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

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

May 26, 2023

Abstract

In this paper, we propose a new model for forecasting time series data distributed on a matrix-shaped spatial grid, using the historical spatio-temporal data together with auxiliary vector-valued time series data. We model the matrix time series as an auto-regressive process, where a future matrix is jointly predicted by the historical values of the matrix time series as well as an auxiliary vector time series. The matrix predictors are associated with row/column-specific autoregressive matrix coefficients that map the predictors to the future matrices via a bi-linear transformation. The vector predictors are mapped to matrices by taking mode product with a 3D coefficient tensor. Given the high dimensionality of the tensor coefficient and the underlying spatial structure of the data, we propose to estimate the tensor coefficient by estimating one functional coefficient for each covariate, with 2D input domain, from a Reproducing Kernel Hilbert Space. We jointly estimate the autoregressive matrix coefficients and the functional coefficients under a penalized maximum likelihood estimation framework, and couple it with an alternating minimization algorithm. Large sample asymptotics of the estimators are established and performances of the model are validated with extensive simulation studies and a real data application to forecast the global total electron content distributions.

Keywords: Matrix Autoregressive Model, 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, geophysics and environmental science, where scientists are interested in modeling the joint dynamics of data observed on a 2D grid across time. In this work, we focus specifically on the data whose 2D grid contains spatial or geographical information of the individual observations. As a concrete example for such spatio-temporal data, we visualize the global total electron content (TEC) distribution in Figure 1 from geophysics. Total electron content 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 181×361 matrix, distributed on a 1°-latitude-by-1°-longitude spatio-temporal grid. The statistical challenge here is to make forecasts of the future TEC maps with the historical, high-dimensional TEC time series.

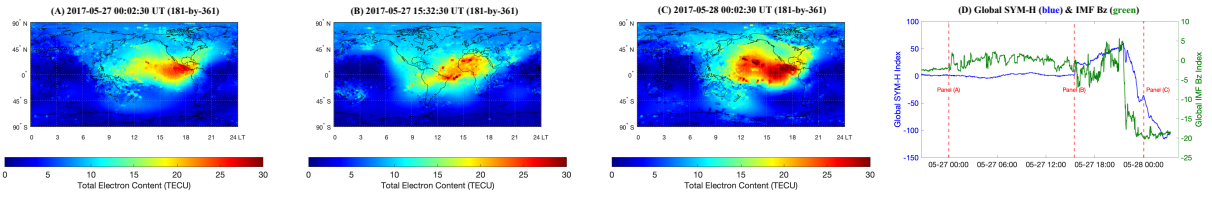

Figure 1: *Example of matrix time series & auxiliary vector time series. Panel (A)-(C) show the global Total Electron Content (TEC) distribution at three timestamps on the latitude-local time grid (source: the VISTA TEC database [Sun et al., 2023]). Panel (D) plots the auxiliary vector time series, including the global SYM-H & IMF Bz index.*

Additional modeling challenge is brought up by the presence of auxiliary vector time series covariates for the matrix time series. In Figure 1(D), we plot the global SYM-H

index and the IMF Bz index, both measuring the geomagnetic activity caused by the solar eruptions that can finally impact the Earth's global TEC distribution. These covariates carry additional information related to the dynamics of the matrix time series data. To see this, Figure 1(A) and 1(B) have similar auxiliary covariates and TEC distributions, but 1(C) has a dramatic decrease of both indices and a much higher TEC near the equator.

Adding the auxiliary covariates not only benefits forecasting but also enables domain scientists to understand the interplay between different data modalities and how the information carried by the centralized non-spatial data (e.g. SYM-H and IMF Bz) disseminates to the decentralized spatial data (e.g. TEC maps). Therefore, a statistical methodology that could benefit these modeling contexts is much needed.

In this paper, we consider the problem of forecasting the matrix time series using both lagged matrix predictors and lagged auxiliary vector predictors. We denote the matrix time series as $\mathbf{X}_1, \ldots, \mathbf{X}_T \in \mathbb{R}^{M \times N}$, and the auxiliary vector time series as $\mathbf{z}_1, \ldots, \mathbf{z}_T \in \mathbb{R}^D$. An additive model using the P lagged matrices and Q lagged vectors can be written as:

$$\mathbf{X}_t = \mathbf{f}_1(\mathbf{X}_{t-1}) + \cdots + \mathbf{f}_P(\mathbf{X}_{t-P}) + \mathbf{g}_1(\mathbf{z}_{t-1}) + \cdots + \mathbf{g}_Q(\mathbf{z}_{t-Q}) + \mathbf{E}_t,$$

where $\mathbf{f}_p(.) : \mathbb{R}^{M \times N} \mapsto \mathbb{R}^{M \times N}, \mathbf{g}_q(.) : \mathbb{R}^D \mapsto \mathbb{R}^{M \times N}$ can be arbitrary functions with matrix outputs and $\mathbf{E}_t$ is the error matrix.

A straightforward but undesirable approach is to vectorize the matrices and concatenate all vector predictors together to build a Vector Autoregressive (VAR) model [Stock and Watson, 2001]. This approach is undesirable because the matrix structure of the data, which may correspond to spatial/temporal/other dependency in practice, is neglected.

There are multiple threads of literature concerning regression models with matrix-/tensor-valued covariates. The first thread is the *scalar-on-tensor* regression models where the response variable is scalar and the covariates are n-mode ($n \geq 2$) tensors. Under our

modeling context, such model can be written as *element-wise* regression model:

$$\mathbf{X}_t(i,j) = \sum_{p=1}^{P} \langle \mathbf{W}_p, \mathbf{X}_{t-p} \rangle + \sum_{q=1}^{Q} \boldsymbol{\eta}_q^\top \mathbf{z}_{t-p} + \mathbf{E}_t(i,j), \tag{1}$$

where $\langle ., . \rangle$ is the tensor inner product [Kolda and Bader, 2009], and $\mathbf{W}_1, \ldots, \mathbf{W}_P \in \mathbb{R}^{M \times N}$, $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_Q \in \mathbb{R}^D$ are model parameters. To reduce the complexity of the model, $\mathbf{W}_p$ is assumed to be low-rank, following either the CANDECOMP/PARAFAC (CP) decomposition [Hung and Wang, 2013, Zhou et al., 2013] or the Tucker decomposition [Li et al., 2018, Kossaifi et al., 2020]. However, these methods are built for *scalar* responses and independent samples and are not directly applicable to matrix time series responses.

Directly using the matrix time series $\mathbf{X}_t$ as the responses leads to the second thread of literature on matrix/tensor autoregression model. Similar to (1), such model features a tensor regression coefficient with low-rankness [Lock, 2018]. For matrix autoregression [Chen et al., 2021], in particular, the matrix response $\mathbf{X}_t$ is modeled as:

$$\mathbf{X}_t = \mathbf{A}_1 \mathbf{X}_{t-1} \mathbf{B}_1^\top + \mathbf{A}_2 \mathbf{X}_{t-2} \mathbf{B}_2^\top + \cdots + \mathbf{A}_P \mathbf{X}_{t-P} \mathbf{B}_P^\top + \mathbf{E}_t, \tag{2}$$

where the model parameters are $\mathbf{A}_1, \ldots, \mathbf{A}_P \in \mathbb{R}^{M \times M}$ and $\mathbf{B}_1, \ldots, \mathbf{B}_P \in \mathbb{R}^{N \times N}$. Every pair of $(\mathbf{A}_p, \mathbf{B}_p)$ transforms the lag-p predictor $\mathbf{X}_{t-p}$ via *bi-linear* transformation. Compared with the VAR model, the dimensionality of the "coefficient matrix" for every $\mathbf{X}_{t-p}$ is reduced significantly from $M^2 \times N^2$ to $M^2 + N^2$. For interpretations of such bi-linear form, we refer our reader to Chen et al. [2021] and we will provide concrete interpretations of the parameters in our real data experiment. An extension of (2) to the spatio-temporal time series forecast is proposed in Hsu et al. [2021] and we will also use (2) as the fundamental building block for our method but extend the framework to incorporate the auxiliary vector covariates.

The difficulty of incorporating the auxiliary vector predictors is that the auxiliary data are *vectors* while the responses are *matrices*. The literature related to this topic is more

generally known as the *tensor-on-scalar* regression model [Rabusseau and Kadri, 2016, Sun and Li, 2017, Li and Zhang, 2017, Lock and Li, 2018]. Typically, the tensor responses are factorized into mode-specific factors with the factors being determined by the mode-specific vector predictors. Under our modeling context, by concatenating $\mathbf{X}_1, \ldots, \mathbf{X}_T$ into a spatio-temporal tensor $\mathcal{X} \in \mathbb{R}^{M \times N \times T}$ and denoting its corresponding vector predictors as $\mathbf{Z} \in \mathbb{R}^{QD \times T}$, such regression model would formulate the regression of $\mathcal{X}$ on $\mathbf{Z}$ as:

$$\mathcal{X}_{M \times N \times T} \approx \mathcal{S}_{r_1 \times r_2 \times r_3} \times_1 \mathbf{U}^{(1)}_{M \times r_1} \times_2 \mathbf{U}^{(2)}_{N \times r_2} \times_3 \mathbf{U}^{(3)}_{T \times r_3}(\mathbf{Z}) \tag{3}$$

where $\mathcal{S}$ is the core tensor, $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}$ are the factors for the two spatial modes and $\mathbf{U}^{(3)}$, determined by $\mathbf{Z}$, is the temporal factor. The $\times_j, j = 1, 2, 3$ operator is the tensor mode product [Kolda and Bader, 2009], defined in detail in Section 2.1.

Such a formulation makes the incorporation of vector predictors straightforward. In this paper, we simplify the problem by considering only the linear effect of the vector predictors and thus we have $\mathbf{U}^{(3)}(\mathbf{Z}) = \mathbf{Z}^\top$. By using $\mathcal{G}$ to denote all the coefficients in front of $\mathbf{U}^{(3)}(\mathbf{Z})$ in (3), we end up with an amended version of (3) as $\mathcal{X} \approx \mathcal{G}_{M \times N \times QD} \times_3 \mathbf{Z}^\top$, or equivalently for each $\mathbf{X}_t$:

$$\mathbf{X}_t \approx \mathcal{G}_{M \times N \times QD} \times_3 \left[\mathbf{z}_{t-1}^\top : \cdots : \mathbf{z}_{t-Q}^\top\right] \tag{4}$$

where we drop the third mode of the resulting tensor on the right hand side with dimension one. For any $(i, j)^{\text{th}}$ element of $\mathbf{X}_t$, it is predicted by a linear combination of lagged vector predictors whose coefficients are stored in $\mathcal{G}(i, j, :)$. Our final model combines (2) and (4) to merge the information of the two data modalities.

The estimation of $\mathcal{G}$ can become computationally prohibitive as the spatial resolution of the data grows. In (3), a low-rank structure is imposed on $\mathcal{G}$ as $\mathcal{G} = \mathcal{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)}$, which makes the computation more tractable. However, given the spatio-temporal setting of the data, we want to incorporate the spatial information more explicitly into $\mathcal{G}$ such

that the vector predictors will have spatially-varying coefficients with certain degree of spatial smoothness. This requires one to take a different approach from the sparse or low-ranked assumption [Rabusseau and Kadri, 2016, Sun and Li, 2017, Li and Zhang, 2017]. In Wang et al. [2017], the authors propose the total-variation (TV) regularization for spatially smooth tensor coefficients and Kang et al. [2018] proposes a Gaussian Process prior over the coefficients to incorporate the spatial information. In this paper, we propose a non-parametric method by treating the slices of $\mathcal{G}$, i.e. $\mathcal{G}(:,:,d)$, as discrete evaluations of functional parameters $g_d \in \mathcal{H}_K$ from a Reproducing Kernel Hilbert Space $\mathcal{H}_K$ (RKHS). This makes the joint estimation of $\mathcal{G}$ and the autoregressive coefficients in (2) much easier than the other approaches mentioned above. The spatial information is then incorporated into the choice of the spatial kernel K.

To enable efficient estimation of the functional parameters, we introduce a functional norm penalty over all g_d. Functional norm penalties have been widely used for estimating smooth functions in classic semi/non-parametric learning in which data variables are either scalar- or 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 the authors' knowledge, this present article is the first time using this technique when the observed time series are matrix-valued and the predictors are vector time series. Similar assumptions can be found in works on spatio-temporal process factor model [Chen et al., 2020] and non-parametric scalar-on-tensor regression model [Hao et al., 2021]. Different from these works, we consider the estimation problem jointly with a parametric matrix autoregressive model in a time series forecasting setting.

The remainder of the paper is organized as follows. Section 2 introduces the notations and outlines our methodology called MARAC, namely Matrix Auto-Regression with

Auxiliary Covariates. Section 3 and 4 discuss the estimating algorithm and the theoretical properties of the estimators. Simulation experiments on validating the algorithms and model selection criterion are in Section 5. A real data application to the TEC forecasting problem is detailed in Section 6. Section 7 concludes. Technical proofs and additional results are provided in the appendices.

2 Method

2.1 Notation

Inherited from the introductions, we adopt the following notation throughout the article. Tensor (with at least 3 modes) is denoted by calligraphic letters (e.g. $\mathcal{X}$). Matrix is denoted by boldface uppercase letters, such as $\mathbf{A}, \mathbf{E}_t$, and vector is denoted by boldface lowercase (Greek) letters such as $\boldsymbol{\beta}, \mathbf{z}_t$. For an arbitrary n-mode tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_n}$, including matrix and vector as special cases, we use $\mathcal{X}(i_1, \ldots, i_n)$ to denote the $(i_1, \ldots, i_n)^{\text{th}}$ element of $\mathcal{X}$ and thereby the tensor inner product is defined as: $\langle \mathcal{X}_1, \mathcal{X}_2 \rangle = \sum_{i_1, \ldots, i_n} \mathcal{X}_1(i_1, \ldots, i_n) \cdot \mathcal{X}_2(i_1, \ldots, i_n)$, and $\|\mathcal{X}\|_{\text{F}} \coloneqq \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$ is the tensor Frobenius norm. The m-mode tensor product for $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_n}$ with matrix $\mathbf{U} \in \mathbb{R}^{r_m \times I_m}$ is denoted as $\mathcal{X} \times_m \mathbf{U}$ and is of size $I_1 \times \ldots I_{m-1} \times r_m \times I_{m+1} \times \ldots I_n$. Element-wisely, we have:

$$\left(\mathcal{X} \times_m \mathbf{U}\right)(i_1, \ldots, i_{m-1}, j, i_{m+1}, \ldots, i_n) = \sum_{i_m=1}^{I_m} \mathcal{X}(i_1, \ldots, i_{m-1}, i_m, i_{m+1}, \ldots, i_n) \cdot \mathbf{U}(j, i_m).$$

We use $\mathbf{vec}(\cdot)$ to denote tensor vectorization operator where the first mode index change the fastest. The m-mode tensor unfolding for $\mathcal{X}$, denoted by $\mathbf{X}_{(m)}$, is an $I_m \times (\prod_{i \neq m} I_i)$ matrix whose i_m-row is defined as $\mathbf{X}_{(m)}(i_m, :) = \mathbf{vec}(\mathcal{X}(:, \ldots, i_m, \ldots, :))$. Finally, for any positive integer k, we use $[k]$ to denote its index set $\{1, \ldots, k\}$. We refer our readers to Kolda and Bader [2009] for a detailed introduction to the tensor operations.

The matrix time series $\{\mathbf{X}_t\}_{t=1}^T$ is collected on an $M \times N$ spatial grid. Without loss of generality, we restrict the spatial grid within $[0,1]^2$, and denote the collection of the spatial coordinates of all grid points as $\mathbb{S} \coloneqq \left\{\mathbf{s} \in [0,1]^2 \mid \mathbf{s} = \left(\frac{i}{M}, \frac{j}{N}\right), i \in [M], j \in [N]\right\}$. Any $(i,j)^{\text{th}}$ element on the matrix grid has its spatial coordinate $\mathbf{s}_{ij} \in \mathbb{S}$. Spatial kernel function is denoted as $K(.,.) : [0,1]^2 \times [0,1]^2 \mapsto \mathbb{R}$, and the corresponding RKHS is denoted as $\mathcal{H}_K$, endowed with a functional norm $\|\cdot\|_{\mathcal{H}_K}$.

2.2 Matrix Autoregression with Auxiliary Covariates (MARAC)

Let $\{\mathbf{X}_t\}_{t=1}^T$ be a matrix time series with $\mathbf{X}_t$ being an $M \times N$ matrix and $\{\mathbf{z}_t\}_{t=1}^T$ be a D-dimensional auxiliary vector time series, and both are observed at discrete time $t \in [T]$. We formulate our Matrix Autoregression with Auxiliary Covariates, or **MARAC**, as follows:

$$\mathbf{X}_t = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{X}_{t-p} \mathbf{B}_p^\top + \sum_{q=1}^{Q} \mathcal{G}_q \times_3 \mathbf{z}_{t-q}^\top + \mathbf{E}_t, \tag{5}$$

where $(\mathbf{A}_p, \mathbf{B}_p)$ are the autoregressive coefficients for lag-p matrix predictor $\mathbf{X}_{t-p}$, and $\mathcal{G}_q \in \mathbb{R}^{M \times N \times D}$ is the tensor regression coefficient of the lag-q auxiliary vector predictor $\mathbf{z}_{t-q}$. $\{\mathbf{E}_t\}_{t=1}^T$ is an i.i.d. noise process and is independent of the auxiliary vector time series $\{\mathbf{z}_t\}_{t=1}^T$. The lag parameters P and Q are hyperparameters of the model, and we denote the model in (5) briefly as MARAC(P,Q). Note that we drop the tensor dimension with dimensionality one on the right hand side of (5).

As stated in the introduction, the key assumption on $\mathcal{G}_q$ is that any slice of $\mathcal{G}_q$ along the third mode, namely $\mathcal{G}_q(:,:,d), d \in [D]$, or denoted as $\mathbf{G}_{q,d} \in \mathbb{R}^{M \times N}$, is the evaluation of a functional parameter $g_{q,d} : [0,1]^2 \mapsto \mathbb{R}$ on the spatial grid $\mathbb{S}$. Furthermore, $g_{q,d}$ comes from a Reproducing Kernel Hilbert Space $\mathcal{H}_K$ endowed with a spatial kernel function $K(.,.)$.

Element-wisely, the MARAC(P,Q) specifies the following relationship between $\mathbf{X}_t(i,j)$

and the predictors:

$$\mathbf{X}_t(i,j) = \sum_{p=1}^{P} \left\langle \mathbf{A}_p(i,:)^\top \mathbf{B}_p(j,:), \mathbf{X}_{t-p} \right\rangle + \sum_{q=1}^{Q} \mathbf{g}_q(\mathbf{s}_{ij})^\top \mathbf{z}_{t-q} + \mathbf{E}_t(i,j), \tag{6}$$

where $\mathbf{g}_q(\cdot) = [g_{q,1}(\cdot), g_{q,2}(\cdot), \ldots, g_{q,D}(\cdot)]^\top$. The lag-$q$ vector covariates $\mathbf{z}_{t-q}$, though being shared across all matrix entries, has element-specific effect on each $\mathbf{X}_t(i,j)$ characterized by the entry-specific coefficients $\mathbf{g}_q(\mathbf{s}_{ij}) \in \mathbb{R}^D$. The lag-$p$ matrix predictor $\mathbf{X}_{t-p}$ has a rank-1 coefficient matrix determined by the specific rows of $\mathbf{A}_p$ and $\mathbf{B}_p$.

Let $\mathbf{\Sigma}$ denote the covariance matrix of $\mathbf{vec}\,(\mathbf{E}_t)$. The error covariance $\mathbf{\Sigma}$ is assumed to have a separable Kronecker-product covariance structure following Chen et al. [2021]:

$$\mathbf{vec}\,(\mathbf{E}_t) \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r\right), \tag{7}$$

where $\mathbf{\Sigma}_c \in \mathbb{R}^{N\times N}, \mathbf{\Sigma}_r \in \mathbb{R}^{M\times M}$ are the column and row covariance components and are symmetric and positive definite. Such Kronecker-product covariance matrix is commonly seen in the covariance models for multi-way tensor data [Hoff, 2011, Fosdick and Hoff, 2014]. Alternatively, one can assume a Kronecker-sum covariance matrix [Greenewald et al., 2019, Wang et al., 2020] which could introduce a sparser covariance matrix. We stick to the Kronecker-product form here for simplicity. Note that the error process is serially-independent and is also assumed to be independent of $\{\mathbf{z}_t\}_{t=1}^T$.

Combining (5) and (7) yields the full MARAC(P,Q) model and the collection of model parameters $\mathbf{\Theta} \coloneqq \left\{\{\mathbf{A}_p, \mathbf{B}_p\}_{p\in[P]}, \{g_{q,d}\}_{q\in[Q],d\in[D]}, \{\mathbf{\Sigma}_r, \mathbf{\Sigma}_c\}\right\}$ contain the autoregression parameters, the auxiliary covariate parameters and the error covariance parameters. We will introduce the estimating algorithm for $\mathbf{\Theta}$ in the next section. Before that, we want to make a comparison of our approach against the existing matrix autoregression models (MAR) in Chen et al. [2021] and Hsu et al. [2021].

At first glance, our model improves the modeling capability of the MAR models by incorporating the auxiliary covariates. More deeply, if one combines the term that involves

the auxiliary covariates and the error term in (5) as a new error term, say $\mathbf{E}_t^{'}$, then one ends up with just another MAR model like those in Chen et al. [2021], Hsu et al. [2021]:

$$\mathbf{X}_t = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{X}_{t-p} \mathbf{B}_p^\top + \mathbf{E}_t^{'},$$

And the covariance of $\mathbf{vec}(\mathbf{E}_t^{'})$ can be derived based on the independence of $\mathbf{E}_t$ and $\mathbf{z}_t$:

$$\mathrm{Var}\left(\mathbf{vec}(\mathbf{E}_t^{'})\right) = \mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r + \mathbf{F}\mathbf{M}\mathbf{F}^\top \tag{8}$$

where $\mathbf{F} \in \mathbb{R}^{MN \times QD}$ is defined as $\mathbf{F} = [(\mathbf{G}_1)_{(3)}^\top : \ldots : (\mathbf{G}_Q)_{(3)}^\top]$ and $(\mathbf{G}_q)_{(3)}$ is the third-mode unfolding of the tensor $\mathcal{G}_q$. $\mathbf{M} \in \mathbb{R}^{QD \times QD}$ is a Q-by-Q block matrix with its $(q_1, q_2)^{\text{th}}$ block being $\mathrm{Cov}(\mathbf{z}_{t-q_1}, \mathbf{z}_{t-q_2})$. Essentially, each column of $\mathbf{F}$ contains the coefficient of one vector covariate at one specific lag. As compared to Chen et al. [2021], we have an additive term that accounts for the spatial correlation of the error process. As compared to Hsu et al. [2021], whose covariance matrix shares the same form as that in (8), we do not take a fixed-rank co-kriging approach [Cressie and Johannesson, 2008] by fixing $\mathbf{F}$ as some pre-specified spatial bases and let $\mathbf{M}$ be the model parameter. On the contrary, we let $\mathbf{M}$ to be determined by the auxiliary covariate and estimate columns of $\mathbf{F}$ from an RKHS.

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

$$\mathbf{vec}\left(\mathbf{X}_t\right) = \sum_{p=1}^{P} \left(\mathbf{B}_p \otimes \mathbf{A}_p\right) \mathbf{vec}\left(\mathbf{X}_{t-p}\right) + \sum_{q=1}^{Q} \left(\mathbf{G}_q\right)_{(3)}^\top \mathbf{z}_{t-q} + \mathbf{vec}\left(\mathbf{E}_t\right), \tag{9}$$

For clarity of notation, we will use $\mathbf{G}_q \in \mathbb{R}^{MN \times D}$ to denote $(\mathbf{G}_q)_{(3)}^\top$ for the remainder of the paper. The Kronecker-product form of the autoregressive coefficients and the error covariance reduces the model complexity significantly and a functional norm penalty over columns of $\mathbf{G}_q$, or essentially over all $g_{q,d} \in \mathcal{H}_K, d \in [D]$, that we introduce later, reduces the semi-parametric estimation problem to a finite dimensional problem. And all of these assumptions result in an efficient estimation algorithm that we will show in the next section.

3 Model Estimation

In this section, we discuss the parameter estimation of the MARAC model. We first propose a penalized maximum likelihood estimator (PMLE) for the semi-parametric model in (5) in Section 3.1 and then propose an approximation to the penalized MLE in Section 3.2 for faster computation with high-dimensional matrix data.

3.1 Penalized Maximum Likelihood Estimator (PMLE)

In order to estimate the MARAC(P, Q) model parameters $\boldsymbol{\Theta}$, we propose a penalized maximum likelihood estimation (PMLE) approach. Based on the Gaussianity assumption on $\mathbf{vec}(\mathbf{E}_t)$ in (7), we define the loss function, denoted by $\mathrm{L}(\boldsymbol{\Theta})$, as the negative log-likelihood with a penalization term and solve the following minimization problem:

$$\min_{\boldsymbol{\Theta}} \left\{ \mathrm{L}(\boldsymbol{\Theta}) \coloneqq \frac{1}{2} \log |\boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r| + \frac{1}{2T} \sum_{t=1}^{T} \mathbf{r}_t^\top \left(\boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r\right)^{-1} \mathbf{r}_t + \frac{\lambda}{2} \sum_{q=1}^{Q} \sum_{d=1}^{D} \left\| g_{q,d} \right\|_{\mathcal{H}_K}^2 \right\}, \quad (10)$$

where $\mathbf{r}_t = \mathbf{vec}(\mathbf{R}_t)$ is the vectorized residual at time t. Using (9), we can write $\mathbf{r}_t$ as:

$$\mathbf{r}_t = \mathbf{vec}\left(\mathbf{X}_t\right) - \sum_{p=1}^{P} \left(\mathbf{B}_p \otimes \mathbf{A}_p\right) \mathbf{vec}\left(\mathbf{X}_{t-p}\right) - \sum_{q=1}^{Q} \mathbf{G}_q \mathbf{z}_{t-q}. \quad (11)$$

Note that we define $\mathrm{L}(\boldsymbol{\Theta})$ based on residuals at $t \in [T]$, and one can simply extend the time series of $\mathbf{X}_t$ and $\mathbf{z}_t$ to allow for negative time indices such that $\mathbf{r}_t$ in (11) is well defined for all $t \in [T]$. It is evident from (10) that the loss function $\mathrm{L}(\boldsymbol{\Theta})$ consists of a normalized negative log-likelihood for the residual process and a functional norm penalty over the functional parameters for the auxiliary covariates.

Now we define the functional norm, i.e. $\left\| g_{q,d} \right\|_{\mathcal{H}_K}$, explicitly. We assume that the spatial kernel $K(\cdot, \cdot)$ is continuous and square integrable, thus it has an eigen-decomposition following the Mercer's Theorem [Williams and Rasmussen, 2006]:

$$K(\mathbf{s}_{ij}, \mathbf{s}_{uv}) = \sum_{k=1}^{\infty} \lambda_k \psi_k(\mathbf{s}_{ij}) \psi_k(\mathbf{s}_{uv}), \quad \mathbf{s}_{ij}, \mathbf{s}_{uv} \in [0,1]^2, \quad (12)$$

where $\{\lambda_k\}_{k=1}^{\infty}$ is a non-increasing sequence of non-negative eigenvalues and $\{\psi_k(\cdot)\}_{k=1}^{\infty}$ is a set of orthonormal basis functions defined on $[0,1]^2$. The functional norm from the RKHS endowed with kernel $K(\cdot,\cdot)$ is defined based on van Zanten and van der Vaart [2008]:

$$\|g_{q,d}\|_{\mathcal{H}_K} = \sqrt{\sum_{k=1}^{\infty} \frac{\beta_{q,d,k}^2}{\lambda_k}}, \quad \text{where } g_{q,d}(\mathbf{s}_{ij}) = \sum_{k=1}^{\infty} \beta_{q,d,k}\psi_k(\mathbf{s}_{ij}). \tag{13}$$

For any $\lambda > 0$ in (10), the generalized representer theorem [Schölkopf et al., 2001] suggests that the solution for the functional parameters, denoted as $\widetilde{g}_{q,d}$, of the minimization problem (10), given any fixed $\{(\mathbf{A}_p, \mathbf{B}_p)\}_{p\in[P]}$ and $\mathbf{\Sigma}_r, \mathbf{\Sigma}_c$, is a linear combination of the representers $K(., \mathbf{s}_{uv})$ (the null space contains only the zero function), namely:

$$\widetilde{g}_{q,d}(\mathbf{s}_{ij}) = \sum_{u=1}^{M}\sum_{v=1}^{N} \widetilde{\gamma}_{q,d}^{(u,v)} K(\mathbf{s}_{ij}, \mathbf{s}_{uv}), \tag{14}$$

where $\widetilde{\gamma}_{q,d}^{(u,v)}$ is the optimal linear combination coefficient. The minimization problem in (10) can thus be reduced to a finite-dimensional Kernel Ridge Regression (KRR) problem:

$$\min_{\substack{\{\mathbf{A}_p,\mathbf{B}_p\}_{p\in[P]} \\ \{\boldsymbol{\gamma}_q\}_{q\in[Q]},\mathbf{\Sigma}_c,\mathbf{\Sigma}_r}} \left\{ \frac{1}{2}\log|\mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r| + \frac{1}{2T}\sum_{t=1}^{T} \mathbf{r}_t^\top \left(\mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r\right)^{-1} \mathbf{r}_t + \frac{\lambda}{2}\sum_{q=1}^{Q} \mathbf{tr}\left(\boldsymbol{\gamma}_q^\top \mathbf{K}\boldsymbol{\gamma}_q\right) \right\}, \tag{15}$$

where $\mathbf{tr}(\cdot)$ takes the trace of a square matrix, $\boldsymbol{\gamma}_q \in \mathbb{R}^{MN\times D}$ contains the coefficients of the representers for each of the D covariates in $\mathbf{z}_{t-q}$ across all MN locations. Given the definitions, $\mathbf{r}_t$ in (11) can now be expressed as:

$$\mathbf{r}_t = \mathbf{vec}\left(\mathbf{X}_t\right) - \sum_{p=1}^{P}\left(\mathbf{B}_p \otimes \mathbf{A}_p\right)\mathbf{vec}\left(\mathbf{X}_{t-p}\right) - \sum_{q=1}^{Q}\mathbf{K}\boldsymbol{\gamma}_q\mathbf{z}_{t-q},$$

and $\mathbf{K} \in \mathbb{R}^{MN\times MN}$ is the kernel gram matrix where $K(\mathbf{s}_{ij}, \mathbf{s}_{uv}) = \mathbf{K}((i-1)M + j, (u-1)M + v), \forall i, u \in [M], j, v \in [N]$.

We attempt to solve the minimization problem in (15) with a cyclical minimization algorithm 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: $\mathbf{A}_1 \to \mathbf{B}_1 \to \ldots \to \mathbf{A}_P \to \mathbf{B}_P \to \boldsymbol{\gamma}_1 \to \ldots \to \boldsymbol{\gamma}_Q \to \mathbf{\Sigma}_c \to \mathbf{\Sigma}_r \to \mathbf{A}_1 \to \mathbf{B}_1 \to \ldots$.

To update $\mathbf{A}_p$, we use:

$$\mathbf{A}_p \leftarrow \left[\sum_t \widetilde{\mathbf{X}}_{t,-p}\boldsymbol{\Sigma}_c^{-1}\mathbf{B}_p\mathbf{X}_{t-p}^\top\right]\left[\sum_t \mathbf{X}_{t-p}\mathbf{B}_p^\top\boldsymbol{\Sigma}_c^{-1}\mathbf{B}_p\mathbf{X}_{t-p}^\top\right]^{-1}, \tag{16}$$

where $\widetilde{\mathbf{X}}_{t,-p}$ is the partial matrix-shaped residual defined as:

$$\widetilde{\mathbf{X}}_{t,-p} = \mathbf{X}_t - \sum_{p'\neq p}\mathbf{A}_{p'}\mathbf{X}_{t-p'}\mathbf{B}_{p'}^\top - \sum_{q=1}^{Q}\mathcal{G}_q \times_3 \mathbf{z}_{t-q}^\top. \tag{17}$$

where $\mathcal{G}_q$ now is the "tensorized" version of $\mathbf{K}\boldsymbol{\gamma}_q$ with $\mathbf{vec}(\mathcal{G}_q(:,:,d)) = \mathbf{K}\boldsymbol{\gamma}_q(:,d), \forall d \in [D]$. Note that all the model parameters in (16) and (17) should be their current values in the iterative algorithm, but we omit the notations for the iteration counter for the simplicity of the presentation without essential loss of clarity.

Similarly, one can update $\mathbf{B}_p$ as follows:

$$\mathbf{B}_p \leftarrow \left[\sum_t \widetilde{\mathbf{X}}_{t,-p}^\top\boldsymbol{\Sigma}_r^{-1}\mathbf{A}_p\mathbf{X}_{t-p}\right]\left[\sum_t \mathbf{X}_{t-p}^\top\mathbf{A}_p^\top\boldsymbol{\Sigma}_r^{-1}\mathbf{A}_p\mathbf{X}_{t-p}\right]^{-1}. \tag{18}$$

To estimate $\boldsymbol{\gamma}_q$, we can rewrite the minimization problem in (15) in terms of $\mathbf{vec}(\boldsymbol{\gamma}_q) \in \mathbb{R}^{MND}$ by only keeping the terms involving $\boldsymbol{\gamma}_q$ as:

$$\min_{\mathbf{vec}(\boldsymbol{\gamma}_q)}\left\{\frac{1}{2T}\sum_{t=1}^{T}\mathbf{r}_t(\boldsymbol{\gamma}_q)^\top\left(\boldsymbol{\Sigma}_c\otimes\boldsymbol{\Sigma}_r\right)^{-1}\mathbf{r}_t(\boldsymbol{\gamma}_q) + \frac{\lambda}{2}\mathbf{vec}\left(\boldsymbol{\gamma}_q\right)^\top\left(\mathbf{I}_D\otimes\mathbf{K}\right)\mathbf{vec}\left(\boldsymbol{\gamma}_q\right)\right\}, \tag{19}$$

where $\mathbf{r}_t(\boldsymbol{\gamma}_q) = \mathbf{vec}(\widetilde{\mathbf{X}}_{t,-q}) - (\mathbf{z}_{t-q}^\top\otimes\mathbf{K})\mathbf{vec}(\boldsymbol{\gamma}_q)$ with the vectorized partial residual $\mathbf{vec}(\widetilde{\mathbf{X}}_{t,-q})$ defined as:

$$\mathbf{vec}(\widetilde{\mathbf{X}}_{t,-q}) = \mathbf{vec}\left(\mathbf{X}_t\right) - \sum_{p=1}^{P}\left(\mathbf{B}_p\otimes\mathbf{A}_p\right)\mathbf{vec}\left(\mathbf{X}_{t-p}\right) - \sum_{q'\neq q}\mathbf{K}\boldsymbol{\gamma}_{q'}\mathbf{z}_{t-q'}, \tag{20}$$

and $\mathbf{I}_D$ is a $D\times D$ identity matrix. The updating rule for $\mathbf{vec}(\boldsymbol{\gamma}_q)$ follows from (19):

$$\mathbf{vec}\left(\boldsymbol{\gamma}_q\right) \leftarrow \left[\left(\frac{\sum_t \mathbf{z}_{t-q}\mathbf{z}_{t-q}^\top}{T}\right)\otimes\boldsymbol{\Sigma}^{-1}\mathbf{K} + \lambda\mathbf{I}_{DMN}\right]^{-1}\left[\frac{1}{T}\sum_t\left(\mathbf{z}_{t-q}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbf{vec}\left(\widetilde{\mathbf{X}}_{t,-q}\right)\right], \tag{21}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_c\otimes\boldsymbol{\Sigma}_r$. Note that the computation of the matrix inverse in (21) only requires calculating the autocovariance of $\mathbf{z}_{t-q}$ across time. The computational complexity of the

matrix inversion is $\mathcal{O}\left((MND)^3\right)$, in the current form, since all MN representer functions are being used for the estimation. In section 3.2, we introduce a finite truncation of the spatial kernel function to reduce computational complexity to $\mathcal{O}((RD)^3)$, with $R << MN$.

The updating rules for $\mathbf{\Sigma}_c$ and $\mathbf{\Sigma}_r$ can be easily derived by calculating their gradients of the loss function (15) and setting them to zero, and we state the results here directly:

$$\mathbf{\Sigma}_c \leftarrow \frac{\sum_t \mathbf{R}_t^\top \mathbf{\Sigma}_r^{-1} \mathbf{R}_t}{MT}, \quad \mathbf{\Sigma}_r \leftarrow \frac{\sum_t \mathbf{R}_t \mathbf{\Sigma}_c^{-1} \mathbf{R}_t^\top}{NT} \tag{22}$$

where $\mathbf{R}_t$ is the full residual matrix at time t:

$$\mathbf{R}_t = \mathbf{X}_t - \sum_{p=1}^{P} \mathbf{A}_p \mathbf{X}_{t-p} \mathbf{B}_p^\top - \sum_{q=1}^{Q} \mathcal{G}_q \times_3 \mathbf{z}_{t-q}^\top, \quad \text{where } \mathbf{vec}(\mathcal{G}_q(:,:,d)) = \mathbf{K}\boldsymbol{\gamma}_q(:,d), \forall d \in [D].$$

The entire algorithm iterates the updating formula in (16), (18), (21), (22) until convergence. For each pair of $\mathbf{A}_p$ and $\mathbf{B}_p$, they are only identifiable up to a change of scale and sign since $(\mathbf{B}_p \otimes \mathbf{A}_p) = [(c^{-1} \cdot \mathbf{B}_p) \otimes (c \cdot \mathbf{A}_p)], \forall c \neq 0$. Therefore, we re-normalize the sign and scale of each pair of $\mathbf{A}_p, \mathbf{B}_p$ after all estimators converge such that $\|\mathbf{A}_p\|_{\mathrm{F}} = 1$ and $\mathrm{sign}(\mathbf{tr}\left(\mathbf{A}_p\right)) = 1$. We summarize the entire algorithm for solving (15) in Algorithm 1 in Section A of the appendix.

After the i^{th} round of the cyclic updates, we compute the following statistics for checking the algorithm convergence:

$$\frac{\sum_{p=1}^{P} \left\| \Delta(\mathbf{B}_p^{(i)} \otimes \mathbf{A}_p^{(i)}) \right\|_{\mathrm{F}}^2}{PM^2N^2}, \quad \frac{\sum_{q=1}^{Q} \|\Delta \boldsymbol{\gamma}_q^{(i)}\|_{\mathrm{F}}^2}{QMND}, \quad \frac{\left\| \Delta(\mathbf{\Sigma}_c^{(i)} \otimes \mathbf{\Sigma}_r^{(i)}) \right\|_{\mathrm{F}}^2}{M^2N^2}, \tag{23}$$

where Δ is the inter-iteration difference operator that computes the relative change of any parameter from iteration $i-1$ to i. Each statistics in (23) simply computes the relative change of the parameters via the mean squared error. The algorithm terminates when all three statistics in (23) are lower than a pre-specified threshold.

When dealing with high-dimensional matrix data, it is not computationally efficient to compute the Kronecker products, and thus we use the following upper bound as a proxy

for convergence check instead:

$$\left\|\Delta(\mathbf{B}_p^{(i)} \otimes \mathbf{A}_p^{(i)})\right\|_{\mathrm{F}} \leq \|\mathbf{B}_p^{(i-1)}\|_{\mathrm{F}} \cdot \|\Delta\mathbf{A}_p^{(i)}\|_{\mathrm{F}} + \|\mathbf{A}_p^{(i)}\|_{\mathrm{F}} \cdot \|\Delta\mathbf{B}_p^{(i)}\|_{\mathrm{F}}.$$

A similar bound applies to the convergence statistics for the error covariance matrix.

The cyclic minimization algorithm guarantees that the loss function (15) does not increase after each update. The optimization problem is not convex for all parameters but is convex for each block of parameter while keeping the other blocks fixed. Such an algorithm typically converges to a stationary point with a convergence rate of $\mathcal{O}(1/k)$ where k is the total number of iterations, see Saha and Tewari [2013] for detailed discussions.

3.2 PMLE with Kernel Truncation

The penalized MLE proposed in the previous section requires matrix inversion in multiple parameter updating steps, and the updating formula for $\{\boldsymbol{\gamma}_q\}_{q=1}^{Q}$ in (21) requires inverting a matrix of size $MND \times MND$, which is computationally demanding for large matrices such as the TEC time series with $M = 181, N = 361$. The high dimensionality in (21) is due to the presence of MN representers: $\{K(., \mathbf{s}_{uv})\}_{u\in[M],v\in[N]}$. In practice, we can approximate the linear combination of all MN representers using a sparser set of basis functions. For example, one can reduce the spatial resolution and pick a subset of the representers as basis functions. In this section, we consider truncating the eigen-decomposition of the spatial kernel $K(.,.)$ instead. A similar technique can be found in Kang et al. [2018].

Given the Mercer decomposition of the spatial kernel $K(.,.)$ in (12), one can truncate the eigen-decomposition at the R^{th} largest eigenvalue: $K(\mathbf{s}_{ij}, \mathbf{s}_{iv}) \approx \sum_{k=1}^{R} \lambda_k \psi_k(\mathbf{s}_{ij})\psi_k(\mathbf{s}_{uv})$, and retains a set of basis functions $\{\psi_1(\cdot), \psi_2(\cdot), \ldots, \psi_R(\cdot)\}$. The choice of R depends on the decaying rate of the eigenvalue sequence $\{\lambda_k\}_{k=1}^{\infty}$, and thus the choice of the kernel. Our simulation result shows that the estimation and prediction errors of the model shrink

monotonically as $R \to \infty$, so R can be chosen based on computational resources available.

Intuitively from (13), when λ_k is very small, any non-zero coefficient for the eigenfunction $\psi_k(\cdot)$ will incur a high functional norm penalty. Typically, these eigenfunctions feature more local "wiggliness" or unsmoothness and thus the truncation of the kernel yields a smoother fit, as we will demonstrate empirically in section 5.2.

Given this basis truncation, any functional parameter $g_{q,d}$ is now a linear combination of R basis functions, and the MARAC(P, Q) model becomes:

$$\mathbf{vec}(\mathbf{X}_t) = \sum_{p=1}^{P} (\mathbf{B}_p \otimes \mathbf{A}_p)\, \mathbf{vec}(\mathbf{X}_{t-p}) + \sum_{q=1}^{Q} \mathbf{K}_R \boldsymbol{\gamma}_q \mathbf{z}_{t-q} + \mathbf{vec}(\mathbf{E}_t), \tag{24}$$

where each column of $\mathbf{K}_R \in \mathbb{R}^{MN \times R}$ is the one of the R basis functions evaluated on the spatial grid and $\boldsymbol{\gamma}_q \in \mathbb{R}^{R \times D}$ becomes the basis coefficients. We name the model in (24) as the MARAC(P, Q, R) model, or the *truncated model*.

The estimation of the truncated model is largely the same as the MARAC(P, Q) model and is aimed at solving the following optimization problem:

$$\min_{\substack{\{\mathbf{A}_p, \mathbf{B}_p\}_{p\in[P]} \\ \{\boldsymbol{\gamma}_q\}_{q\in[Q]}, \boldsymbol{\Sigma}_c, \boldsymbol{\Sigma}_r}} \left\{ \frac{1}{2} \log |\boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r| + \frac{1}{2T} \sum_{t=1}^{T} \mathbf{r}_t^\top (\boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r)^{-1} \mathbf{r}_t + \frac{\lambda}{2} \sum_{q=1}^{Q} \mathbf{tr}\left(\boldsymbol{\gamma}_q^\top \boldsymbol{\Lambda}_R^{-1} \boldsymbol{\gamma}_q\right) \right\} \tag{25}$$

where $\boldsymbol{\Lambda}_R = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_R)$ is a diagonal matrix containing the eigenvalues of the leading R eigenfunctions and $\mathbf{r}_t$ is defined in a similar way as (11). The updating formulae for $\{\mathbf{A}_p, \mathbf{B}_p\}_{p\in[P]}, \boldsymbol{\Sigma}_c, \boldsymbol{\Sigma}_r$ still follow those in (16), (18), (22), and the only adjustment is to replace $\mathbf{K}$ with $\mathbf{K}_R$. The key difference is the updating formula for $\boldsymbol{\gamma}_q$, which is different from that in (21), but is still a kernel ridge regression estimator:

$$\mathbf{vec}\left(\boldsymbol{\gamma}_q\right) \leftarrow \left[\widehat{\boldsymbol{\Upsilon}}_q \otimes \left(\mathbf{K}_R^\top \boldsymbol{\Sigma}^{-1} \mathbf{K}_R\right) + \lambda \left(\mathbf{I}_D \otimes \boldsymbol{\Lambda}_R^{-1}\right)\right]^{-1} \left[\frac{1}{T} \sum_t \mathbf{z}_{t-q} \otimes \left(\mathbf{K}_R^\top \boldsymbol{\Sigma}^{-1} \mathbf{vec}\left(\widetilde{\mathbf{X}}_{t,-q}\right)\right)\right], \tag{26}$$

where $\widehat{\boldsymbol{\Upsilon}}_q = \sum_t \mathbf{z}_{t-q} \mathbf{z}_{t-q}^\top / T$. We omit the full algorithm for the truncated model here but the overall updating schemes follow (16), (18), (22), (26).

4 Theoretical Analysis

In this section, we present two major theoretical results of the MARAC model. We first derive the stationarity condition of the MARAC(P, Q) model, and then establish the consistency and the asymptotic normality of the penalized maximum likelihood estimator output by Algorithm 1 given a fixed dimensionality of the matrix.

4.1 Stationarity Condition

To derive the stationarity condition for the matrix time series $\{\mathbf{X}_t\}_{t=1}^{\infty}$ generated by the MARAC(P, Q) model, we first make an assumption on the auxiliary time series $\{\mathbf{z}_t\}_{t=1}^{\infty}$.

Assumption 4.1.1 (Exogenous Variable Condition)**.** *The auxiliary time series* $\{\mathbf{z}_t\}_{z=1}^{\infty}$ *follows a VAR*($\widetilde{Q}$) *process:*

$$\mathbf{z}_t = \mathbf{C}_1\mathbf{z}_{t-1} + \mathbf{C}_2\mathbf{z}_{t-2} + \ldots + \mathbf{C}_{\widetilde{Q}}\mathbf{z}_{t-\widetilde{Q}} + \boldsymbol{\nu}_t, \tag{27}$$

where $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_{\widetilde{Q}}$ *are the autoregressive coefficient matrices and* $\{\boldsymbol{\nu}_t\}_{t=1}^{\infty}$ *is a serially-independent noise process with bounded fourth-order moments and is independent of* $\{\mathbf{E}_t\}_{t=1}^{\infty}$*, the noise process of* $\{\mathbf{X}_t\}_{t=1}^{\infty}$*. In other words, conditional on* $\mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots, \mathbf{z}_{t-\widetilde{Q}}$*,* $\mathbf{z}_t$ *is independent of the matrix time series up to time* t*, i.e.* $\{\mathbf{X}_s\}_{s\leq t}$*.*

Assumption 4.1.1 states that the auxiliary time series $\mathbf{z}_t$ is generated from a lag-$\widetilde{Q}$ vector autoregressive process and affects the matrix time series $\mathbf{X}_t$ as an *exogenous* variable. To put it in the context of TEC forecasting, the IMF Bz index, which measures the strength of the solar energy, is affecting the Earth's TEC distribution as an external driving force and is not directly impacted by the Earth's TEC. With this assumption, we derive the *joint* stationarity condition for the matrix time series $\{\mathbf{X}_t\}_{t=1}^{\infty}$ and $\{\mathbf{z}_t\}_{t=1}^{\infty}$ in Theorem 4.1.2.

Theorem 4.1.2 (MARAC Stationarity Condition)**.** *With the assumption 4.1.1 that the D-dimensional* $\{\mathbf{z}_t\}_{t=1}^{\infty}$ *follows a VAR*$(\widetilde{Q})$ *process, suppose that the* $M \times N$ *matrix time series* $\{\mathbf{X}_t\}_{t=1}^{\infty}$ *is generated from an MARAC*(P, Q) *process:*

$$\mathbf{vec}\left(\mathbf{X}_t\right) = \sum_{p=1}^{P} \left(\mathbf{B}_p \otimes \mathbf{A}_p\right) \mathbf{vec}\left(\mathbf{X}_{t-p}\right) + \sum_{q=1}^{Q} \mathbf{G}_q \mathbf{z}_{t-q} + \mathbf{vec}\left(\mathbf{E}_t\right).$$

The matrix time series and the auxiliary vector time series are jointly stationary if and only if both $\det(\mathbf{\Phi}_1(y)) \neq 0$ *and* $\det(\mathbf{\Phi}_2(y)) \neq 0$ *for any* $y \in \mathbb{C}, |y| \leq 1$*, where the characteristic polynomials* $\mathbf{\Phi}_1(y)$ *and* $\mathbf{\Phi}_2(y)$ *are defined as:*

$$\mathbf{\Phi}_1(y) \coloneqq \mathbf{I}_{MN} - \sum_{p=1}^{P} \left(\mathbf{B}_p \otimes \mathbf{A}_p\right) y^p, \quad \mathbf{\Phi}_2(y) \coloneqq \mathbf{I}_D - \sum_{q=1}^{\widetilde{Q}} \mathbf{C}_q y^q.$$

The proof is provided in Section B of the appendix. Theorem 4.1.2 states that the stationarity of the MARAC model requires the stationarity of the autoregressive part of the MARAC(P, Q) model and the stationarity of the VAR$(\widetilde{Q})$ model, and is not impacted by the coefficients $\{\mathbf{G}_q\}_{q \in [Q]}$. As a special case when $P = Q = \widetilde{Q} = 1$, the stationarity condition requires that $\rho(\mathbf{A}_1) \cdot \rho(\mathbf{B}_1) < 1$ and $\rho(\mathbf{C}_1) < 1$, where $\rho(.)$ is the spectral radius of a matrix. The condition on $\mathbf{A}_1, \mathbf{B}_1$ echoes the result in Proposition 1 of Chen et al. [2021].

4.2 Consistency and Asymptotic Normality

In this subsection, we establish the consistency and asymptotic normality of the model estimators for MARAC(P, Q) model, given fixed matrix dimensionality. We start with the consistency of the error covariance estimator $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}_c \otimes \widehat{\mathbf{\Sigma}}_r$ in proposition 4.2.1.

Proposition 4.2.1. *Assume that* $\lambda \to 0$ *as* $T \to \infty$ *at any rate, then* $\widehat{\mathbf{\Sigma}}$*, the estimation of the error covariance, converges in probability to the ground truth* $\mathbf{\Sigma}$*.*

We leave the proof to Section C of the appendix. We further establish the asymptotic

normality for the penalized MLE estimators: $\{\widehat{\mathbf{A}}_p, \widehat{\mathbf{B}}_p\}_{p\in[P]}$, $\{\widehat{\gamma}_q\}_{q\in[Q]}$ in Theorem 4.2.2 based on Proposition 4.2.1.

Theorem 4.2.2. *With the assumption that $\lambda = o(1/\sqrt{T})$ and that the matrix time series $\{\mathbf{X}_t\}_{t=1}^T$ is generated by MARAC(P,Q) model in* (5) *and is jointly stationary with the vector time series $\{\mathbf{z}_t\}_{t=1}^T$, and the error process $\{\mathbf{E}_t\}_{t=1}^T$ is i.i.d. Gaussian following* (7)*. Then given fixed dimensionality of $\mathbf{X}_t, \mathbf{z}_t$ and known P and Q, the penalized maximum likelihood estimator obtained by solving* (10) *follows an asymptotic normal distribution:*

$$\sqrt{T}\begin{bmatrix}\mathbf{vec}(\widehat{\mathcal{A}}-\mathcal{A})\\ \mathbf{vec}(\widehat{\mathcal{B}}-\mathcal{B})\\ \mathbf{vec}(\widehat{\mathcal{R}}-\mathcal{R})\end{bmatrix} \xrightarrow{d.} \mathcal{N}\left(\mathbf{0}, \Xi_1\right), \tag{28}$$

where $\mathcal{A}\in\mathbb{R}^{M\times M\times P}, \mathcal{B}\in\mathbb{R}^{N\times N\times P}, \mathcal{R}\in\mathbb{R}^{MN\times D\times Q}$ are tensors with $\mathcal{A}[:,:,p] = \mathbf{A}_p, \mathcal{B}[:,:,p] = \mathbf{B}_p^\top, \mathcal{R}[:,:,q] = \boldsymbol{\gamma}_q$, and $\widehat{\mathcal{A}}, \widehat{\mathcal{B}}, \widehat{\mathcal{R}}$ are the corresponding estimators. Furthermore, $\Xi_1 = \mathbf{H}^{-1}\mathbb{E}[\mathbf{W}_t^\top\boldsymbol{\Sigma}^{-1}\mathbf{W}_t]\mathbf{H}^{-1}$, where $\mathbf{W}_t$ is defined as:

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

and $\mathbf{H} = \mathbb{E}[\mathbf{W}_t^\top\boldsymbol{\Sigma}^{-1}\mathbf{W}_t] + \boldsymbol{\eta}\boldsymbol{\eta}^\top$ with $\boldsymbol{\eta} = [\mathbf{vec}(\mathbf{A}_1)^\top,\ldots,\mathbf{vec}(\mathbf{A}_P)^\top,\mathbf{0}^\top]^\top$.

We leave the proof to Section D of the appendix. In summary, all autoregression coefficients and auxiliary covariate coefficients share the same convergence rate of $\sqrt{T}$, which is similar to the result obtained for the matrix autoregression model in Chen et al. [2021]. The regime discussed here is the fixed dimensionality case, where (M,N) are fixed. The case when $M,N\to\infty$ as $T\to\infty$ for partial linear model has been discussed when the samples are independent instead of being serially-dependent, see for example Cui et al. [2018]. We leave the theoretical discussion for the high-dimensional case for future work but will demonstrate the empirical performance of the model estimators under growing dimensionality in the next section.

5 Simulation Study

In this section, we conduct a series of systematically designed simulation experiments to first validate the estimators obtained by the penalized maximum likelihood estimation introduced in Section 3, under finite sample scenarios; and then we validate the efficacy of two information criteria for choosing the lag orders of the MARAC model. We start the section by providing a detailed explanation of the simulation setup.

5.1 Simulation Study Design

We generate a simulated dataset according to the MARAC(P, Q) model specified by (5) and (7). We start with the simple case where $(P, Q) = (1, 1)$ and we specify D, i.e. the dimension of the auxiliary time series, as 3. We simulate the autoregressive coefficients $(\mathbf{A}_1, \mathbf{B}_1)$ such that they satisfy the stationarity condition specified in Theorem 4.1.2 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 2, we plot the simulated $\mathbf{A}_1, \mathbf{B}_1, \mathbf{\Sigma}_r, \mathbf{\Sigma}_c$ when $(M, N) = (20, 20)$.

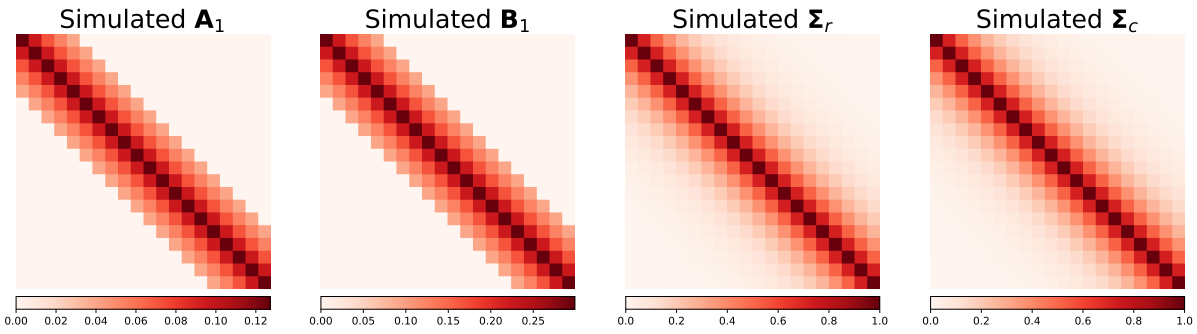

Figure 2: *Heatmap of simulated* $\mathbf{A}_1, \mathbf{B}_1, \mathbf{\Sigma}_r, \mathbf{\Sigma}_c$ *with* $(M, N) = (20, 20)$.

To simulate the functional parameters $g_1, g_2, g_3 \in \mathcal{H}_K$,we select the Lebedev kernel [Kennedy et al., 2013] as the kernel for $\mathcal{H}_K$. We provide details of the Lebedev kernel and its eigen-decomposition in Section E.1 of the appendix. We generate the random functions

g_1, g_2, g_3 by simulating from a zero-mean Gaussian Process with the Lebedev kernel and simulate the associated 3-dimensional vector time series $\{\mathbf{z}_t\}_{t=1}^{T}$ from a stationary VAR(1) process. In Figure 3, we visualize a sample of the functional parameters $g_{1:3}$ (evaluated on a 20×20 spatial grid) and the corresponding 3-dimensional vector time series ($T = 1000$).

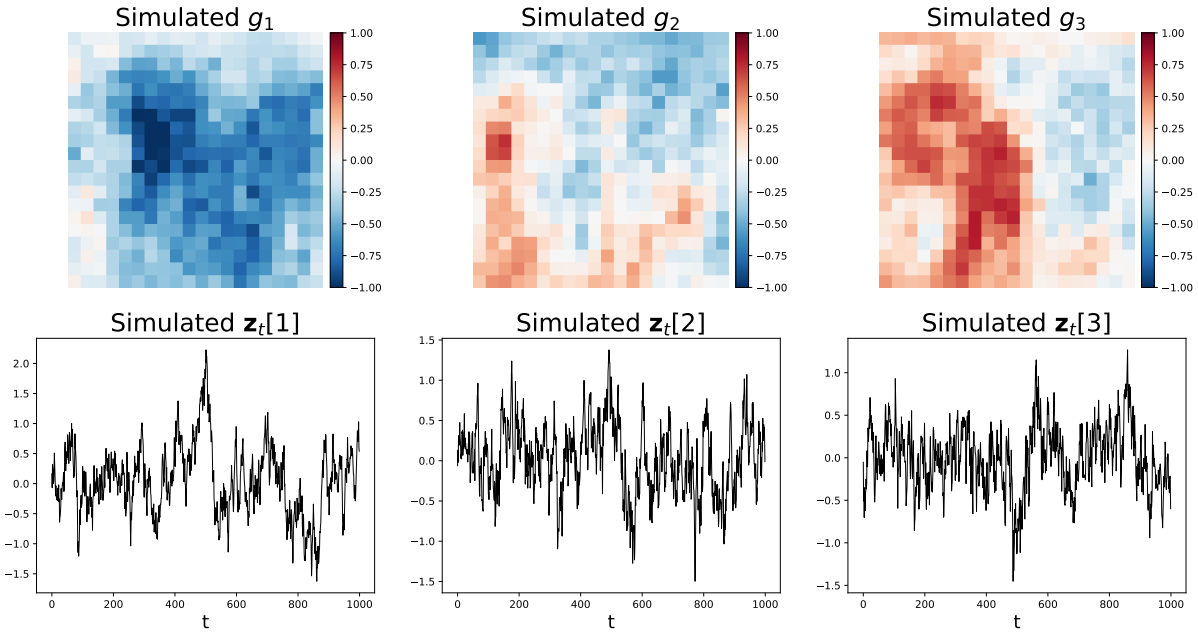

Figure 3: *Simulated functional parameters $g_{1:3}$ evaluated on a 20×20 spatial grid (top row) and the corresponding 3-dimensional auxiliary vector time series (bottom row).*

5.2 Model Estimator Validation

To investigate the impact of growing spatial dimensionality (M, N) and sample size T on the accuracy of the inference and prediction, we consider generating data from a MARAC(1,1) model with $(M, N) \in \{(5, 5), (10, 10), (20, 20), (40, 40)\}$ and $T_{\text{train}} \in \{100, 500, 1000, 2000, 4000, 8000, 16000\}$. We fit the model with the penalized maximum likelihood approach detailed in Section 3.1. Apart from the training set with varying T, we set aside a validation set and a testing set with $T_{\text{val}} = T_{\text{train}}/2$ and $T_{\text{test}} = 5000$, for choosing the tuning parameter λ and evaluating the prediction performances respectively. We evaluate the element-wise

root mean-squared error of the model estimates and predictions:

$$\mathrm{RMSE}(\widehat{\mathbf{A}}_1) = \frac{\|\widehat{\mathbf{A}}_1 - \mathbf{A}_1\|_{\mathrm{F}}}{M}, \quad \mathrm{RMSE}(\widehat{\mathbf{B}}_1) = \frac{\|\widehat{\mathbf{B}}_1 - \mathbf{B}_1\|_{\mathrm{F}}}{N}, \quad \mathrm{RMSE}(\widehat{g}) = \frac{1}{3}\sum_{d=1}^{3} \frac{\|\widehat{g}_d - g_d\|_2}{\sqrt{MN}};$$

$$\mathrm{RMSE}(\widehat{\mathbf{\Sigma}}_c \otimes \widehat{\mathbf{\Sigma}}_r) = \frac{\|\widehat{\mathbf{\Sigma}}_c \otimes \widehat{\mathbf{\Sigma}}_r - \mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r\|_{\mathrm{F}}}{MN}, \quad \mathrm{RMSE}_{\mathrm{pred}} = \sum_{t=2}^{T_{\mathrm{test}}} \frac{\|\widehat{\mathbf{X}}_t - \mathbf{X}_t\|_{\mathrm{F}}}{(T_{\mathrm{test}}-1)\sqrt{MN}};$$

where $\|\widehat{g}_d - g_d\|_2$ is the ℓ_2 norm of the functional parameters over the $M \times N$ spatial grid. We show the five RMSE metrics (rows) for the four choices of (M, N) (columns) with growing sample sizes T (x-axis for each plot) in Figure 4.

Figure 4 reveals that the parameter estimators $\widehat{\mathbf{A}}_1, \widehat{\mathbf{B}}_1$, $\widehat{\mathbf{\Sigma}}_c \otimes \widehat{\mathbf{\Sigma}}_r$ and $\widehat{g}$ have higher accuracy as sample sizes grows from $T_{\mathrm{train}} = 100$ to $T_{\mathrm{train}} = 16,000$; and for $\widehat{\mathbf{B}}_1$, higher spatial resolution leads to higher errors given the same sample size. With fixed spatial resolution, empirically, the convergence rate of $\widehat{\mathbf{B}}_1 \otimes \widehat{\mathbf{A}}_1$ is at the order of $1/\sqrt{T}$, coinciding with our theoretical result. The same convergence rate is achieved for $\widehat{g}$ and $\widehat{\mathbf{\Sigma}}_c \otimes \widehat{\mathbf{\Sigma}}_r$.

The test set performance improves as sample size grows. Given relatively large sample size ($T_{\mathrm{train}} \geq 1000$), all spatial dimensions show near-optimal prediction performance (the error matrices have their diagonals fixed at unity, so the optimal RMSE is 1).

The numerical experiments conducted above use the penalized maximum likelihood estimation (PMLE) instead of the truncated PMLE introduced in Section 3.2 for exact estimation of the functional parameter $g_{1:3}$. In Figure 5, we compare the PMLE estimator and the truncated PMLE estimator with different choices of the number of basis functions R against the ground truth of g_3. 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.

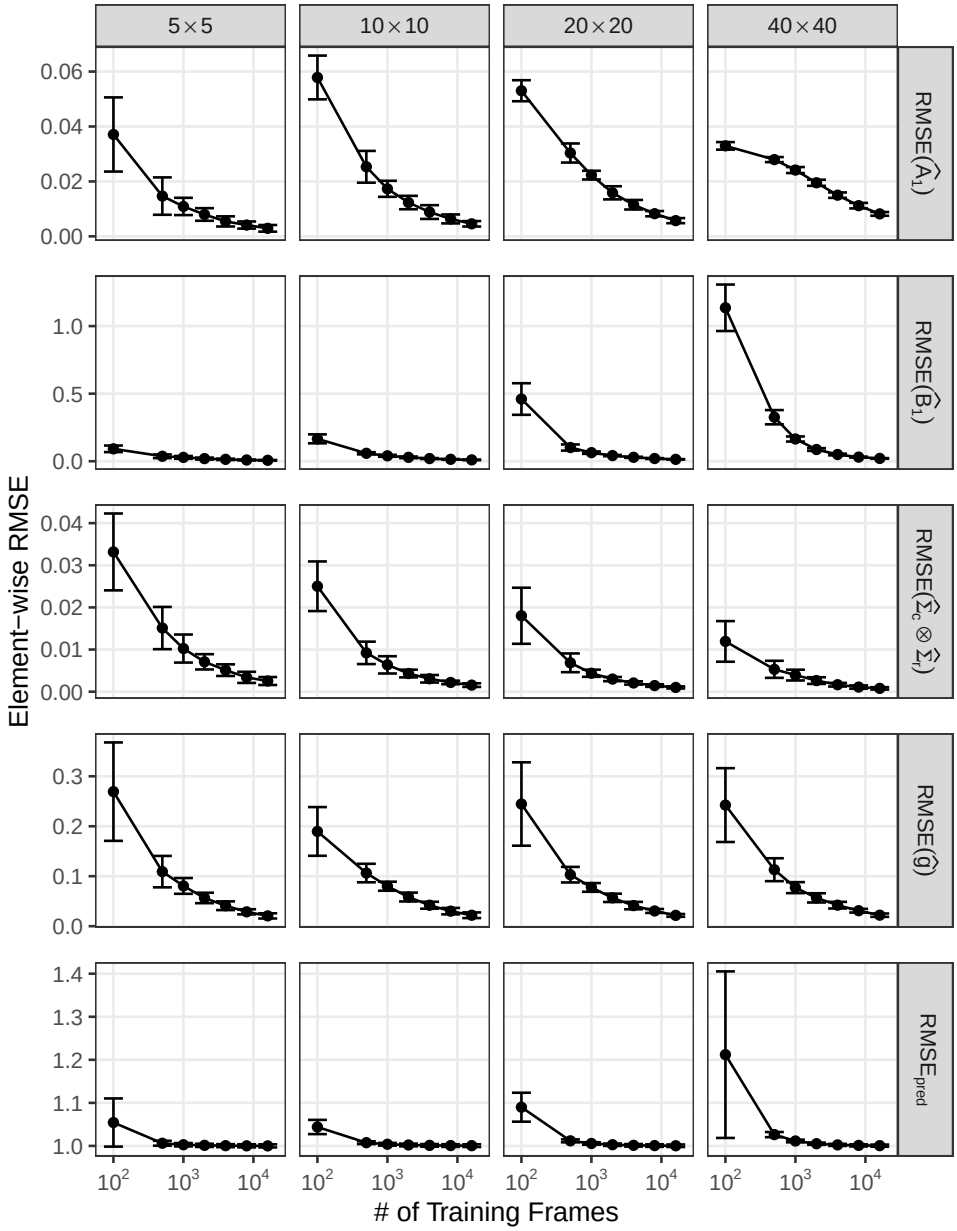

Figure 4: *MARAC(1,1) model estimator validation results. Element-wise RMSE for parameter estimators* $\widehat{\mathbf{A}}_1, \widehat{\mathbf{B}}_1$, $\widehat{\mathbf{\Sigma}}_c \otimes \widehat{\mathbf{\Sigma}}_r$, $\widehat{g}$ *and the test set prediction RMSE under four spatial resolutions* $(M, N) \in \{(5,5), (10,10), (20,20), (40,40)\}$ *with growing training sample sizes* $T_{train} \in \{100, 500, 1000, 2000, 4000, 8000, 16000\}$ *are plotted. Errorbars are plotted for* ± 1.96 *standard deviation, and results are based on 20 independent iterations.*

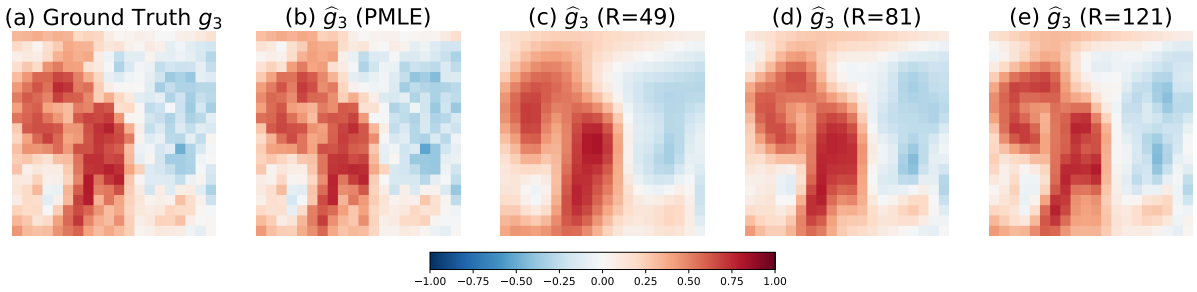

Figure 5: *Ground truth g_3 (panel (a)) against the PMLE estimator $\widehat{g}_3$ (panel (b)) and the truncated PMLE estimator $\widehat{g}_3$ with $R \in \{49, 81, 121\}$ basis functions.* $(M, N) = (20, 20)$.

5.3 Lag Selection with Information Criteria

The MARAC model has three tuning parameters: autoregressive lag P, auxiliary covariate lag Q, and the functional norm penalty weight λ. In practice, λ is chosen based on the prediction performance on the validation set. The choice of (P, Q) requires a more formal model selection procedure. 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]; and we validate their efficacy with simulations. In the context of the MARAC(P, Q) model, whose parameter can be summarised as $\boldsymbol{\Theta} \coloneqq \big\{\{\mathbf{A}_p, \mathbf{B}_p\}_{p\in[P]}, \{g_{q,d}\}_{q\in[Q],d\in[D]}, \{\boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_c\}\big\}$, the AIC and BIC can be written as:

$$\mathrm{AIC}(P, Q, \lambda) = -\frac{2}{T}\ell(\widehat{\boldsymbol{\Theta}}) + \frac{2}{T}\mathbf{df}(P, Q, \lambda), \tag{29}$$

$$\mathrm{BIC}(P, Q, \lambda) = -\frac{2}{T}\ell(\widehat{\boldsymbol{\Theta}}) + \frac{\log(T)}{T}\mathbf{df}(P, Q, \lambda), \tag{30}$$

where $\ell(\cdot)$ is the log likelihood of the training data, T is the total number of training time points and $\mathbf{df}(P, Q, \lambda)$ is the *effective degrees of freedom* of the MARAC(P, Q) model.

Given fixed matrix dimensionality at $M \times N$, the autoregressive parameters $\{\mathbf{A}_p, \mathbf{B}_p\}_{p=1}^{P}$ has $P \times (M^2 + N^2 - 1)$ degrees of freedom, where the one subtracted accounts for the rescaling step for parameter identifiability. The error covariance components $\boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_c$ have

$(M^2 + N^2)$ degrees of freedom. For the functional parameters $\{g_{q,d}\}_{q\in[Q],d\in[D]}$, we use the effective degrees of freedom for kernel ridge regression and define the degrees of freedom for each set of D functions $\mathbf{g}_q = (g_{q,1}, g_{q,2}, \ldots, g_{q,D})$ as:

$$\mathbf{df}(\mathbf{g}_q|\lambda) = \mathbf{tr}\left(\left[\left(\frac{\sum_t \mathbf{z}_{t-q}\mathbf{z}_{t-q}^\top}{T}\right) \otimes \mathbf{\Sigma}^{-1}\mathbf{K} + \lambda \mathbf{I}_{DMN}\right]^{-1}\left[\left(\frac{\sum_t \mathbf{z}_{t-q}\mathbf{z}_{t-q}^\top}{T}\right) \otimes \mathbf{\Sigma}^{-1}\mathbf{K}\right]\right).$$

Thus the total degrees of freedom for all functional parameters are $\sum_{q=1}^{Q} \mathbf{df}(\mathbf{g}_q|\lambda)$. A special case is when $\lambda \to 0$, we have $\mathbf{df}(\mathbf{g}_q|\lambda) \to DMN$. This implies that each auxiliary covariate at lag q has MN regression coefficients, equivalent to assigning *element-wise* regression coefficient without any spatial correlation.

We set up a simulation experiment to demonstrate the consistency of the two information criteria for selecting the correct model. We set the ground truth $(P, Q) = (2, 2)$, and choose the candidate models from $\{(P, Q) \in \mathbb{N}^2 | 1 \leq P, Q \leq 4\}$, and $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$. We fix the matrix dimension as 5×5. We fit the model with all possible combinations of (P, Q, λ) with training sample size $T \in \{1000, 2000, 4000, 8000\}$, and repeat the experiments 100 times. In Table 1, we list the probability that the true model, i.e. $(P, Q) = (2, 2)$, is selected by the AIC and BIC under different sample sizes. We find that AIC tends to select the model with more autoregressive lags but BIC performs 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), but BIC can be a good alternative.

In Table 3 of Section E.2 of the appendix, we also compare the test set prediction RMSE of models with different (P, Q) and show that the prediction is significantly worse when either P or Q is less than the correct lag order 2.

	$T = 1000$	$T = 2000$	$T = 4000$	$T = 8000$
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 1: *Probability that AIC and BIC select the correct P (first number), Q (second number) and (P, Q) (third number) from 100 repeated runs.*

6 Application: Global TEC Forecast

As a real data application, we choose the problem of forecasting the global total electron content (TEC).Ionosphere TEC is defined as the density of electrons along the path connecting the satellite of the Global Navigation Satellite Systems (GNSS) radio transmitter and a ground-based receiver, and 1 TEC unit (TECu) is 10^{16} electrons/m^2. TEC affects the propagation of radio waves, leading to positioning error in the GNSS Positioning, Navigation and Timing (PNT) services. Better knowledge of the distribution of the global TEC can make the PNT services more accurate. At a particular timestamp, the TEC is measured with a spatial resolution of 1°latitude $\times$ 1°longitude, and the temporal cadence is 5 minutes. The geophysical community is interested in forecasting the future TEC map given the history of observed TEC maps [Liu et al., 2020], together with multiple geophysical parameters related to the global ionosphere activity (see Figure 1D as an example). The original TEC data from the Madrigal TEC database [Rideout and Coster, 2006] has more than 80% of the data missing, on average, for every 181×361 spatial grid during $2005 \sim 2020$. In Sun et al. [2022], a matrix completion method called VISTA is proposed to fill in the missing entries with high accuracy, and in this paper, we use the imputed global TEC database named the VISTA database [Sun et al., 2023]. Figure 1 gives examples of the VISTA TEC maps along with two selected geophysical parameters.

The TEC forecast problem is typically more challenging when the geomagnetic activity magnitude is high. Such periods are termed stormy periods and the counterpart where the activity magnitude is low is termed non-stormy periods. We train our MARAC models for predicting the global TEC for stormy and non-stormy periods, separately. In Table 2, we list the periods of time from which we collect the training, validation and testing data, based on the geomagnetic activity magnitude of these periods. Note that stormy data are from year 2015 and non-stormy data are from year 2017.

Status	Train (24 days)	Validation (6 days)	Test (14 days)
Storm	$1.7 \sim 1.12$ $3.17 \sim 3.25$ $6.22 \sim 6.30$	$8.26 \sim 8.31$	$10.6 \sim 10.12$ $12.19 \sim 12.25$
Nonstorm	$3.1 \sim 3.24$	$4.1 \sim 4.6$	$5.1 \sim 5.14$

Table 2: *Periods for training, validating and testing the MARAC models for stormy (from 2015) and non-stormy periods (from 2017).*

To speed up the computation and reduce the consumption of the memory, we downsample the TEC time series from its original 5-min cadence to 1-hour cadence. Consequently, for both stormy and non-stormy periods, we have a training data of size $181 \times 361 \times 576$. Note that the training data for the stormy periods have three discontinuous sub-periods, containing three separate geomagnetic storms of 2015. We can easily accommodate this discontinuity of the training data by modifying the code for all parameter updates. We omit the details here but provide the implementation in our online repository.

We consider forecasting the first-order derivative of the TEC time series, denoted as $\{\Delta\text{TEC}_t\}_{t=1}^{T}$, by using the lagged ΔTEC and four auxiliary covariates. For a MARAC(P, Q)

model, we predict the ΔTEC_t as:

$$\Delta\text{TEC}_t = \sum_{p=1}^{P} \mathbf{A}_p \left(\Delta\text{TEC}_{t-p}\right) \mathbf{B}_p^\top + \sum_{q=1}^{Q} \mathcal{G}_q \times_3 \mathbf{z}_{t-q}^\top + \mathbf{E}_t, \tag{31}$$

where $\mathbf{z}_{t-q}$ contains the IMF Bz, SYM-H, $\cos(2\pi\text{ToD}/24)$, $\sin(2\pi\text{ToD}/24)$, with ToD being the Time-of-Day, taking values from [0 UT, 24 UT] (UT: Universal Time). The sine and cosine component of ToD accounts for the within-day cycle of the TEC. The IMF Bz and SYM-H time series are from the OMNIWeb database from NASA SPDF [Papitashvili and King, 2020], with an original cadence at 5-min. For any particular time where TEC is measured, we take the average of these two quantities over the previous 1 hour as the auxiliary time series. To make prediction of the next hour's TEC, we simply add the predicted ΔTEC for the next hour to the current hour's observed TEC.

For both stormy and non-stormy periods, we train MARAC models by choosing different (P, Q, λ), where $P \in \{1, 2, 3\}, Q \in \{0, 1, 2, 3\}$ and $\lambda \in [0.001, 0.01, 0.1, 1, 10]$. We choose λ based on the validation set prediction performance and (P, Q) based on the BIC, as discussed in Section 5.3. We use the truncated PMLE method with $R = 81$ spherical harmonic basis functions from the Lebedev kernel with $\eta = 3$. We can only use the truncated PMLE here since the spatial resolution of the data, 181×361, is ultra-high.

For stormy periods, we end up with MARAC$(2, 1)$ with $\lambda = 0.01$, and for non-stormy periods we finally choose MARAC$(3, 0)$, which reduced to the MAR model. The AIC also selects the same set of models. The model selection result indicates that the auxiliary covariates are more useful for TEC prediction when there are geomagnetic storms during which the Earth's ionosphere is impacted more heavily by the solar wind.

In Figure 6, we show, in the first row, the day-by-day prediction RMSE for the two stormy periods and the one non-stormy periods of the testing set. As a comparison, we also plot the RMSE of the persistence model, which predicts the next hour's TEC using

the current TEC, as the baseline. It is evident that the MARAC model in all three cases outperform the baseline. Also, the scale of the RMSE is very different for the stormy and non-stormy periods. In the second and third row, we select one sample frame of the TEC from each of the three testing periods and visualize its ground truth (second row) and the prediction made by the MARAC model (third row). Overall, the predicted TEC map resembles the ground truth, but the predictions seems smoother than the ground truth. We visualize the model parameters in Section F of the appendix.

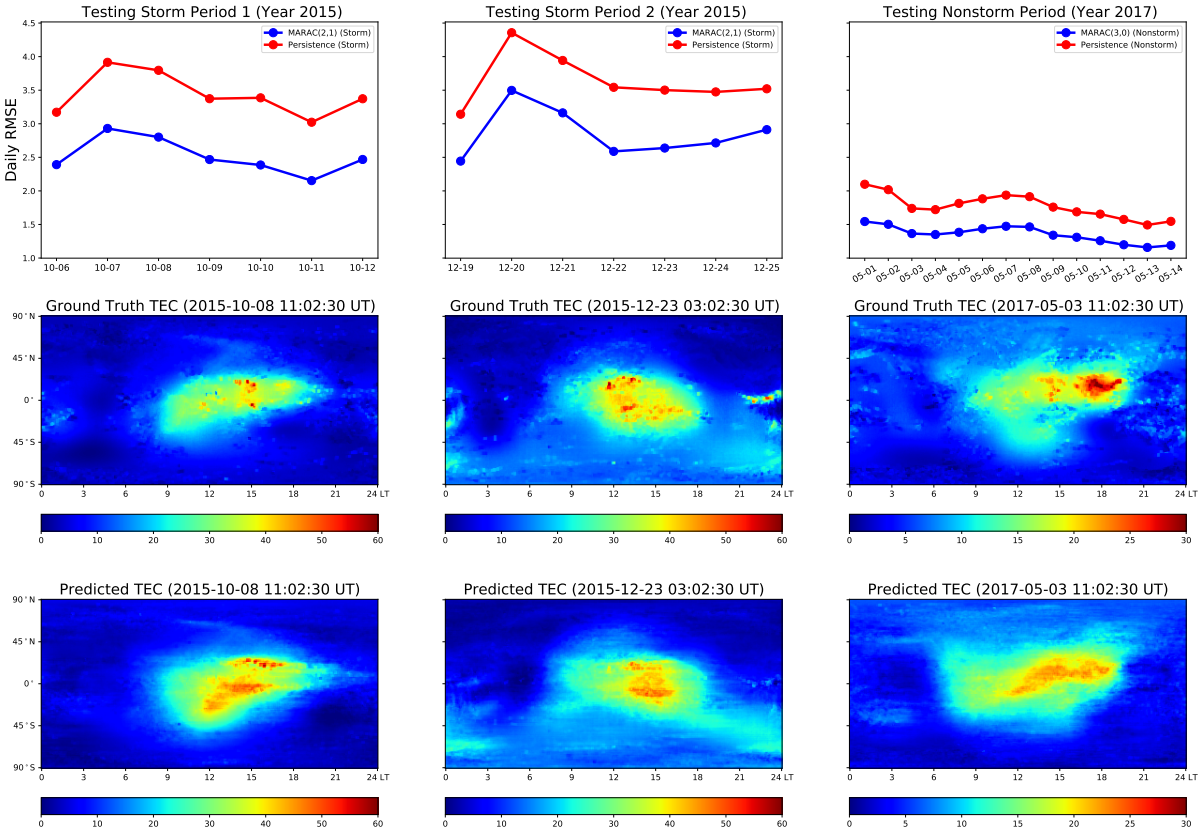

Figure 6: *Row 1: MARAC model prediction RMSE for the three testing periods against the RMSE of the persistence model. Row 2: Ground truths of selected frames of TEC from the three testing periods. Row 3: Predictions made by the MARAC model.*

7 Conclusion

In this paper, we introduce a new methodology for matrix-valued time series autoregression with spatio-temporal data under the presence of auxiliary vector time series covariates. The method, called MARAC, is a matrix counterpart of the ARMAX model where the ARMA model is augmented by exogenous covariates. The model contains an autoregressive component with bi-linear transformation on the lagged matrix predictors and an additive auxiliary component that transforms vector predictors to matrix outputs via tensor mode-product. The tensor coefficients of the auxiliary covariates are assumed to be discrete evaluations of functions from an RKHS with 2D domain and we propose to estimate the functional parameters, together with the autoregression parameters under a unified penalized MLE approach. We further establish the stationarity condition of the model and the large sample asymptotics of the model estimators. Both simulations and real data application to global TEC forecast show the effectiveness of our methodology.

To conclude, we list a few extensions of our work. First, it would be worthwhile to establish the optimal tuning order of λ as the spatial resolution of the data grows with sample size T. Second, our algorithm is not yet scalable to massive datasets and it would be beneficial to convert it to an alternative where data can be loaded in batches. Finally, the high dimensionality of the parameters make the inference of the model computationally intractable and we leave the uncertainty quantification of the model for future work.

8 Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant NSF-PHY 2027555 and NSF-DMS 113397.

References

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

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.

Elynn Y Chen, Xin Yun, Rong Chen, and Qiwei Yao. Modeling Multivariate Spatial-Temporal Data with Latent Low-dimensional Dynamics. *arXiv preprint arXiv:2002.01305*, 2020.

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

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

Noel Cressie 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.

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

Kristjan Greenewald, Shuheng Zhou, Alfred Hero, et al. Tensor Graphical Lasso (TeraLasso). *Journal of the Royal Statistical Society Series B*, 81(5):901–931, 2019.

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

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

Botao Hao, Boxiang Wang, Pengyuan Wang, Jingfei Zhang, Jian Yang, and Will Wei Sun. Sparse Tensor Additive Regression. *Journal of machine learning research*, 22, 2021.

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.

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.

Hung Hung and Chen-Chien Wang. Matrix Variate Logistic Regression Model with Application to EEG Data. *Biostatistics*, 14(1):189–202, 2013.

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.

Jean Kossaifi, Zachary C Lipton, Arinbjörn Kolbeinsson, Aran Khanna, Tommaso Furlanello, and Anima Anandkumar. Tensor Regression Networks. *The Journal of Machine Learning Research*, 21(1):4862–4882, 2020.

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.

Lei Liu, Shasha Zou, Yibin Yao, and Zihan Wang. Forecasting Global Ionospheric TEC using Deep Learning Approach. *Space Weather*, 18(11):e2020SW002501, 2020.

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

Eric F Lock and Gen Li. Supervised Multiway Factorization. *Electronic journal of statistics*, 12(1):1150, 2018.

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`.

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

William Rideout and Anthea Coster. Automated GPS Processing for Global Total Electron Content Data. *GPS solutions*, 10:219–228, 2006.

Ankan Saha and Ambuj Tewari. On the Nonasymptotic Convergence of Cyclic Coordinate Descent Methods. *SIAM Journal on Optimization*, 23(1):576–601, 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.

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, Yang Chen, Shasha Zou, Jiaen Ren, Yurui Chang, Zihan Wang, and Anthea Coster. Complete Global Total Electron Content Map Dataset based on a Video Imputation Algorithm VISTA. *Scientific Data*, 10(1):236, 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.

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.

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.

Yu Wang, Byoungwook Jang, and Alfred Hero. The Sylvester Graphical Lasso (syglasso). In *International Conference on Artificial Intelligence and Statistics*, pages 1943–1953. PMLR, 2020.

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 Complete Algorithm for Penalized MLE

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

Input: Matrix time series: $\mathbf{X}_t \in \mathbb{R}^{M \times N}$; Auxiliary vector time series: $\mathbf{z}_t \in \mathbb{R}^D$.

Parameter: Autoregression coefficients: $\{\mathbf{A}_p, \mathbf{B}_p\}_{p\in[P]}$; Auxiliary covariates coefficients: $\{\boldsymbol{\gamma}_q\}_{q\in[Q]}$; Error covariance components: $\boldsymbol{\Sigma}_c, \boldsymbol{\Sigma}_r$.

Randomly initialize parameters $\{\mathbf{A}_p^{(0)}, \mathbf{B}_p^{(0)}\}_{p\in[P]}$, $\{\boldsymbol{\gamma}_q^{(0)}\}_{q\in[Q]}$. Randomly initialize $\boldsymbol{\Sigma}_c^{(0)}, \boldsymbol{\Sigma}_r^{(0)}$ with positive definite matrices. Set iteration counter $i = 1$.

while not converge and $i \leq$ max-iter **do**

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

    Update $\mathbf{A}_p^{(i+1)}$ with (16)

    Update $\mathbf{B}_p^{(i+1)}$ with (18)

  end for

  for $q = 1, 2, \ldots, Q$ **do**

    Update $\boldsymbol{\gamma}_q^{(i+1)}$ with (21)

  end for

  Update $\boldsymbol{\Sigma}_c^{(i+1)}$ and $\boldsymbol{\Sigma}_r^{(i+1)}$ with (22)

  $i \leftarrow i + 1$

end while

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

  $c_p \leftarrow \text{sign}\left(\mathbf{tr}(\mathbf{A}_p^{(i)})\right) \cdot \|\mathbf{A}_p^{(i)}\|_{\text{F}}$

  $\mathbf{A}_p^{(i)} \leftarrow c_p^{-1} \cdot \mathbf{A}_p^{(i)}$; $\mathbf{B}_p^{(i)} \leftarrow c_p \cdot \mathbf{B}_p^{(i)}$

end for

Output: $\{\mathbf{A}_p^{(i)}, \mathbf{B}_p^{(i)}\}_{p\in[P]}, \{\boldsymbol{\gamma}_q^{(i)}\}_{q\in[Q]}, \boldsymbol{\Sigma}_r^{(i)}, \boldsymbol{\Sigma}_c^{(i)}$.

B Proof of Theorem 4.1.2

Proof. Depending the relationship between Q and $\widetilde{Q}$, the proof will be slightly different but all cases lead to the same result. For simplicity, we give a proof for the case where $\widetilde{Q} > Q$ and the case where $\widetilde{Q} \leq Q$ can be proved similarly.

Given that $\widetilde{Q} > Q$, one can combine the MARAC(P, Q) process of $\{\mathbf{X}_t\}_{t=1}^{\infty}$ with the VAR$(\widetilde{Q})$ process of $\{\mathbf{z}_t\}_{t=1}^{\infty}$ and re-define a new vector time series $\mathbf{y}_t$:

$$\mathbf{y}_t = \left[\mathbf{vec}(\mathbf{X}_t)^\top, \ldots, \mathbf{vec}(\mathbf{X}_{t-P+1})^\top, \mathbf{z}_t^\top, \ldots, \mathbf{z}_{t-\widetilde{Q}+1}^\top\right]^\top$$

Then one can write down the VAR(1) process for $\{\mathbf{y}_t\}$ as:

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{AR}_{P\times P} & \mathbf{AC}_{P\times\widetilde{Q}} \\ \mathbf{O}_{\widetilde{Q}\times P} & \mathbf{C}_{\widetilde{Q}\times\widetilde{Q}} \end{bmatrix} \mathbf{y}_{t-1} + \mathbf{u}_t = \mathbf{F}\mathbf{y}_{t-1} + \mathbf{u}_t \tag{32}$$

where $\mathbf{U}_t$ is a noise process. The VAR(1) process has a coefficient matrix $\mathbf{F}$ that can be divided into 4 distinct blocks: $\mathbf{AR}_{P\times P}$, $\mathbf{AC}_{P\times\widetilde{Q}}$, $\mathbf{O}_{\widetilde{Q}\times P}$ and $\mathbf{C}_{\widetilde{Q}\times\widetilde{Q}}$, where all four blocks are block matrices by themselves and the block structure is shown as their subscripts. The four blocks are the matrix auto-regressive part of MARAC(P, Q), the auxiliary covariates part of MARAC(P, Q), the exogenous variable condition of $\mathbf{z}_t$ on $\mathbf{X}_t$ based on assumption 4.1.1 and the auto-regressive part of VAR$(\widetilde{Q})$, respectively.

Specifically for $\mathbf{AR}_{P\times P}$ and $\mathbf{C}_{\widetilde{Q}\times\widetilde{Q}}$:

$$\mathbf{AR}_{P\times P} = \begin{bmatrix} \mathbf{B}_1 \otimes \mathbf{A}_1 & \mathbf{B}_2 \otimes \mathbf{A}_2 & \ldots & \ldots & \mathbf{B}_P \otimes \mathbf{A}_P \\ \mathbf{I}_{MN} & \mathbf{O}_{MN\times MN} & \ldots & \ldots & \mathbf{O}_{MN\times MN} \\ \mathbf{O}_{MN\times MN} & \mathbf{I}_{MN} & \mathbf{O}_{MN\times MN} & \ldots & \mathbf{O}_{MN\times MN} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{O}_{MN\times MN} & \mathbf{O}_{MN\times MN} & \ldots & \mathbf{I}_{MN} & \mathbf{O}_{MN\times MN} \end{bmatrix}$$

$$\mathbf{C}_{\widetilde{Q}\times\widetilde{Q}} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 & \dots & \dots & \mathbf{C}_{\widetilde{Q}} \\ \mathbf{I}_D & \mathbf{O}_{D\times D} & \dots & \dots & \mathbf{O}_{D\times D} \\ \mathbf{O}_{D\times D} & \mathbf{I}_D & \mathbf{O}_{D\times D} & \dots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{O}_{D\times D} & \mathbf{O}_{D\times D} & \dots & \mathbf{I}_D & \mathbf{O}_{D\times D} \end{bmatrix}$$

And the stationarity of $\mathbf{y}_t$ requires that $\det(\lambda\mathbf{I} - \mathbf{F}) = 0$ has all the solutions within the unit circle in the complex plane. The determinant of $(\lambda\mathbf{I} - \mathbf{F})$ can be easily calculated by removing all identity matrices in $\mathbf{AR}_{P\times P}$ and $\mathbf{C}_{\widetilde{Q}\times\widetilde{Q}}$ via column operations. After the column operations, $(\lambda\mathbf{I} - \mathbf{F})$ becomes an upper triangular matrix whose diagonal is:

$$\text{diag}(\lambda\mathbf{I}-\mathbf{F}) = \left[\lambda\mathbf{I}_{MN} - \sum_{p=1}^{P}(\mathbf{B}_p \otimes \mathbf{A}_p)\,\lambda^{1-p}; \overbrace{\lambda\mathbf{I}_{MN}; \dots; \lambda\mathbf{I}_{MN}}^{P-1 \text{ blocks}}; \lambda\mathbf{I}_D - \sum_{q=1}^{\widetilde{Q}}\mathbf{C}_q\lambda^{1-q}; \overbrace{\lambda\mathbf{I}_D; \dots; \lambda\mathbf{I}_D}^{\widetilde{Q}-1 \text{ blocks}}\right]^\top$$

The determinant can then be evaluated by the product of the determinants of all matrices along the diagonal and replacing λ with $1/y$ gives the result in Theorem 4.1.2. □

C Proof of Proposition 4.2.1

Proof. For a MARAC(P, Q) model, by applying vectorization to both sides, one obtains:

$$\mathbf{x}_t = \mathbf{y}_t\boldsymbol{\theta} + \mathbf{e}_t \tag{33}$$

where $\mathbf{x}_t = \mathbf{vec}(\mathbf{X}_t), \mathbf{e}_t = \mathbf{vec}(\mathbf{E}_t)$ and $\mathbf{y}_t$ and $\boldsymbol{\theta}$ are defined as:

$$\mathbf{y}_t = \left[\mathbf{x}_{t-1}^\top \otimes \mathbf{I}; \dots; \mathbf{x}_{t-P}^\top \otimes \mathbf{I}; \mathbf{z}_{t-1}^\top \otimes \mathbf{K}; \dots; \mathbf{z}_{t-Q}^\top \otimes \mathbf{K}\right]$$

$$\boldsymbol{\theta} = \left[\mathbf{vec}(\boldsymbol{\Phi}_1)^\top; \dots; \mathbf{vec}(\boldsymbol{\Phi}_P)^\top; \mathbf{vec}(\boldsymbol{\gamma}_1)^\top; \dots; \mathbf{vec}(\boldsymbol{\gamma}_Q)^\top\right]^\top$$

The penalized negative log-likelihood loss function can thus be expressed in terms of an arbitrary choice of $\boldsymbol{\Sigma}$ and $\boldsymbol{\theta}$ as $\mathrm{L}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$:

$$\mathrm{L}(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{T}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\sum_t (\mathbf{x}_t - \mathbf{y}_t\boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_t - \mathbf{y}_t\boldsymbol{\theta}) + \frac{\lambda}{2}\boldsymbol{\theta}^\top\widetilde{\mathbf{K}}\boldsymbol{\theta} \tag{34}$$

where $\widetilde{\mathbf{K}}$ is:

$$\widetilde{\mathbf{K}} = \begin{bmatrix} \mathbf{O}_1 & \mathbf{O}_2 \\ \mathbf{O}_3 & \mathbf{I}_{QD} \otimes \mathbf{K} \end{bmatrix}$$

where $\mathbf{O}_1, \mathbf{O}_2, \mathbf{O}_3$ are zero matrices with proper dimensionality. Alternatively, we use the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ to replace $\boldsymbol{\Sigma}$ in (34) and then divide the loss function by T, we obtain the penalized loss function as:

$$h(\boldsymbol{\theta}, \boldsymbol{\Omega}) = -\frac{1}{2}\log\det(\boldsymbol{\Omega}) + \frac{1}{2}\mathbf{tr}\left(\boldsymbol{\Omega}\mathbf{S}(\boldsymbol{\theta})\right) + \frac{\lambda}{2T}\boldsymbol{\theta}^\top\widetilde{\mathbf{K}}\boldsymbol{\theta} \tag{35}$$

where $\mathbf{S}(\boldsymbol{\theta}) = \sum_t (\mathbf{x}_t - \mathbf{y}_t\boldsymbol{\theta})(\mathbf{x}_t - \mathbf{y}_t\boldsymbol{\theta})^\top/T$.

We use $\mathcal{H}_{\boldsymbol{\theta}}$ to denote the feasible set of $\boldsymbol{\theta}$ that satisfies the Kronecker product constraint over $\boldsymbol{\Phi}_{1:P}$:

$$\mathcal{H}_{\boldsymbol{\theta}} = \{\boldsymbol{\theta} | \boldsymbol{\theta} = \left[\mathbf{vec}(\boldsymbol{\Phi}_1)^\top; \ldots; \mathbf{vec}(\boldsymbol{\Phi}_P)^\top; \mathbf{vec}(\boldsymbol{\gamma}_1)^\top; \ldots; \mathbf{vec}(\boldsymbol{\gamma}_Q)^\top\right]^\top, \text{where } \boldsymbol{\Phi}_p = \mathbf{B}_p \otimes \mathbf{A}_p\}$$

and similarly we use $\mathcal{H}_{\boldsymbol{\Omega}}$ to denote the feasible set of $\boldsymbol{\Omega}$ that satisfies the Kronecker product constraint of $\boldsymbol{\Omega}$:

$$\mathcal{H}_{\boldsymbol{\Omega}} = \{\boldsymbol{\Omega} | \boldsymbol{\Omega} = \boldsymbol{\Sigma}_c^{-1} \otimes \boldsymbol{\Sigma}_r^{-1}, \text{where } \boldsymbol{\Sigma}_c, \boldsymbol{\Sigma}_r \text{ are positive definite}\}$$

To prove the consistency of $\widehat{\boldsymbol{\Sigma}}$, we seek to prove that:

$$\mathrm{P}\left(\inf_{\|\bar{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\| \geq c} \inf_{\bar{\boldsymbol{\theta}} \in \mathcal{H}_{\boldsymbol{\theta}}} h(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\Omega}}) \leq h(\boldsymbol{\theta}, \boldsymbol{\Omega})\right) \to 0, \quad \text{as } T \to \infty \tag{36}$$

where we now use $\boldsymbol{\theta}, \boldsymbol{\Omega}$ to represent the ground truth. To prove (36), we find a lower bound of the term on the left hand side by optimizing over $\bar{\boldsymbol{\theta}}$, instead of within $\mathcal{H}_{\boldsymbol{\theta}}$, in an *unconstrained* manner. Given any arbitrary $\bar{\boldsymbol{\Omega}}$, one can find the optimal *unconstrained* $\bar{\boldsymbol{\theta}}$ by:

$$\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}(\bar{\boldsymbol{\Omega}}) = \left(\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{y}_t + \lambda\widetilde{\mathbf{K}}\right)^{-1} \left(\sum_t \mathbf{y}_t^\top \bar{\boldsymbol{\Omega}} \mathbf{x}_t\right) \tag{37}$$

Plugging this into the loss function $h(\cdot,\cdot)$, we obtain the profile likelihood with respect to $\bar{\mathbf{\Omega}}$ as $\ell(\bar{\mathbf{\Omega}})$:

$$\ell(\bar{\mathbf{\Omega}}) = -\frac{1}{2}\log\det(\bar{\mathbf{\Omega}}) + \frac{1}{2}\mathbf{tr}\left(\bar{\mathbf{\Omega}}\frac{\sum_t(\mathbf{x}_t - \mathbf{y}_t\bar{\boldsymbol{\theta}})\mathbf{x}_t^\top}{T}\right) \tag{38}$$

We now convert the proposition in (36) to:

$$\mathrm{P}\left(\inf_{\|\bar{\mathbf{\Omega}}-\mathbf{\Omega}\|\geq c}\ell(\bar{\mathbf{\Omega}}) \leq h(\boldsymbol{\theta},\mathbf{\Omega})\right) \to 0, \quad \text{as } T\to\infty$$

Note that $\bar{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ under the assumption that $\lambda/T \to 0$ and the matrix/vector time series are jointly stationary. Therefore, one can re-express the profile likelihood $\ell(\bar{\mathbf{\Omega}})$ as:

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

Using $\widetilde{\mathbf{\Omega}}$ to denote the *unconstrained* minimizer of $\widetilde{\ell}(\cdot)$, then in Theorem 4 of Chen et al. [2021], it is shown that with probability approaching one:

$$\widetilde{\ell}(\bar{\mathbf{\Omega}}) \geq \widetilde{\ell}(\widetilde{\mathbf{\Omega}}) + \frac{c^2}{32}\lambda^2_{\min}(\mathbf{\Omega}^{-1})$$

Therefore, we have $\mathrm{P}\left(\ell(\bar{\mathbf{\Omega}}) > \ell(\widetilde{\mathbf{\Omega}})\right) \to 1, \quad \text{as } T\to\infty$. Finally, notice that $\widetilde{\mathbf{\Omega}} \xrightarrow{p} \mathbf{\Omega}$ and $\bar{\boldsymbol{\theta}}(\widetilde{\mathbf{\Omega}}) \xrightarrow{p} \boldsymbol{\theta}$, we have $\ell(\widetilde{\mathbf{\Omega}}) \xrightarrow{p} h(\boldsymbol{\theta},\mathbf{\Omega})$. Thus, we have:

$$\mathrm{P}\left(\ell(\bar{\mathbf{\Omega}}) \leq \inf_{\|\bar{\mathbf{\Omega}}-\mathbf{\Omega}\|\geq c}\inf_{\bar{\boldsymbol{\theta}}\in\mathcal{H}_{\boldsymbol{\theta}}} h(\bar{\boldsymbol{\theta}},\bar{\mathbf{\Omega}}) \leq h(\boldsymbol{\theta},\mathbf{\Omega})\right) \to 0, \quad \text{as } T\to\infty$$

and thereby completes the proof. □

D Proof of Theorem 4.2.2

We start by proving the convergence rate of the penalized maximum likelihood estimators (PMLE) in the following lemma:

Lemma D.0.1. *Given the same assumption as Theorem 4.2.2, one can establish the convergence rate for the model estimators* $\{(\widehat{\mathbf{A}}_p,\widehat{\mathbf{B}}_p)\}_{p=1}^P, \{\widehat{\boldsymbol{\gamma}_q}\}_{q=1}^Q$ *as:*

$$\widehat{\mathbf{A}}_p = \mathbf{A}_p + O_P(T^{-\frac{1}{2}}), \widehat{\mathbf{B}}_p = \mathbf{B}_p + O_P(T^{-\frac{1}{2}}), \widehat{\boldsymbol{\gamma}_q} = \boldsymbol{\gamma}_q + O_P(T^{-\frac{1}{2}}), \quad \forall 1\leq p\leq P, 1\leq q\leq Q$$

Proof. We begin the proof by using the same notation as the proof of Proposition 4.2.1, namely we use $\mathbf{y}_t$ and $\boldsymbol{\theta}$ to denote:

$$\mathbf{y}_t = \left[\mathbf{x}_{t-1}^\top \otimes \mathbf{I}; \ldots; \mathbf{x}_{t-P}^\top \otimes \mathbf{I}; \mathbf{z}_{t-1}^\top \otimes \mathbf{K}; \ldots; \mathbf{z}_{t-Q}^\top \otimes \mathbf{K}\right], \quad \mathbf{x}_t \coloneqq \mathbf{vec}(\mathbf{X}_t)$$

$$\boldsymbol{\theta} = \left[\mathbf{vec}(\boldsymbol{\Phi}_1)^\top; \ldots; \mathbf{vec}(\boldsymbol{\Phi}_P)^\top; \mathbf{vec}(\boldsymbol{\gamma}_1)^\top; \ldots; \mathbf{vec}(\boldsymbol{\gamma}_Q)^\top\right]^\top$$

and thus the MARAC(P, Q) model has the form of $\mathbf{x}_t = \mathbf{y}_t\boldsymbol{\theta} + \mathbf{e}_t$, where $\mathbf{e}_t = \mathbf{vec}(\mathbf{E}_t)$. To show the convergence rate result, we first show that the following result holds for the loss function $h(\boldsymbol{\theta}, \boldsymbol{\Omega})$ in (35):

$$\mathrm{P}\left(\inf_{\sqrt{T}\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\|\geq c_T} \inf_{\bar{\boldsymbol{\Omega}}\in\mathcal{H}_{\boldsymbol{\Omega}}} h(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\Omega}}) \leq h(\boldsymbol{\theta}, \boldsymbol{\Omega})\right) \longrightarrow 0, \quad \forall c_T \to \infty, \quad \text{as } T \to \infty \tag{39}$$

To show this, we modify the notation of $\mathbf{y}_t$ and $\boldsymbol{\theta}$ a bit by applying vectorization on both sides of $\mathbf{x}_t = \mathbf{y}_t\boldsymbol{\theta} + \mathbf{e}_t$ and get:

$$\mathbf{x}_t = \left(\boldsymbol{\theta}^\top \otimes \mathbf{I}_{MN}\right) \mathbf{vec}\left(\mathbf{y}_t\right) + \mathbf{e}_t = \boldsymbol{\Theta}\mathbf{vec}\left(\mathbf{y}_t\right) + \mathbf{e}_t$$

Then let $\mathcal{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$ and $\mathcal{Y} = [\mathbf{vec}\left(\mathbf{y}_1\right), \ldots, \mathbf{vec}\left(\mathbf{y}_T\right)]$, we can rewrite the loss function $h(\boldsymbol{\theta}, \boldsymbol{\Omega})$ as $\widetilde{h}(\boldsymbol{\Theta}, \boldsymbol{\Omega})$, where:

$$\widetilde{h}(\boldsymbol{\Theta}, \boldsymbol{\Omega}) = -\frac{1}{2}\log\det(\boldsymbol{\Omega}) + \frac{1}{2}\mathbf{tr}\left(\boldsymbol{\Omega}(\mathcal{X} - \boldsymbol{\Theta}\mathcal{Y})(\mathcal{X} - \boldsymbol{\Theta}\mathcal{Y})^\top/T\right) + O_P(T^{-1})$$

where we represent the functional penalty term as $O_P(T^{-1})$ since we assume that λ is bounded as $T \to \infty$. To prove (39), we first introduce $\widetilde{\boldsymbol{\Theta}}$ to denote the unconstrained OLS solution to the regression problem, namely $\widetilde{\boldsymbol{\Theta}} = \arg\min \|\mathcal{X} - \boldsymbol{\Theta}\mathcal{Y}\|^2$ and note that $\widetilde{\boldsymbol{\Theta}} = \boldsymbol{\Theta} + O_P(T^{-1/2})$. The left hand side of (39) can be rewritten as:

$$\inf_{\sqrt{T}\|\bar{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|\geq c_T} \inf_{\bar{\boldsymbol{\Omega}}\in\mathcal{H}_{\boldsymbol{\Omega}}} \widetilde{h}(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{\Omega}}) = \inf_{\bar{\boldsymbol{\Omega}}\in\mathcal{H}_{\boldsymbol{\Omega}}} \widetilde{h}(\widetilde{\boldsymbol{\Theta}}, \bar{\boldsymbol{\Omega}}) + \mathbf{tr}\left(\bar{\boldsymbol{\Omega}}(\bar{\boldsymbol{\Theta}} - \widetilde{\boldsymbol{\Theta}})\widehat{\boldsymbol{\Gamma}}(\bar{\boldsymbol{\Theta}} - \widetilde{\boldsymbol{\Theta}})^\top\right) \tag{40}$$

where $\widehat{\boldsymbol{\Gamma}} = \mathcal{Y}\mathcal{Y}^\top/T$, and $\widehat{\boldsymbol{\Gamma}} \xrightarrow{a.s.} \boldsymbol{\Gamma} = \mathrm{Cov}(\mathbf{vec}(\mathbf{y}_t), \mathbf{vec}(\mathbf{y}_t))$ by the assumption of joint stationarity (and thus ergodicity). The trace term on the right hand side of (40) can be

lower bounded as follows:

$$\begin{aligned}\mathbf{tr}\left(\bar{\mathbf{\Omega}}(\bar{\mathbf{\Theta}}-\widetilde{\mathbf{\Theta}})\widehat{\mathbf{\Gamma}}(\bar{\mathbf{\Theta}}-\widetilde{\mathbf{\Theta}})^\top\right) &\geq \lambda_{\min}(\bar{\mathbf{\Omega}})\cdot\lambda_{\min}(\widehat{\mathbf{\Gamma}})\cdot\|\bar{\mathbf{\Theta}}-\widetilde{\mathbf{\Theta}}\|_F^2 \\ &\geq \left(\frac{1}{2}\lambda_{\min}(\mathbf{\Omega})\right)\cdot\lambda_{\min}(\widehat{\mathbf{\Gamma}})\cdot\|\bar{\mathbf{\Theta}}-\widetilde{\mathbf{\Theta}}\|_F^2\end{aligned}$$

given that $\|\bar{\mathbf{\Omega}}-\mathbf{\Omega}\|\leq c$. Together with the consistency of $\widetilde{\mathbf{\Theta}}$, we have:

$$\mathrm{P}\left(\mathbf{tr}\left(\bar{\mathbf{\Omega}}(\bar{\mathbf{\Theta}}-\widetilde{\mathbf{\Theta}})\widehat{\mathbf{\Gamma}}(\bar{\mathbf{\Theta}}-\widetilde{\mathbf{\Theta}})^\top\right)\geq\frac{1}{4}\lambda_{\min}(\mathbf{\Omega})\cdot\lambda_{\min}(\mathbf{\Gamma})\cdot\frac{c_T^2}{T}\right)\to 1,\quad\text{as } T\to\infty \tag{41}$$

Similarly, we can show that when $\|\bar{\mathbf{\Omega}}-\mathbf{\Omega}\|\leq c$:

$$\inf_{\bar{\mathbf{\Omega}}\in\mathcal{H}_{\mathbf{\Omega}}}\widetilde{h}(\mathbf{\Theta},\bar{\mathbf{\Omega}})=\inf_{\bar{\mathbf{\Omega}}\in\mathcal{H}_{\mathbf{\Omega}}}\widetilde{h}(\widetilde{\mathbf{\Theta}},\bar{\mathbf{\Omega}})+\mathbf{tr}\left(\bar{\mathbf{\Omega}}(\mathbf{\Theta}-\widetilde{\mathbf{\Theta}})\widehat{\mathbf{\Gamma}}(\mathbf{\Theta}-\widetilde{\mathbf{\Theta}})^\top\right)=\inf_{\bar{\mathbf{\Omega}}\in\mathcal{H}_{\mathbf{\Omega}}}\widetilde{h}(\widetilde{\mathbf{\Theta}},\bar{\mathbf{\Omega}})+O_P(T^{-1}) \tag{42}$$

Combining (40), (41) and (42), we end up having the result that when $\|\bar{\mathbf{\Omega}}-\mathbf{\Omega}\|\leq c$:

$$\mathrm{P}\left(\inf_{\sqrt{T}\|\bar{\mathbf{\Theta}}-\mathbf{\Theta}\|\geq c_T}\inf_{\bar{\mathbf{\Omega}}\in\mathcal{H}_{\mathbf{\Omega}}}\widetilde{h}(\bar{\mathbf{\Theta}},\bar{\mathbf{\Omega}})\geq\inf_{\bar{\mathbf{\Omega}}\in\mathcal{H}_{\mathbf{\Omega}}}\widetilde{h}(\mathbf{\Theta},\bar{\mathbf{\Omega}})\right)\to 1,\quad\text{as } T\to\infty \tag{43}$$

Finally, notice that $\inf_{\bar{\mathbf{\Omega}}\in\mathcal{H}_{\mathbf{\Omega}}}\widetilde{h}(\mathbf{\Theta},\bar{\mathbf{\Omega}})\xrightarrow{p}\widetilde{h}(\mathbf{\Theta},\mathbf{\Omega})$, we thus eventually proved (39). The result states that any global minimizer of the loss function $\widetilde{h}(\cdot,\cdot)$ must have $\|\bar{\mathbf{\Theta}}-\mathbf{\Theta}\|\leq c_T/\sqrt{T}$, for any sequence of $c_T\to\infty$, with probability approaching one. This establishes the convergence rate result for all model estimators in the lemma. □

Equipped with the convergence rate result, we now move on to derive the asymptotic distribution of the model estimators in Theorem 4.2.2:

Proof. We start the proof by revisiting the updating formula of $\widehat{\mathbf{A}}_p$ in (16). One can re-arrange (16) and plug in the data-generating model for $\mathbf{X}_t$ based on MARAC(P,Q) and obtain:

$$\sum_t\left\{\sum_{i=1}^{P}\left[\left(\widehat{\mathbf{A}}_i-\mathbf{A}_i\right)\mathbf{X}_{t-i}\mathbf{B}_i^\top+\widehat{\mathbf{A}}_i\mathbf{X}_{t-i}\left(\widehat{\mathbf{B}}_i-\mathbf{B}_i\right)^\top\right]+\sum_{j=1}^{Q}\left(\widehat{\mathcal{G}}_j-\mathcal{G}_j\right)\times_3\mathbf{z}_{t-j}^\top-\mathbf{E}_t\right\}\widehat{\mathbf{\Sigma}}_c^{-1}\widehat{\mathbf{B}}_p\mathbf{X}_{t-p}^\top=\mathbf{O}$$

Left-multiplying both sides by $\widehat{\mathbf{\Sigma}}_r$ and using the convergence rate of $\{(\widehat{\mathbf{A}}_p, \widehat{\mathbf{B}}_p)\}_{p=1}^P$ and the consistency of $\widehat{\mathbf{\Sigma}}$, one can simplify the equation above by vectorizing both sides and get:

$$\begin{aligned}
&\sum_{i=1}^{P}\sum_{t}\left(\mathbf{X}_{t-p}\mathbf{B}_p^\top\mathbf{\Sigma}_c^{-1}\mathbf{B}_i\mathbf{X}_{t-i}^\top\otimes\mathbf{\Sigma}_r^{-1}\right)\mathbf{vec}\left(\widehat{\mathbf{A}}_i-\mathbf{A}_i\right)\\
+&\sum_{i=1}^{P}\sum_{t}\left(\mathbf{X}_{t-p}\mathbf{B}_p^\top\mathbf{\Sigma}_c^{-1}\otimes\mathbf{\Sigma}_r^{-1}\mathbf{A}_i\mathbf{X}_{t-i}\right)\mathbf{vec}\left(\widehat{\mathbf{B}}_i^\top-\mathbf{B}_i^\top\right)\\
+&\sum_{j=1}^{Q}\sum_{t}\left[\mathbf{z}_{t-j}^\top\otimes\left(\mathbf{X}_{t-p}\mathbf{B}_p^\top\mathbf{\Sigma}_c^{-1}\otimes\mathbf{\Sigma}_r^{-1}\right)\mathbf{K}\right]\mathbf{vec}\left(\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)=\sum_{t}\left(\mathbf{X}_{t-p}\mathbf{B}_p^\top\mathbf{\Sigma}_c^{-1}\otimes\mathbf{\Sigma}_r^{-1}\right)\mathbf{vec}\left(\mathbf{E}_t\right)+o_P(\sqrt{T})
\end{aligned}$$

Similarly, for all updating formula for $\widehat{\mathbf{B}}_p$, we have:

$$\begin{aligned}
&\sum_{i=1}^{P}\sum_{t}\left(\mathbf{\Sigma}_c^{-1}\mathbf{B}_i\mathbf{X}_{t-i}^\top\otimes\mathbf{X}_{t-p}^\top\mathbf{A}_p^\top\mathbf{\Sigma}_r^{-1}\right)\mathbf{vec}\left(\widehat{\mathbf{A}}_i-\mathbf{A}_i\right)\\
+&\sum_{i=1}^{P}\sum_{t}\left(\mathbf{\Sigma}_c^{-1}\otimes\mathbf{X}_{t-p}^\top\mathbf{A}_p^\top\mathbf{\Sigma}_r^{-1}\mathbf{A}_i\mathbf{X}_{t-i}\right)\mathbf{vec}\left(\widehat{\mathbf{B}}_i^\top-\mathbf{B}_i^\top\right)\\
+&\sum_{j=1}^{Q}\sum_{t}\left[\mathbf{z}_{t-j}^\top\otimes\left(\mathbf{\Sigma}_c^{-1}\otimes\mathbf{X}_{t-p}^\top\mathbf{A}_p^\top\mathbf{\Sigma}_r^{-1}\right)\mathbf{K}\right]\mathbf{vec}\left(\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)=\sum_{t}\left(\mathbf{\Sigma}_c^{-1}\otimes\mathbf{X}_{t-p}^\top\mathbf{A}_p^\top\mathbf{\Sigma}_r^{-1}\right)\mathbf{vec}\left(\mathbf{E}_t\right)+o_P(\sqrt{T})
\end{aligned}$$

And similarly, for all updating formula for $\widehat{\boldsymbol{\gamma}}_q$, we have:

$$\begin{aligned}
&\sum_{i=1}^{P}\sum_{t}\left(\mathbf{z}_{t-q}\otimes\mathbf{K}\mathbf{\Sigma}^{-1}\right)\left(\mathbf{B}_i\mathbf{X}_{t-i}^\top\otimes\mathbf{I}_M\right)\mathbf{vec}\left(\widehat{\mathbf{A}}_i-\mathbf{A}_i\right)\\
+&\sum_{i=1}^{P}\sum_{t}\left(\mathbf{z}_{t-q}\otimes\mathbf{K}\mathbf{\Sigma}^{-1}\right)\left(\mathbf{I}_N\otimes\mathbf{A}_i\mathbf{X}_{t-i}\right)\mathbf{vec}\left(\widehat{\mathbf{B}}_i^\top-\mathbf{B}_i^\top\right)\\
+&\sum_{j=1}^{Q}\sum_{t}\left(\mathbf{z}_{t-q}\mathbf{z}_{t-j}^\top\otimes\mathbf{K}\mathbf{\Sigma}^{-1}\mathbf{K}\right)\mathbf{vec}\left(\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)=\sum_{t}\left(\mathbf{z}_{t-q}\otimes\mathbf{K}\mathbf{\Sigma}^{-1}\right)\mathbf{vec}\left(\mathbf{E}_t\right)+o_P(\sqrt{T})
\end{aligned}$$

where we use the assumption that $\lambda = o(\sqrt{T})$ to include the penalty term in $o_P(\sqrt{T})$. Now, we define $\mathbf{U}_{it} = \mathbf{I}_N \otimes \mathbf{A}_i\mathbf{X}_{t-i}$, $\mathbf{V}_{it} = \mathbf{B}_i\mathbf{X}_{t-i}^\top \otimes \mathbf{I}_M$, and $\mathbf{Y}_{jt} = \mathbf{z}_{t-j}^\top \otimes \mathbf{K}$, and correspondingly $\mathbf{U}_t = [\mathbf{U}_{1t}, \ldots, \mathbf{U}_{Pt}]$, $\mathbf{V}_t = [\mathbf{V}_{1t}, \ldots, \mathbf{V}_{Pt}]$ and $\mathbf{Y}_t = [\mathbf{Y}_{1t}, \ldots, \mathbf{Y}_{Qt}]$. Then the above estimating equations can be grouped together as:

$$\sum_t \begin{bmatrix}\mathbf{V}_t^\top\\ \mathbf{U}_t^\top\\ \mathbf{Y}_t^\top\end{bmatrix}\mathbf{\Sigma}^{-1}[\mathbf{V}_t;\mathbf{U}_t;\mathbf{Y}_t]\begin{bmatrix}\mathbf{vec}(\widehat{\mathcal{A}}-\mathcal{A})\\ \mathbf{vec}(\widehat{\mathcal{B}}-\mathcal{B})\\ \mathbf{vec}(\widehat{\mathcal{R}}-\mathcal{R})\end{bmatrix} = \sum_t \begin{bmatrix}\mathbf{V}_t^\top\\ \mathbf{U}_t^\top\\ \mathbf{Y}_t^\top\end{bmatrix}\mathbf{\Sigma}^{-1}\mathbf{vec}(\mathbf{E}_t) + o_P(\sqrt{T}) \qquad (44)$$

By defining $\mathbf{W}_t = [\mathbf{V}_t; \mathbf{U}_t; \mathbf{Y}_t]$, and using the Martingale central limit theorem [Hall and Heyde, 2014], as $T \to \infty$, (44) can be converted to:

$$\mathbb{E}\left[\mathbf{W}_t^\top \mathbf{\Sigma}^{-1} \mathbf{W}_t\right] \begin{bmatrix} \mathbf{vec}(\widehat{\mathcal{A}} - \mathcal{A}) \\ \mathbf{vec}(\widehat{\mathcal{B}} - \mathcal{B}) \\ \mathbf{vec}(\widehat{\mathcal{R}} - \mathcal{R}) \end{bmatrix} \xrightarrow{d.} \mathcal{N}\left(\mathbf{0}, \mathbb{E}\left[\mathbf{W}_t^\top \mathbf{\Sigma}^{-1} \mathbf{W}_t\right]\right) \tag{45}$$

But $\mathbb{E}\left[\mathbf{W}_t^\top \mathbf{\Sigma}^{-1} \mathbf{W}_t\right]$ is not full-ranked, since $\mathbb{E}\left[\mathbf{W}_t^\top \mathbf{\Sigma}^{-1} \mathbf{W}_t\right] [\mathbf{vec}(\mathcal{A})^\top, -\mathbf{vec}(\mathcal{B})^\top, \mathbf{0}^\top]^\top = \mathbf{0}$. Given the identifiability constraint that $\|\mathbf{A}_p\|_F = 1, \forall p$, we have $\mathbf{vec}(\mathcal{A})^\top \mathbf{vec}(\widehat{\mathcal{A}} - \mathcal{A}) = o_P(T^{-1/2})$, thus we can replace the matrix $\mathbb{E}\left[\mathbf{W}_t^\top \mathbf{\Sigma}^{-1} \mathbf{W}_t\right]$ on the left hand side of (45) by $\mathbf{H} = \mathbb{E}\left[\mathbf{W}_t^\top \mathbf{\Sigma}^{-1} \mathbf{W}_t\right] + \boldsymbol{\eta}\boldsymbol{\eta}^\top$, where $\boldsymbol{\eta} = [\mathbf{vec}(\mathcal{A})^\top, \mathbf{0}^\top]^\top$ without changing the asymptotic normal distribution on the right hand side of (45) and thereby completes the proof of the asymptotic distribution of the model estimators. □

E Additional Details & Results of the Simulation

In this part of the supplemental materials, we provide additional technical details about the spatial kernel used for generating the functional parameters in Section 5.1. Then we provide additional results on the test set prediction performances under different lag choices following the experiment in Section 5.3.

E.1 Lebedev Kernel & Spherical Harmonics Basis

To motivate the real data applications, we consider a matrix time series distributed on a 2D spatial grid with the two dimensions being latitude and longitude of points on a sphere $\mathbb{S}^2$ in a three dimensional space. 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 corresponding Euclidean coordinates $(x, y, z) = (\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta))$.

We generate the functional parameters g_1, g_2, g_3 from an RKHS endowed with the Lebedev kernel:

$$K_\eta(s_1, s_2) = \left(\frac{1}{4\pi} + \frac{\eta}{12\pi}\right) - \frac{\eta}{8\pi}\sqrt{\frac{1 - \langle s_1, s_2\rangle}{2}}, \quad s_1, s_2 \in \mathbb{S}^2, \tag{46}$$

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 spherical harmonics as its eigenfunction:

$$K_\eta(s_1, s_2) = \frac{1}{4\pi} + \sum_{l=1}^{\infty} \frac{\eta}{(4l^2 - 1)(2l + 3)} \sum_{m=-l}^{l} Y_l^m(s_1) Y_l^m(s_2), \tag{47}$$

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

$$Y_l^m(s) = Y_l^m(\theta, \phi) = \begin{cases} \sqrt{2} N_{lm} P_l^m(\cos(\theta)) \cos(m\phi) & \text{if } m > 0 \\ N_{l0} P_l^0(\cos(\theta)) & \text{if } m = 0 \\ \sqrt{2} N_{l|m|} P_l^{|m|}(\cos(\theta)) \sin(|m|\phi) & \text{if } m < 0 \end{cases}, \tag{48}$$

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.

E.2 Test Prediction Comparison under Different (P, Q)

In Table 3, we conduct one-sided paired T-tests between the test prediction RMSE of the model with $(P, Q) = (2, 2)$ and the model with an arbitrary choice of (P, Q), and list the p-value of these tests under different training sample size.

As the sample size grows, the true model, i.e. MARAC$(2,2)$, has superior test prediction performance against those under-specified models with lower P or Q. Models with higher P or Q can have equally good prediction performance when sample size is large.

(P,Q)	$T=1000$	$T=2000$	$T=4000$	$T=8000$
$(1,1)$	5.24×10^{-54}	2.08×10^{-121}	2.88×10^{-100}	6.55×10^{-121}
$(1,2)$	1.45×10^{-50}	3.20×10^{-133}	1.90×10^{-100}	7.63×10^{-129}
$(1,3)$	2.85×10^{-50}	5.72×10^{-134}	4.20×10^{-100}	3.25×10^{-129}
$(1,4)$	2.49×10^{-50}	8.08×10^{-135}	4.90×10^{-100}	2.04×10^{-129}
$(2,1)$	8.72×10^{-3}	4.59×10^{-36}	1.96×10^{-19}	3.52×10^{-29}
$(2,2)$	–	–	–	–
$(2,3)$	**0.674**	8.34×10^{-3}	**0.243**	**0.075**
$(2,4)$	**0.483**	5.16×10^{-6}	**0.264**	**0.233**
$(3,1)$	1.63×10^{-3}	4.99×10^{-38}	8.29×10^{-21}	2.50×10^{-41}
$(3,2)$	**0.753**	4.54×10^{-12}	**0.700**	**0.297**
$(3,3)$	**0.497**	1.63×10^{-8}	**0.351**	0.013
$(3,4)$	**0.183**	1.11×10^{-7}	0.027	1.61×10^{-3}
$(4,1)$	1.46×10^{-6}	7.11×10^{-39}	9.00×10^{-27}	1.97×10^{-38}
$(4,2)$	**0.270**	5.93×10^{-21}	**0.051**	0.028
$(4,3)$	**0.101**	6.59×10^{-26}	0.023	1.76×10^{-6}
$(4,4)$	5.82×10^{-3}	2.71×10^{-6}	1.06×10^{-5}	1.32×10^{-6}

Table 3: *p-values of two-sample, one-sided, paired T-tests for test set RMSE under the correct $(P,Q)=(2,2)$ against other choices. The null hypothesis is that the test RMSE for both models have the same mean, the alternative is that the correct specification has a lower average RMSE. λ is chosen for each model based on a separate validation set of size $T_{val}\in\{500,1000,2000,4000\}$. The experiment is based on 100 repeated runs. The case where we* ***cannot*** *reject the null at 95% level is in boldface.*

F Additional Results on TEC Forecast

In Figure 7, we visualize a subset of the key model parameters for the MARAC$(2, 1)$ model trained from the stormy periods data. The fitted $\widehat{\mathbf{A}}_1, \widehat{\mathbf{B}}_1$ shows strong local auto-regressive effects, as the non-zero parameters concentrate around the main diagonal. The covariance components $\widehat{\boldsymbol{\Sigma}}_r$ revealed that the more noisy regions are near the equator in both north and south hemispheres (highlighted in red boxes), which coincides with the domain knowledge on the bands of latitudes with high TEC uncertainties. The $\widehat{\boldsymbol{\Sigma}}_c$ indicates that the noisy columns are those with local time in between 9 UT and 15 UT (highlighted in red boxes). The functional parameter estimates for IMF Bz and SYM-H, i.e. $\widehat{g}_{\text{IMF Bz}}$ and $\widehat{g}_{\text{SYM-H}}$, highlight the regions whose ΔTEC can be positively and negatively affected by both auxiliary covariates, which can be helpful for understanding the dynamics of the global TEC driven by these auxiliary scalar quantities. We leave the scientific discussion for future interdisciplinary research and simply demonstrate the output of the fitted model here for completeness.

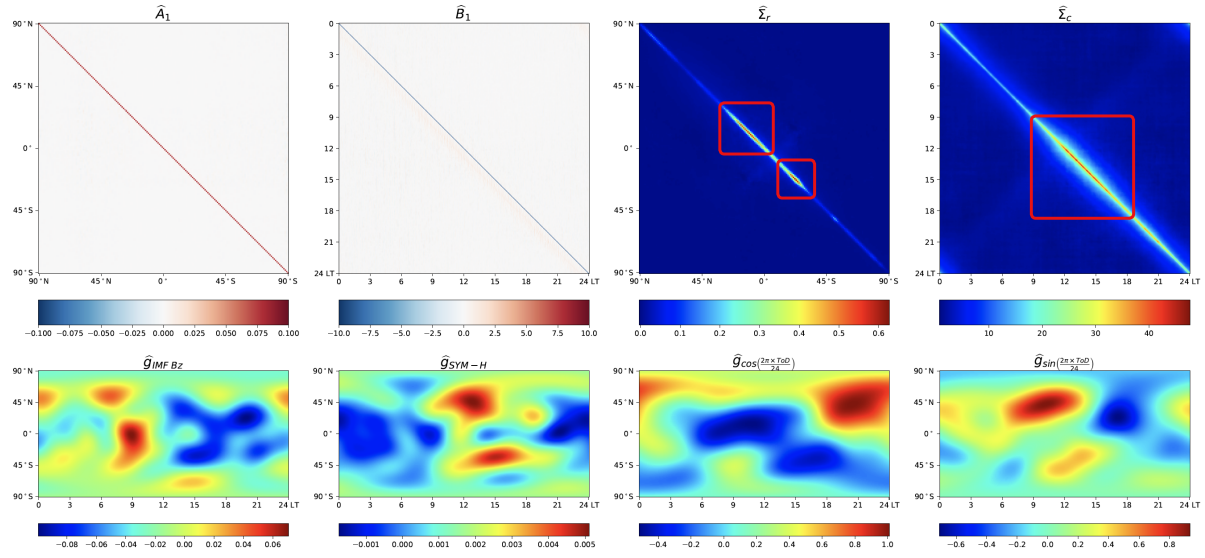

Figure 7: Selected MARAC$(2, 1)$ model parameters trained from the data of stormy periods.