac{1}{\sqrt{PS}} \sqrt{\sum_{p=1}^{P} \left\| \widehat{\mathbf{B}}_p \otimes \widehat{\mathbf{A}}_p - \mathbf{B}^*_p \otimes \mathbf{A}^*_p \right\|_{\mathrm{F}}^2} \lesssim O_P\left(\sqrt{\frac{C_g \cdot \gamma_S}{c_{0,S} \cdot S}}\right) + O_P\left(\sqrt{\frac{c_{1,S} \cdot D}{c_{0,S} \cdot TS}}\right), \tag{4.18}$$

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

$$\sqrt{(TS)^{-1}\sum_{t=1}^{T}\left\|\sum_{q=1}^{Q}\left(\widehat{\boldsymbol{\mathcal{G}}}_q-\boldsymbol{\mathcal{G}}_q^*\right)\bar{\times}\mathbf{z}_{t-q}\right\|_{\mathrm{F}}^2}$$
$$\lesssim O_P\left(\frac{\sqrt{c_{1,S}\cdot\gamma_S^{-1/2r_0}}}{\sqrt{T}\sqrt[4]{S}}\right)+O_P(\sqrt{\gamma_S})+O_P\left(\frac{1}{\sqrt{S}}\right)+O_P\left(\sqrt{\frac{c_{1,S}}{T}}\right)+O_P\left(\frac{\sqrt{c_{1,S}\cdot\gamma_S^{-1}}}{\sqrt{TS}}\right). \tag{4.19}$$

In Theorem 3, (4.18) gives the error bound of the autoregressive coefficients and (4.19) gives the error bound of the prediction made by the auxiliary time series, which contains the functional parameter estimators. As a special case of (4.18) where $\gamma_S = 0$ and S is fixed, the convergence rate for the autoregressive coefficients is $O_P(T^{-1/2})$, which reproduces the result in Theorem 2. For the discussion below, we use AR_{err} and AC_{err} as acronyms for the quantity on the left-hand side of (4.18) and (4.19).

Remark 2 (Optimal Choice of λ and Phase Transition)**.** According to our proof, the error bound (4.19) can be decomposed into the sum of:

- nonparametric error: $O_P\left(\frac{\sqrt{c_{1,S}\cdot\gamma_S^{-1/2r_0}}}{\sqrt{T}\sqrt[4]{S}}\right)+O_P(\sqrt{\gamma_S})$,
- autoregressive error: $O_P\left(\sqrt{\gamma_S}\right)+O_P\left(S^{-1/2}\right)+O_P\left(\sqrt{\frac{c_{1,S}}{T}}\right)+O_P\left(\frac{\sqrt{c_{1,S}\cdot\gamma_S^{-1}}}{\sqrt{TS}}\right)$,

where the autoregressive error stems from the estimation error of $\widehat{\mathbf{B}}_p\otimes\widehat{\mathbf{A}}_p$. In our model, if there is no autoregressive error, the optimal tuning parameter satisfies $\gamma_S \asymp (TS^{1/2}c_{1,S}^{-1})^{-2r_0/(2r_0+1)}$. Compared to the classical semi-parametric regression result (Cui et al., 2018), our optimal rate does not exactly scale with the number of data points TS, but scales with $T\sqrt{S}$. This is a special result for matrix-shaped data. Also, the optimal rate depends upon the correlations

among the errors, which is a special result for spatial data. Notably, under $S \log S/T \to 0$, the autoregressive error dominates the nonparametric error.

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

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

Table 1: Summary of optimal tuning parameter γ_S and estimators error following (4.18) and (4.19), under the assumption that $c_{0,S} \geq c_0 > 0$, $c_{1,S} \leq c_1 < \infty$, for all S and $S = T^c$ for some constant $0 < c < 1$ such that $S \log S/T \to 0$. AR_{err} and AC_{err} are the quantity on the left-hand side of (4.18) and (4.19).

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

In (4.18), the constant $c_{0,S}$ appears in the error bound of the autoregressive term. This constant characterizes the spatial correlation of the matrix time series $\mathbf{X}_t$, conditioning on the auxiliary vector time series $\mathbf{z}_t$, and can vary across different assumptions made on the covariances of $\mathbf{E}_t$ and $\boldsymbol{\nu}_t$. In Table 1, we assume that $c_{0,S} \geq c_0 > 0$ for some universal constant c_0. Unfortunately, in practice, it is common to have $c_{0,S} \to 0$ as $S \to \infty$, which makes the autoregressive coefficient converge at a slower rate but does not affect the functional parameter convergence. We leave the constant $c_{0,S}$ here in (4.18) to give a general result and leave the characterization of $c_{0,S}$ under specific assumptions for future works.

5. Simulation Experiments

5.1 Consistency, Convergence Rate and Hypothesis Testing

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

The evaluation metric is the rooted mean squared error (RMSE), defined as $\text{RMSE}(\widehat{\boldsymbol{\Theta}}) = \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\text{F}}/\sqrt{d(\boldsymbol{\Theta}^*)}$, where $d(\boldsymbol{\Theta}^*)$ is the number of elements in $\boldsymbol{\Theta}^*$. We consider $\boldsymbol{\Theta} \in \{\mathbf{B}_1 \otimes \mathbf{A}_1, \boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r, \boldsymbol{\mathcal{G}}_1, \boldsymbol{\mathcal{G}}_2, \boldsymbol{\mathcal{G}}_3\}$ and we report the average RMSE for $\boldsymbol{\mathcal{G}}_1, \boldsymbol{\mathcal{G}}_2, \boldsymbol{\mathcal{G}}_3$. The dataset

is configured with $M \in \{5, 10, 20, 40\}$ and $N = M$. For each M, we train the MARAC model with $P = Q = 1$ over $T_{\text{train}} \in \{1, 5, 10, 20, 40, 80, 160\} \times 10^2$ frames of the matrix time series and choose the tuning parameter λ based on the prediction RMSE over a held-out validation set with $T_{\text{val}} = T_{\text{train}}/2$. We validate the prediction performance over a $5,000$-frame testing set. All results are reported with 20 repetitions in Figure 3.

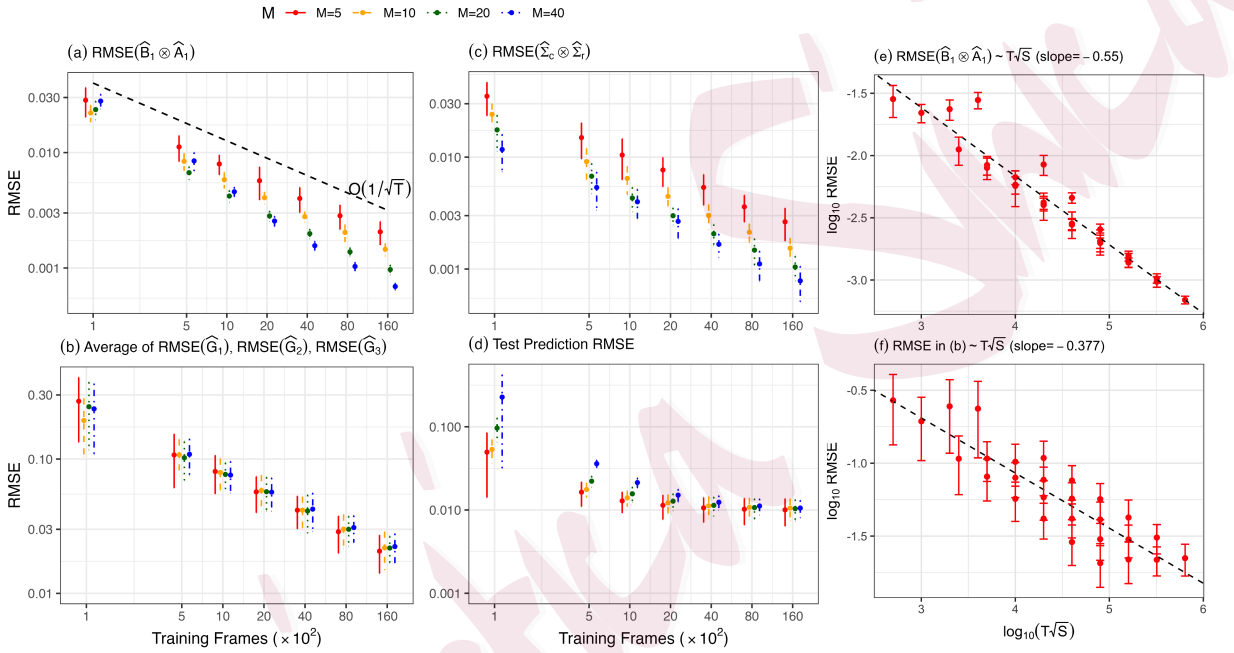

Figure 3: Panel (a), (b), (c) show the RMSE of the penalized MLE of the MARAC model. Panel (d) shows the testing set prediction RMSE subtracted by 1, where 1 is the noise variance of the simulated time series. Panels (a)-(d) have both axes plotted in $\log_{10}$ scale. (e) and (f) are the RMSE of the autoregressive parameters and auxiliary covariates parameters under different $T\sqrt{S}$, plotted with both axes in $\log_{10}$ scale together with a fitted linear regression line.

The result shows that all model estimators are consistent. The convergence rate, under a fixed spatial dimensionality, is close to $1/\sqrt{T}$ (the black line in panel (a) shows a reference line of $O(1/\sqrt{T})$), echoing the result in Theorem 2. As the spatial dimensionality S increases, the RMSE for $\widehat{\mathbf{B}}_1 \otimes \widehat{\mathbf{A}}_1$ becomes even smaller, echoing the result in (4.18) and Table 1. The RMSE of the nonparametric estimators $\widehat{g}_1, \widehat{g}_2, \widehat{g}_3$, under a fixed spatial dimensionality,

also decay at a rate of $1/\sqrt{T}$, echoing the result in Theorem 2 as well. The RMSE of the covariance matrix estimator $\widehat{\mathbf{\Sigma}}_c \otimes \widehat{\mathbf{\Sigma}}_r$ suggests that it is consistent, confirming the result of Proposition 2 and showing a convergence rate similar to $\widehat{\mathbf{B}}_1 \otimes \widehat{\mathbf{A}}_1$, though we did not provide the exact convergence rate theoretically.

In this simulation, we fix the variance of each element of $\mathbf{vec}\,(\mathbf{E}_t)$ to be unity. Therefore, the optimal testing set prediction RMSE should be unity. When plotting the test prediction RMSE in (d), we subtract 1 from all RMSE results, and thus, the RMSE should be interpreted as the RMSE for the *signal* part of the matrix time series. The test prediction RMSE for all cases converges to zero, and for matrices of higher dimensionality, we typically require more training frames to reach the same prediction performance.

To validate the theoretical result of the high-dimensional MARAC in Theorem 3, we also plot the RMSE of $\widehat{\mathbf{B}}_1 \otimes \widehat{\mathbf{A}}_1$ and $\widehat{g}_1, \widehat{g}_2, \widehat{g}_3$ against $T\sqrt{S}$ in panel (e) and (f) of Figure 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 setup. It also shows that the functional parameter's convergence rate is around $(T\sqrt{S})^{-3/8}$, which coincides with our simulation setup where $r_0 \approx 3/4$ and the theoretical result in the last row of Table 1.

We also conduct finite-sample simulations for the hypothesis testing discussed in Corollary 1. We set $(M, N) \in \{(5,5), (10,10)\}$, $P = 1$ and test for both the scenarios of $Q = 0$ (H_0 is true) and $Q = 1$ (H_0 is False). For $Q = 1$, we further introduce a scaling factor η that controls the scale of $\boldsymbol{\mathcal{G}}_1^*$, and thus a smaller η makes the alternative hypothesis less distinguishable from the null hypothesis. We run the simulation for different sample sizes T with 1000 repetitions and report the Type I Error rate and power of the test in Figure 4.

In Corollary 1, we lower bound the degrees of freedom of the chi-square distribution by

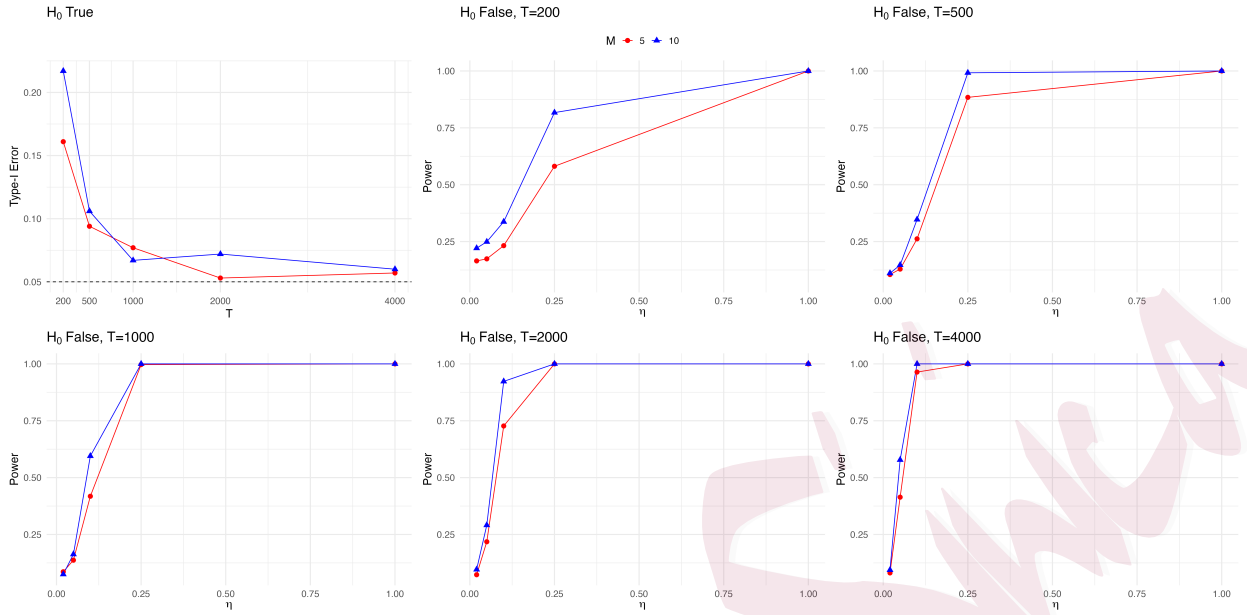

Figure 4: Specification test simulation results. The top left panel shows the Type-I Error rate out of 1000 repetitions across different sample sizes T, when H_0 is True ($Q = 0$). The remaining panels show the power of the test, under different T, when the alternative hypothesis is true ($Q = 1$) but with varying scaling factor η (smaller η means the norm of $\boldsymbol{\mathcal{G}}_1^*$ is smaller).

$MNQD-1$ and use it to determine the critical value, and thus our tests could lead to higher Type-I Error and power. However, this lower bound makes very little difference, even given the smaller matrix setup here, and we do see that the Type-I error rate reaches the specified level ($\alpha = 0.05$) with relatively larger sample size T. The power of the test approaches unity as the sample size T grows or the alternative hypothesis becomes more distinguishable (larger η). These results justify the applications of the proposed test in practice, and we will be using it in our data application in Section 6.

5.2 Lag Selection Consistency

In Section 3.2, we propose to select the lag parameters P and Q of the MARAC model using information criteria such as AIC and BIC. To validate the consistency of these model selection

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

	$T = 1000$	$T = 2000$	$T = 4000$	$T = 8000$
AIC	$(.54, .99, .53)$	$(.55, .97, .53)$	$(.59, .96, .55)$	$(.65, .94, .59)$
BIC	$(1.00, .09, .09)$	$(.99, .56, .56)$	$(.97, .97, .94)$	$(.96, .99, .95)$

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

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

5.3 Comparison with Alternative Methods

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

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

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

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

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

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

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

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

$$\mathbf{X}_t - \sum_{p=1}^{P} \widehat{\mathbf{A}}_p \mathbf{X}_{t-p} \widehat{\mathbf{B}}_p^\top = \sum_{q=1}^{Q} \boldsymbol{\mathcal{G}}_q \bar{\times} \mathbf{z}_{t-q} + \mathbf{E}_t, \mathbf{vec}\,(\mathbf{E}_t) \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2 \mathbf{I}), \tag{5.20}$$

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

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

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

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

The results of the average prediction RMSE obtained from the 20 repeated runs are plotted

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

6. Application to Global Total Electron Content Forecast

For real data applications, we consider the problem of predicting the global total electron content (TEC) distribution, which we briefly introduce in Section 1. The TEC data we use is the IGS (International GNSS Service) TEC data, which are freely available from the National Aeronautics and Space Administration (NASA) Crustal Dynamics Data Information System (Hernández-Pajares et al., 2009). The spatial-temporal resolution of the data is $2.5°(\text{latitude}) \times 5°(\text{longitude}) \times 15(\text{minutes})$. We use whole-month data for September 2017, a matrix time series with $T = 2880$ and $M = 71, N = 73$. We use the 15-minute resolution IMF Bz and Sym-H time series for the auxiliary covariates, which are parameters related

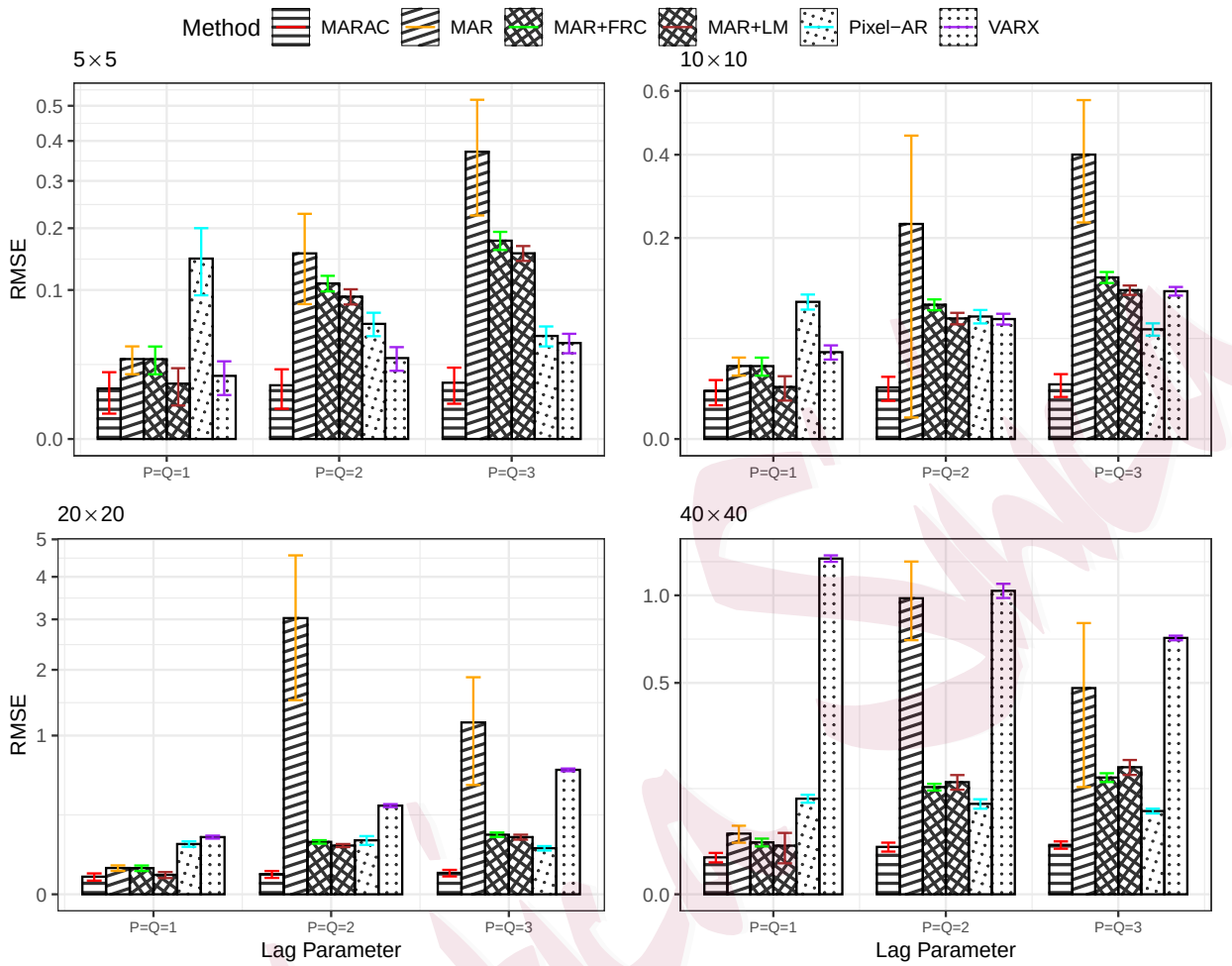

Figure 5: Testing set prediction RMSE comparison across six competing methods on the matrix autoregression task. Four panels correspond to four different matrix dimensionality (labeled on the top-left corner of each panel). Test prediction RMSE is subtracted by 1 for better visualization, where 1 is the noise variance of the simulated data. Error bar shows 95% CI of the 20 repeated runs. For better visualization, we rearrange the spacing between ticks along the y-axis using a square root transformation.

to the near-Earth magnetic field and plasma (Papitashvili et al., 2014). 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:

$$\Delta \mathrm{TEC}_{t+h} = \sum_{p=1}^{P} \mathbf{A}_p \Delta \mathrm{TEC}_{t-p} \mathbf{B}_p^\top + \sum_{q=1}^{Q} \boldsymbol{\mathcal{G}}_q \bar{\times} \Delta \mathbf{z}_{t-q} + \mathbf{E}_t, \tag{6.21}$$

where h is the forecast latency time and $\Delta\mathbf{z}_t = \mathbf{z}_t - \mathbf{z}_{t-1}$ includes the change of IMF Bz, and Sym-H indices from time $t-1$ to t. We chose to forecast the $\Delta\text{TEC}_t = \text{TEC}_t - \text{TEC}_{t-1}$ series and use $\Delta\mathbf{z}_t$ instead of the raw $\mathbf{z}_t$ series as the auxiliary covariates to satisfy the joint stationarity condition in Theorem 1. To ensure we have better estimator convergence and valid inference, we downsample each matrix from $(71, 73)$ to $(12, 12)$ via averaging each local 6×6 patch.

We consider the forecasting scenario with $h \in \{4, 8, 12, \ldots, 72\}$, corresponding to making forecasts from 1 hour to 18 hours ahead. For each h, we fit our MARAC(P, Q) model following (6.21) with $1 \leq P \leq 5$ and $1 \leq Q \leq 3$. As a comparison, we also fit the MAR model with $1 \leq P \leq 5$ and the MAR+LM model with $1 \leq P \leq 5$ and $1 \leq Q \leq 3$, see the definition of MAR+LM model in (5.20). 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 selected for all models based on the BIC.

We report the results in Figure 6. From the left panel, it is clear that the MARAC model is consistently outperforming the other two competing methods across all forecast horizons. The addition of the auxiliary covariates improves the prediction accuracy, and this is also confirmed in the right panel, where all specification tests reject the null of $Q = 0$.

To visualize the difference that the auxiliary covariates can make, in Figure 7, we fit a MARAC$(1, 1)$ model to predict the ΔTEC_{t+12}, namely the ΔTEC 3 hours later. For better visibility, we only downsample the data to 24×24. To further distinguish the predictions made by different models, we take the sum of 90 consecutive predictions/ground truth, and plot the results. It is clear from the results that the MARAC prediction tracks the

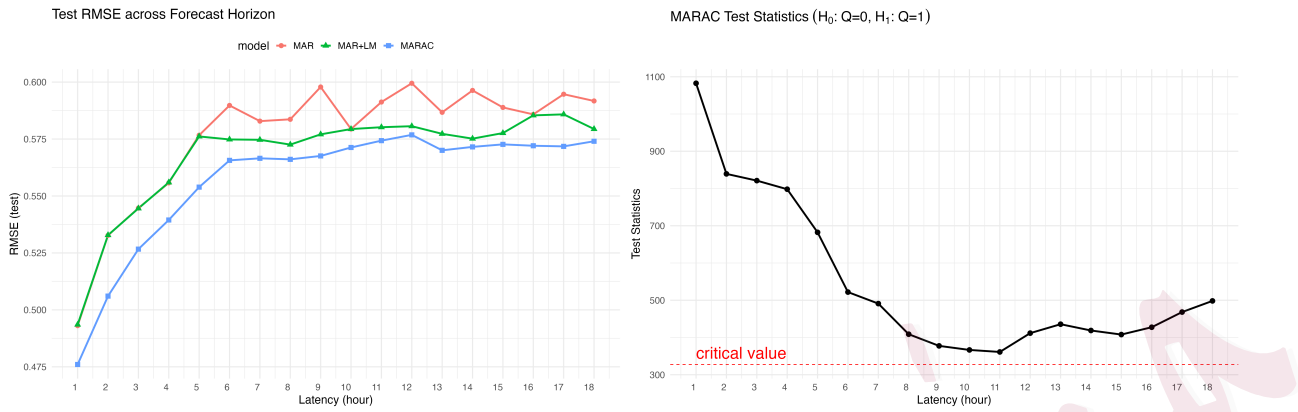

Figure 6: TEC prediction results. Left: test set prediction RMSE across three methods. Right: MARAC model test statistics, with model fitted with $Q = 1$ and P selected by BIC.

target better than the competing methods, and the auxiliary covariates contribute to better predictions near the equatorial region (middle band of the plot), which is also the primary region of scientific interest. We believe that our method can help domain scientists to determine if scalar time series can predict spatial responses in similar application contexts.

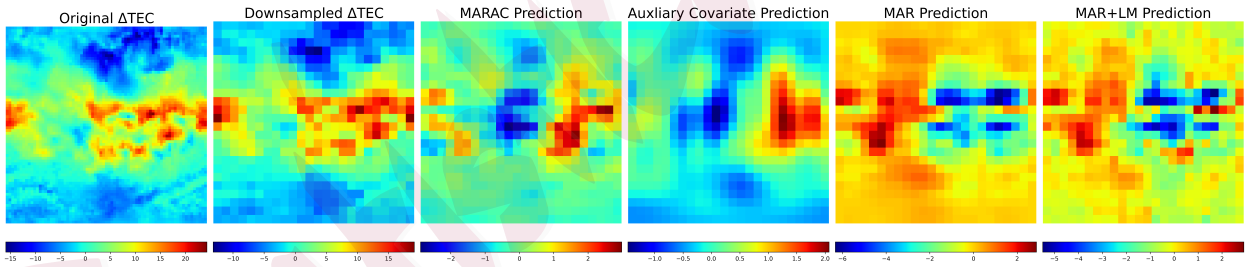

Figure 7: Example visualization of ΔTEC prediction with a forecast horizon of 3 hours. Left to right: original ΔTEC of size 71×73; downsampled ΔTEC of size 24×24; MARAC$(1, 1)$ prediction result; Auxiliary covariate term prediction result from MARAC$(1, 1)$; MAR(1) prediction result; MAR+LM$(1, 1)$ prediction result. The results are the sum across 90 consecutive frames.

7. Summary

This paper proposes a new methodology for spatial-temporal matrix autoregression with non-spatial exogenous vector covariates. The model has an autoregressive component with bilinear transformations on the lagged matrix predictors and an additive auxiliary covariate component with a tensor-vector product between a tensor coefficient and the lagged vector covariates. We propose a penalized MLE estimation approach with a squared RKHS norm penalty and establish the estimator asymptotics under fixed and high matrix dimensionality. The model efficacy has been validated using both numerical experiments and an application to the global TEC forecast.

The application of our model can be extended to other spatial data with exogenous, non-spatial predictors and is not restricted to matrix-valued data but can be generalized to the tensor setting and potentially data without a grid structure or containing missing data. Furthermore, our model nests a simpler model that does not include the autoregressive term, i.e., $P = 0$, and thus can be applied to matrix-on-scalar regression with spatial data. Also, it is a natural extension of our paper to consider the case where D, the dimension of the auxiliary covariates, grows together with M and N, and thus enables the modeling of high-dimensional auxiliary covariates for matrix/tensor response. We leave the discussions for these setups to future research.

Supplementary Materials

The supplemental material contains details of the alternating minimization algorithm, technical proofs of all theorems and propositions of the paper, additional details of the simulation experiments, and the approximated estimating algorithm based on kernel truncation. Our

code is available at `https://github.com/husun0822/MARAC`.

Acknowledgements

The authors thank Shasha Zou, Zihan Wang, and Yizhou Zhang for helpful discussions on the TEC data. YC acknowledges support from NSF DMS 2113397, NSF PHY 2027555, NASA Federal Award No. 80NSSC23M0192, and No. 80NSSC23M0191.

References

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected papers of Hirotugu Akaike*, pp. 199–213. Springer.

Attouch, H., J. Bolte, and B. F. Svaiter (2013). 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.

Banerjee, A., I. S. Dhillon, J. Ghosh, S. Sra, and G. Ridgeway (2005). Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning Research 6*(9), 1345–1382.

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

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

Chen, E. Y. and J. Fan (2023). Statistical Inference for High-Dimensional Matrix-Variate Factor Models. *Journal of the American Statistical Association 118*(542), 1038–1055.

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

Cheng, G. and Z. Shang (2015). Joint Asymptotics for Semi-nonparametric Regression Models with Partially Linear Structure. *The Annals of Statistics 43*, 1351–1390.

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

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

Cressie, N. and C. K. Wikle (2015). *Statistics for Spatio-Temporal Data*. John Wiley & Sons.

Cui, W., H. Cheng, and J. Sun (2018). An RKHS-based Approach to Double-Penalized Regression in High-dimensional Partially Linear Models. *Journal of Multivariate Analysis 168*, 201–210.

Dong, M., L. Huang, X. Wu, and Q. Zeng (2020, feb). Application of Least-Squares Method to Time Series Analysis for 4dpm Matrix. *IOP Conference Series: Earth and Environmental Science 455*(1), 012200.

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

Gao, Z. and R. S. Tsay (2023). A Two-way Transformed Factor Model for Matrix-Variate Time Series. *Econometrics and Statistics 27*, 83–101.

Gao, Z. and R. S. Tsay (2025). Denoising and Multilinear Projected-Estimation of High-Dimensional Matrix-Variate Factor Time Series. *IEEE Transactions on Information Theory*, in press.

Gu, C. (2013). *Smoothing Spline ANOVA models, 2nd edition*. Springer, New York.

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

Guhaniyogi, R., S. Qamar, and D. B. Dunson (2017). Bayesian Tensor Regression. *The Journal of Machine Learning Research 18*(1), 2733–2763.

Guo, S., Y. Wang, and Q. Yao (2016, 10). High-dimensional and Banded Vector Autoregressions. *Biometrika 103*(4),

889–903.

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

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition*. Springer, New York.

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

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

Hsu, N.-J., H.-C. Huang, and R. S. Tsay (2021). Matrix Autoregressive Spatio-Temporal Models. *Journal of Computational and Graphical Statistics 30*(4), 1143–1155.

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

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

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

Koltchinskii, V. and E. Giné (2000). Random Matrix Approximation of Spectra of Integral Operators. *Bernoulli 6*(1), 113–167.

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

Li, X., D. Xu, H. Zhou, and L. Li (2018). Tucker Tensor Regression and Neuroimaging Analysis. *Statistics in Biosciences 10*(3), 520–545.

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

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

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

Luo, Y. and A. R. Zhang (2024). Tensor-on-tensor Regression: Riemannian Optimization, Over-parameterization, Statistical-Computational Gap and Their Interplay. *The Annals of Statistics 52*(6), 2583–2612.

Lyu, X., W. W. Sun, Z. Wang, H. Liu, J. Yang, and G. Cheng (2019). Tensor Graphical Model: Non-convex Optimization and Statistical Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence 42*(8), 2024–2037.

Papadogeorgou, G., Z. Zhang, and D. B. Dunson (2021). Soft Tensor Regression. *The Journal of Machine Learning Research 22*, 219–1.

Papitashvili, N., D. Bilitza, and J. King (2014). OMNI: a Description of Near-Earth Solar Wind Environment. *40th COSPAR Scientific Assembly 40*, C0–1.

Rabusseau, G. and H. Kadri (2016). Low-rank Regression with Tensor Responses. *Advances in Neural Information Processing Systems 29*.

Schölkopf, B., R. Herbrich, and A. J. Smola (2001). A Generalized Representer Theorem. In *International Conference on Computational Learning Theory*, pp. 416–426. Springer.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461–464.

Shang, Z. and G. Cheng (2013). Local and Global Asymptotic Inference in Smoothing Spline Models. *The Annals of Statistics 41*, 2608–2638.

Shang, Z. and G. Cheng (2015). Nonparametric Inference in Generalized Functional Linear Models. *The Annals of Statistics 43*, 1742–1773.

Shen, B., W. Xie, and Z. Kong (2022). 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.

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

Sun, H., Z. Hua, J. Ren, S. Zou, Y. Sun, and Y. Chen (2022). Matrix Completion Methods for the Total Electron Content Video Reconstruction. *The Annals of Applied Statistics 16*(3), 1333–1358.

Sun, H., W. Manchester, M. Jin, Y. Liu, and Y. Chen (2023). Tensor Gaussian Process with Contraction for Multi-Channel Imaging Analysis. In *International Conference on Machine Learning*, pp. 32913–32935. PMLR.

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

Tsiligkaridis, T., A. O. Hero III, and S. Zhou (2013). On Convergence of Kronecker Graphical Lasso Algorithms. *IEEE Transactions on Signal Processing 61*(7), 1743–1755.

Tzeng, S. and H.-C. Huang (2018). Resolution Adaptive Fixed Rank Kriging. *Technometrics 60*(2), 198–208.

van Zanten, J. and A. W. van der Vaart (2008). Reproducing Kernel Hilbert Spaces of Gaussian Priors. In *Pushing the Limits of Contemporary Statistics: Contributions in honor of Jayanta K. Ghosh*, pp. 200–222. Institute of Mathematical Statistics.

Wang, D., X. Liu, and R. Chen (2019). Factor Models for Matrix-valued High-dimensional Time Series. *Journal of Econometrics 208*(1), 231–248.

Wang, D., Y. Zheng, and G. Li (2024). High-Dimensional Low-rank Tensor Autoregressive Time Series Modeling. *Journal of Econometrics 238*(1), 105544.

Wang, D., Y. Zheng, H. Lian, and G. Li (2022). High-dimensional Vector Autoregressive Time Series Modeling via Tensor Decomposition. *Journal of the American Statistical Association 117*(539), 1338–1356.

Wang, X., H. Zhu, and A. D. N. Initiative (2017). Generalized Scalar-on-Image Regression Models via Total Variation.

Journal of the American Statistical Association 112(519), 1156–1168.

Wang, Z., S. Zou, L. Liu, J. Ren, and E. Aa (2021). 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.

Williams, C. K. and C. E. Rasmussen (2006). *Gaussian Processes for Machine Learning*, Volume 2. MIT press Cambridge, MA.

Xiao, H., Y. Han, R. Chen, and C. Liu (2022). Reduced Rank Autoregressive Models for Matrix Time Series. *Journal of Business and Economic Statistics*.

Yang, Y., Z. Shang, and G. Cheng (2020). Non-asymptotic Analysis for Nonparametric Testing. In *33rd Annual Conference on Learning Theory*, pp. 1–47. ACM.

Younas, W., M. Khan, C. Amory-Mazaudier, P. O. Amaechi, and R. Fleury (2022). 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.

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

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

Zhou, S. (2014). GEMINI: Graph Estimation with Matrix Variate Normal Instances. *The Annals of Statistics 42*(2), 532–562.

Hu Sun

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

E-mail: husun@umich.edu

Zuofeng Shang

Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07103, USA

E-mail: zshang@njit.edu

Yang Chen

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

E-mail: ychenang@umich.edu