UNIVERSITY of CALIFORNIA
Santa Barbara

Modeling Dependence in Large and Complex Data Sets

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy

in

Statistics and Applied Probability

by

Chao Zhang

Committee in charge:

Professor Alexander Petersen, Chair

Professor Wendy Meiring

Professor Tomoyuki Ichiba

June 2022

The dissertation of Chao Zhang is approved:

Professor Wendy Meiring

Professor Tomoyuki Ichiba

Professor Alexander Petersen, Chair

May 2022

To my family.

Acknowledgements

This dissertation and the graduate career which it represents have been influenced in a number of ways by a number of people and institutions, and I hope I am forgiven if I neglect to thank any of them here.

First and foremost, I extend my profoundest gratitude to my Ph.D. advisor Prof. Alexander Petersen for guiding me through my graduate career with his generous support, immense knowledge, and deep enthusiasm for research that has greatly shaped my passion towards research work.

I would also like to thank the rest of my dissertation committee: Prof. Wendy Meiring and Prof. Tomoyuki Ichiba, for their contribution to this dissertation.

A special thanks to Prof. Harvey Stein for the inspiration to expand my research domain and the guidance throughout.

To my fellow cohorts, friends, faculty, and staff members in the department of Statistics and Applied Probability at the University of California, Santa Barbara, who made this 5-year journey an unforgettable experience.

Finally, to my loving wife, Dr. Anqi Hu, and my family for their unwavering support.

Curriculum Vitæ

Chao Zhang

Education

2022	Ph.D., Statistics and Applied Probability, University of California, Santa Barbara, CA
2017	M.S., Mathematics, Ohio University, Athens, OH
2017	Master of Financial Economics, Ohio University, Athens, OH
2014	L.L.B., International Economic Law, East China University of Political Science and Law, Shanghai, China

Publications/In Preparation

Zhang, C., Kokoszka, P., & Petersen, A. (2022). Wasserstein autoregressive models for density time series. Journal of Time Series Analysis, 43(1), 30-52.

Petersen, A., **Zhang, C.**, & Kokoszka, P. (2022). Modeling probability density functions as data objects. Econometrics and Statistics, 21, 159-178.

Zhang, C., Tran, C., Achard, S., Meiring, W. and Petersen, A. Quantifying brain functional connectivity with noisy loxel level signals. *In preparation.*

Petersen, A. and **Zhang, C.** Concentration inequalties for estimation of mean and covariance functions with high-dimensional functional data. *In preparation.*

Abstract

Modeling Dependence in Large and Complex Data Sets

by

Chao Zhang

Classical statistical theory mostly focuses on independent samples that reside in finite dimensional vector spaces. While such methods are often appropriate and yield fruitful results, practical data analyses often go beyond the scope of these classical settings. In particular, with technological advancements, the computing power to record large volumes of data points at a high frequency is becoming more accessible than ever before. The large volume of data sets makes it possible to produce metadata on sample points—such as distributions, networks, or shapes, to name a few, and the high frequency of data records enables one to model data dependency structures at a fine temporal and/or spatial resolution that would not have been possible with sparsely recorded data. In the age of big data, the study of data atoms which constitute complex data objects and the statistical modeling of high resolution signals endowed with rich dependency structures are hitting their stride.

In this dissertation, we consider two specific instances of such big data. One is time dependent distributional data represented by the corresponding probability density functions. Indeed, data consisting of time-indexed distributions of cross-sectional or intraday returns have been extensively studied in finance, and provide one example in which the data atoms consist of serially dependent probability distributions. Motivated by such data, we propose an autoregressive model for density time series by exploiting the tangent space structure on the space of distributions that is induced by the Wasserstein metric. The densities themselves are not assumed to have any specific parametric form, leading to flexible forecasting of future unobserved densities. The main estimation targets in the order-p Wasserstein autoregressive model are Wasserstein autocorrelations and the vector-valued autoregressive parameter.

We propose suitable estimators and establish their asymptotic normality, which is verified in a simulation study. The new order-p Wasserstein autoregressive model leads to a prediction algorithm, which includes a data driven order selection procedure. Its performance is compared to existing prediction procedures via application to four financial return data sets, where a variety of metrics are used to quantify forecasting accuracy. For most metrics, the proposed model outperforms existing methods in two of the data sets, while the best empirical performance in the other two data sets is attained by existing methods based on functional transformations of the densities.

The second instance is the brain functional magnetic resonance imaging (fMRI) signals that are contaminated by spatiotemporal noise at the voxel level. Such data feature a rich spatiotemporal dependency structure due to a fine acquisition resolution. In neuroscience studies, resting state brain functional connectivity quantifies the similarity between pairs of brain regions, each of which consists of voxels at which dynamic signals are acquired via neuroimaging techniques, for example, the blood-oxygen-level-dependent (BOLD) signals that quantify an fMRI scan. Pearson correlation and similar metrics have been adopted to estimate inter-regional connectivity, often through averaging of signals within regions. However, dependencies between signals within each region and the presence of noise contaminate such inter-regional correlation estimates. We propose a mixed-effects model with a simple spatiotemporal covariance structure that explicitly isolates the different sources of variability in the observed BOLD signals, including correlated regional signals, local spatiotemporal noise, and measurement error. Methods for tackling the computational challenges associated with restricted maximum likelihood estimation will be discussed. Large sample properties are established by posing mild and practically verifiable sufficient conditions. Simulation results demonstrate that the parameters of the proposed model can be accurately estimated and is superior to the Pearson correlation of averages in the presence of spatiotemporal noise. The model was also implemented on data collected from a dead rat and an anesthetized live rat. Brain networks were constructed from estimated model parameters. Large scale parallel computing and GPU acceleration were implemented to speed up connectivity estimation.

Contents

Chapter 1

Introduction

This dissertation is a collection of individual projects that deal with large and complex data sets. Two specific classes of such data sets are studied—1) time-dependent distributional data represented in the form of probability density functions and 2) individual brain functional magnetic resonance imaging (fMRI) signals that are contaminated by spatiotemporal noises at the voxel level. Modeling time-dependent distributional data represented by probability density functions constitutes a rapidly growing subdomain of so-called *next-generation* functional data analysis (FDA), which commonly features dependent data atoms and/or complex geometry structures in the space where data atoms reside (Wang et al., 2016). While the second data class, namely spatially indexed dynamic signals, can be treated as spatial functional data (Delicado et al., 2010; Hörmann and Kokoszka, 2011), we will adopt the more classical linear mixed-effects (LME) models as our goal is to estimate brain connectivity at the individual level with no independent replicates.

Classical statistical inference treats probability density functions as fixed, usually parametric representations of the underlying mechanisms that generate observations. Estimating the probability density functions or the corresponding distributions is among the ultimate goals of inference. However, with technological advancements, nowadays large volumes of data sets are recorded at increasingly high frequencies and it is increasingly the case that observations are associated with their own probability distributions (Petersen et al., 2022). In order to model the heterogeneity among these probability distributions, one can adopt the view that each probability distribution is a data atom. In particular, we will focus on modeling time-dependent probability distributions in the form of probability density functions as data objects, yielding the notion of a *density time series*. The study of density time series can be categorized as next-generation FDA due to the nonlinear space in which the densities reside and the dependence across the time index of the data atoms. These features render the conventional FDA methods and results, such as functional principal component analysis, functional regression, functional central limit theorem, to name a few, inapplicable or at least inappropriate as they assume linear structure and independently, identically distributed (i.i.d.) samples. Readers are referred to the papers and monographs of Gasser et al. (1984); Rice and Silverman (1991); Ramsay and Silverman (2005); Hsing and Eubank (2015) for a comprehensive treatment of such conventional, or *first-generation* FDA topics.

In addition to the dependency across the time index, which is commonly investigated in time series analysis, a growing body of studies are motivated by incorporating data atoms' geographical marks as an integral part of statistical inquires and is commonly known as spatial or geostatistics. In this dissertation, we also work on the blood-oxygen-level-dependent (BOLD) fMRI signals collected at the voxel level which clearly has spatial attributes. Our goal of this study is to quantify brain connectivity between brain regions at the subject level. A large volume of literature has been devoted to brain connectivity studies aiming to shine a light on the evolution of pathologies such as neurodegenerative diseases or consciousness disorders. However, one of the common challenges in these studies is to address the noisy nature of the BOLD signals. In practice, BOLD signals are collected at voxels over a period of measurement time with fixed frequency. Such measurements incur spatiotemporal random perturbations and connectivity estimates can be subject to heavy biases (Achard et al., 2011; Chaimow et al., 2018). Due to the inherent spatial feature of BOLD sginals, the natural starting point of our inquiry is spatial statistics. The monograph by Cressie (1993) provides a detailed account of questions and frameworks in this area for vector-valued spatial data. A common practice is to assume *intrinsic stationarity* on the spatial process, then model the *semivariogram* with parametric covariance kernels. Predictions are commonly carried out by applying *kriging* (linear prediction) with variogram–based covariance estimates. With data processes' progression along the time domain, functional spatial models constitute a potential venue for modeling our data sets. Functional kriging can be developed in an analogous way as in the vector-valued setting to predict processes at unvisited locations.

However, there are several differences between the application scenarios of spatial FDA tools and our case of BOLD signals. First, functional spatial data usually assume a compact time domain, for instance, the Canadian weather data example in Ramsay and Silverman (2005). Most importantly, the asymptotic regimes of spatial functional data, either the *infill domain* or *increasing domain* sampling schemes, assume increasing spatial sampling locations, which is not an appropriate assumption for our BOLD signals. Therefore, spatial functional models are not readily applicable for our investigation on connectivity. However, we note that another principled approach to address data's dependency across space and time is spatiotemporal modeling (Christakos, 2000). This approach models the covariance structure across both space and time to properly represent the spatial and temporal features of the signals. The bright side of this method is that it borrows strength from neighboring signals to reveal dependence structures when data are scarce or irregular, with the downside of being computationally expensive when the number of data points increases. As our goal is to use noisy BOLD signals to quantify brain connectivity for individuals without independent replicates, we will adopt the LME models, which is commonly used in spatiotemporal modeling, to explicitly model the spatiotemporal features of our data and extract the connectivity estimates.

While both of the projects included in this dissertation investigate large dependent and complex data sets, their emphasises are in rather different research fields. The first project focuses on modeling time dependent distributional data as complex data objects, which is considered as next-generation FDA; the second project emphasizes spatiotemporal modeling of large and noisy neuroimaging data sets in order to quantify brain connectivity. The volume of research work devoted to both of these specific fields has enriched statistical literature at a rapid rate in the past decades and is still seeing fast-paced growth, hence it is not the intent of this chapter to catalog a comprehensive coverage of topics in either of these areas. Instead, its aim is to provide the motivations for the projects the author has worked on and their relevant backgrounds information in a self-contained manner.

For the remainder of this chapter, an overview of general settings and notations of functional data analysis will be presented in Section 1.1. This overview provides context for both of the projects included in this dissertation. The density time series is built upon the conventional FDA settings and the BOLD signals can be considered as observations of spatial functional data on a discrete grid, even though we adopt the spatiotemporal modeling approach for our studies of the signals. In Section 1.2, we proceed with a brief introduction to dependent functional data, in particular, conventional functional time series in linear space as it constitutes an important foundation to understand the motivation for density time series modeling that is thoroughly treated in Chapter 2. A quick discussion of the linear mixed-effects models is presented in Section 1.3 in order to provide background information for the methodology that we adopt to develop spatiotemporal models for BOLD signals in Chapter 3.

1.1 Overview

At the early stage of the development of functional data analysis, sometimes referred to as the first-generation functional data analysis, the main focus is on random samples consisting of independent real-valued functions on a compact interval on the real line, that is, $X_1(t), ...X_n(t)$, $t \in I \subset \mathbb{R}$, in particular, $I = [0, T]$. Early analysis of such data can be found in Gasser et al. (1984), where nonparametric smoothing, which is prevalent in FDA, was applied to biomedical data; or Rice and Silverman (1991), who modeled and estimated the mean function and covariance surface of the gaits data of a group of 5-year-old children, etc. Indeed, a common setting where FDA is carried out is Hilbertian structure. The random sample of independent real-valued functions $X_1(t), X_2(t), \ldots, X_n(t)$, $t \in I$ is regarded as realizations of an $L^2(I)$ process $X(t)$, that is, $\mathbb{E}\left[\int_I X^2(t)\mathrm{d}t\right] < \infty$. Such linear structure enables natural and heuristic notions of mean and covariance, whose estimators and their properties are of fundamental importance in FDA literature. For example, see Fan and Gijbels (1992); Yao et al. (2005); Hall et al. (2006); Zhang and Wang (2016); Fan and Gijbels (2018). Monographs by Ramsay and Silverman (2005); Hsing and Eubank (2015) provide thorough details from both theoretical and applied perspectives into the early developments of FDA.

An alternative term for the aforementioned type of functional data is "curve data". From the theoretical perspective, such data type is considered as the sample path of a stochastic process $X(t, \omega) : \Omega \to \mathbb{R}$, $t \in I$, where $(\Omega, \mathcal{A}, P)$ is the underlying probability space. Classical probability theory consctructs suitable filtrations to allow the notion of measurability globally, functional data usually adopts a local perspective where $X(t, \omega)$, $t \in I$ are treated as elements of $\mathbb{R}^I$. Therefore, the measurability is characterized in the sense that $\{\omega : X(\cdot, \omega) \in \mathcal{B}\} \in \mathcal{A}$, where $\mathcal{B}$ is a σ-algebra in a subspace $E \subset R^I$. In order for the measurability to hold, a careful choice of the subspace E and the norm on the space is usually required as $\mathbb{R}^I$ contains too many elements. A detailed treatment can be found in Bosq (2000); Wellner et al. (2013); Billingsley (2013). With this understanding, we reduce the notation from $X(t, \omega)$ to $X(t)$ hereafter.

1.2 Dependent Functional Data

Recent developments aim to tackle more complexity built on top of the first-generation functional data analysis, in particular, dependency in the data structure. A seminal work that models autoregressive dependent structure for functional data is presented by Bosq (2000). More recent monographs (Horváth and Kokoszka, 2012; Kokoszka and Reimherr, 2017a) provide thorough accounts of topics covering basics of functional data analysis, which is usually included in the discussion of first-generation functional data analysis, as well as estimation and modeling problems for dependent functional data. In this section, we will briefly discuss several dependency regimes that are relevant to author's projects.

1.2.1 Functional Time Series

In the case where a collection of functional samples $X_1(t), \ldots, X_n(t)$ are indexed by time, the i.i.d. assumption is no longer viable. In the study of scalar or real vector valued time series, the most general and popular dependence assumptions are various mixing conditions, see Bradley (2005); Doukhan (2012). Another widely adopted venue of dependency modeling is by moments based methods, such as modeling autocovariance functions. While taking different perspective, these two methods are indeed closely related to each other. These two principled ways of dependency quantification has led to fruitful developments on time series models and methodologies (Brockwell and Davis, 1991).

In functional time series analysis, one potential approach of temporal dependency modeling is to model the data dynamics directly, which incorporates the dependency mechanism. The alternative way, which is widely accepted in the statistical field, is to model the dependency structures directly. This approach is analogous to the study of vector-valued time series. Hence, we will adopt the same order of development, from general notions of dependency to specific models, to introduce the dependence modeling techniques that are relavant to author's project.

Notions of Dependency

Among all the relaxation of the independence assumptions, m-dependence has one of the simplest forms and is still flexible enough to model or approximate more convoluted dependent cases. Suppose $\{X_n(t), n \in \mathbb{Z}\}$ are measurable maps from probability space $(\Omega, \mathcal{A}, P)$ to a space $\mathbb{M}$. Let $\mathcal{A}_k^{-\infty}$ and $\mathcal{A}_k^{\infty}$ be the σ-algebras generated up to time k and from time k, respectively. Then the sequence $\{X_n(t), n \in \mathbb{Z}\}$ is said to be m-dependent if for any k, $\mathcal{A}_k^{-\infty}$ $\mathcal{A}_{k+m}^{\infty}$ are independent. One can immediately identify that this in indeed a special case of α-mixing. The idea of the notion of m-dependence is that even though most time series are not m-dependent, they can be approximated by a similar m-dependent process in some way. When the rate of approximation is fast enough, properties that are easy to establish for m-dependent processes usually hold on their approximation targets as well. See for example, Berkes and Horváth (2001); Berkes et al. (2006); Aue et al. (2009).

To formalize the above idea of approximation, let $\{X_n(t), n \in \mathbb{Z}\}$ be a sequence in $L^p[0,1]$, that is, $\mathbb{E}[\|X_n\|^p]^{1/p} < \infty$. X_n is called L^p-m-approximable if each X_n admits the representation

$$X_n = f(\epsilon_n, \epsilon_{n-1}, \dots), \tag{1.1}$$

where ϵ_i are i.i.d. measurable maps from $(\Omega, \mathcal{A}, P)$ to $\mathbb{M}$, then let $\{\epsilon_i'\}$ be independent copies of $\{\epsilon_i\}$ and

$$X_n^{(m)} = f(\epsilon_n, \epsilon_{n-1}, \dots, \epsilon_{n-m+1}, \epsilon_{n-m+1}', \epsilon_{n-m-1}', \dots), \tag{1.2}$$

we have

$$\sum_{m=1}^{\infty} \mathbb{E}\left[\|X_n - X_n^{(m)}\|^p\right]^{1/p} < \infty. \tag{1.3}$$

One can immediately see from the definition (1.1) and (1.2) that the sequence $\{X_n^{(m)}, n \in \mathbb{Z}\}$ is strictly stationary and m-dependent, and $X_n^{(m)}$ is equal in distribution to X_n for each m and n. With (1.3), the L^p-m-approximability provides a powerful tool in the study of various properties of functional times series, for example, functional central limit theorem. It also leads to specific models for functional time series, one of which is the functional autoregressive process of order 1 (FAR(1)).

1.2.2 Functional Autoregressive Process of Order 1

Let $\mathcal{L} = \mathcal{L}(L^p[0,1], L^p[0,1])$, the space of bounded linear operators between $L^p[0,1]$ and itself. Suppose $\Phi \in \mathcal{L}$ with operator norm $\|\Phi\|_{\mathcal{L}} < 1$. Denoted by ϵ_n, the i.i.d. zero-mean elements of $L^2[0,1]$. There exists a unique stationary sequence of $L^2[0,1]$ elements X_n such that

$$X_n(t) = (\Phi X_{n-1})(t) + \epsilon_n(t). \tag{1.4}$$

For details on existence and uniqueness, see Bosq (2000); Horváth and Kokoszka (2012); Kokoszka and Reimherr (2017a). It is straight forward to show that the FAR(1) model admits the representation

$$X_n = \sum_{j=1}^{\infty} \Phi^j \epsilon_{n-j}, \tag{1.5}$$

where Φ^j is applying the operator Φ for j times. Following the idea of (1.2), set

$$X_n^{(m)} = \sum_{j=1}^{m-1} \Phi^j \epsilon_{n-j} + \sum_{j=m}^{\infty} \Phi^j \epsilon'_{n-j}.$$

It is easy to verify that

$$\mathbb{E}\left[\left\|X_n - X_n^{(m)}\right\|^p\right]^{1/p} \leq 2 \sum_{j=m}^{\infty} \|\Phi\|_{\mathcal{L}}^j \, \mathbb{E}\left[\|\epsilon_o\|^p\right]^{1/p}.$$

Therefore, (1.3) is satisfied with $p \geq 2$ provided $\mathbb{E}\|\epsilon_o\|^{p1/p} < \infty$. One can see the processes whose dependence structure satisfies the FAR(1) process are indeed L^p-m-approximable.

1.3 Linear Mixed-Effects Models

Recall that we resort to linear mixed effects models for spatiotemporal modeling as spatial functional data methods are not readily applicable to BOLD signals for our connectivity study, even though our data sets indeed qualify as spatial functional data. Formally, denote the samples observed at locations $\mathbf{s}$ as $\{X(\mathbf{s}), \mathbf{s} \in S \subset \mathbb{R}^d\}$, $d \geq 1$, where $X(\mathbf{s})$ is usually assume to be a random function in $L^2[0, T]$ for some $T > 0$. This dimension of spatial functional data is usually time, hence the notation $X(\mathbf{s}; t)$, the value of the spatial functional process at location $\mathbf{s}$ and time t. When observed over a discrete grid, one can obtain a matrix representation of the data. In the most general setting, one assumes

$$X(\mathbf{s}; t) = \mu(t) + \epsilon(\mathbf{s}; t),\ t \in [0, T] \tag{1.6}$$

where $\mu(t)$ is the deterministic mean signal that is independent of spatial locations $\mathbf{s}$ and the random field $\epsilon(\mathbf{s}; t)$ contains information of spatiotemporal structure of the data processes. Furthermore, the covariance operator defined by

$$C_{\mathbf{s},\mathbf{s}}x = \mathbb{E}\left[\langle X(\mathbf{s}) - \mu, x\rangle \left(X(\mathbf{s}) - \mu\right)\right],\ x \in L^2[0, T], \tag{1.7}$$

is assumed to be homogeneous among spatial locations $\mathbf{s}$ as well since otherwise the problem of mean and covariance estimation would be ill-posed (Hörmann and Kokoszka, 2011). With this understanding, let $\{\mathbf{s}_1, \ldots, \mathbf{s}_l\}$, $l = 1, \ldots, L$, and $\{t_1, \ldots, t_m\}$, $m = 1, \ldots, M$ be the observation grids in space and time. The discrete version of model (1.6) is

$$X_{ilm} = \mu_m + \epsilon_{ilm}, \tag{1.8}$$

where $X_{ilm} = X_i(\mathbf{s}_l; t_m)$, $\mu_m = \mu(t_m)$, $\epsilon_{ilm} = \epsilon_i(\mathbf{s}_l; t_m)$, and $i = 1, \ldots, n$ is the index for subject. To simplify the notation, we implicitly assumed that all objects are observed on the identical spatialtemporal grids in (1.8). However, such simplification is not necessary for the adoption of the LME models and can be easily relaxed.

From (1.8), it is clear why linear mixed-effects models can be readily applied to discretized spatial functional data. In general, let X_i be the n_i vector of observations of subject $i = 1, \ldots, N$, LME models take the form of

$$\boldsymbol{X}_i = \boldsymbol{Z}_i\boldsymbol{\beta} + \boldsymbol{U}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{1.9}$$

where $\boldsymbol{Z}_i$ and $\boldsymbol{\beta}_i$ are the $n_i \times p$ design matrix and p-vector of coefficients of the *fixed effects*, respectively. $\boldsymbol{U}_i$ and $\boldsymbol{b}_i$ are those of the *random effects* of dimension $n_i \times q$ and q. It is conventionally assumed that $\boldsymbol{b}_i$ follows $\mathcal{N}(0, \sigma^2\boldsymbol{D})$ for some covariance matrix $\boldsymbol{D}$ and σ^2 comes from $\epsilon_i \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_{n_i})$, where $\boldsymbol{I}_{n_i}$ is the $n_i \times n_i$ identity matrix. LME models provide a parsimonious way to model the mean and covariance structures simultaneously in the sense that they allow nonindependence among samples which usually arise across different hierarchies or groups. See Chapter 1 of Pinheiro and Bates (2006) for various examples where LME models arise.

1.3.1 Model Estimation

Due to the Gaussian structures on the random components in (1.8), likelihood methods are prevailing in LME model estimation. Conventional maximum likelihood (ML) gives estimators for fixed effects that are invariant to basis changes in data but suffers from biases in covariance estimates. These biases arise from the loss of degrees of freedom from the estimation of fixed effects. To mitigate the situation, Harville (1974) proposed *restricted maximum likelihood (ReML)* estimators for linear mixed-effects models. The main ideas is to fit a joint Gaussian likelihood on the error contrast other than the original data, such that the fixed effects parameters $\boldsymbol{\beta}$ will not appear in the likelihood function. While ReML yields unbiased estimators for covariance component, the fixed-effects estimates are no longer invariant to basis changes. However, Pinheiro and Bates (2006) suggested that it is still reasonable to utilize the ReML estimates of fixed effects.

The choice between ML and ReML depends on one's goal of parameter estimation. In our case, since the covariance structure is of primary interest, we will adopt ReML for model estimation. The detailed formulation of the ReML function will be deferred to Chapter 3 when we develop more details of our BOLD signals. Here we note that common techniques to find ReML estimators are Newton-Raphson methods, EM algorithm, and Fisher scoring (Harville, 1974; Lindstrom and Bates, 1988; Cressie and Lahiri, 1993). The class of Newton-Raphson methods is most popular for its fast convergence rate and easy evaluation. One of the most well-known approach is the l-bfgs optimizer (Liu and Nocedal, 1989), which is a quasi-Newton method that only updates the approximated Hessian matrix with rank two matrices in each iteration other than evaluating the exact Hessian matrix.

1.3.2 Model Inference

Large sample properties of model parameters are critical to uncertainty quantification. The classical asymptotic distributions of ML estimators from i.i.d. data are well established under certain regularity conditions. For more general settings, Sweeting (1980) established general asymptotic normality of maximum likelihood estimators which requires increasing, convergent, and smooth information. It is not surprising that ReML estimators share similar asymptotic properties as the difference between ML and ReML estimators is effectively a set of linear transformations, i.e., error contrasts. In particular, the asymptotic properties for ReML estimators under the spatial regression setting, which is closely related to our project, was investigated by Cressie and Lahiri (1993) where it was shown that the parameters of the covariance structure of LME models converge to a Gaussian vector under mild regularity conditions. While we will get more exposures to LME in Chapter 3 as we develop our spatiotemporal models, we refer readers to Jennrich and Schluchter (1986); Lindstrom and Bates (1988); Pinheiro and Bates (2006); Jiang and Nguyen (2007) for more detailed background information.

Chapter 2

Wasserstein Autoregressive Models for Density Time Series

2.1 Introduction

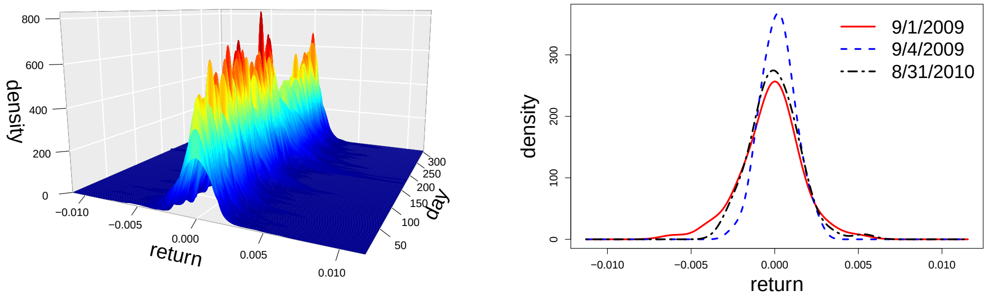
Samples of probability density functions or, more generally, probability distributions arise in a variety of settings. Examples include fertility and mortality data Mazzuco and Scarpa (2015), Shang and Haberman (2020), functional connectivity in the brain Petersen and Müller (2019), distributions of image features from head CT scans Salazar et al. (2019), and distributions of stock returns Harvey et al. (2016), Bekierman and Gribisch (2019), with the above recent references provided for illustration only. This chapter is concerned with modeling, estimation and forecasting of probability density functions which form a time series.

An early approach to the analysis of distributional data by Kneip and Utikal (2001) used cross-sectional averaging and functional principal component analysis (FPCA) applied directly to yearly income densities. In a more recent work, Yang et al. (2020) represented the sample of distributions by their quantile functions, and applied a linear function-on-scalar regression model with quantile functions as response variables. These two approaches are principled alternatives to naively apply methods of functional data analysis (FDA) to density-valued data. Since there are a variety of functional representations that provide unique characterizations of the distributions, including densities, quantile functions, and cumulative distribution functions, one faces the need to choose a representation prior to applying the (typically linear) methods of functional data analysis. Further complicating this dilemma is the fact that these standard functional representations do not constitute linear spaces due to inherent nonlinear constraints (e.g., monotonicity for quantile functions or positivity and mass constraints for densities), so that outputs from models with linear underlying structures are generally inadequate. For this reason, methodological developments for the analysis of distributional data have taken a geometric approach over the last decade. Rather than choosing a functional form under which to analyze the data, one chooses a metric on the space of distributions in order to develop coherent models. Examples of suitable metrics that have been used successfully in the modeling of distributional data include the Fisher-Rao metric Srivastava et al. (2007), an infinite-dimensional version of the Aitchison metric Egozcue et al. (2006); Hron et al. (2016), and the Wasserstein optimal transport metric Panaretos and

Zemel (2016); Petersen and Müller (2019); Bigot et al. (2017).

In many cases, the distributions in a sample are indexed by time, for example annual income, fertility and mortality data, or financial returns or insurance claims at various time resolutions. In this chapter, we will assume that all such distributions possess a density with respect to the Lebesgue measure, and will refer to this type of data as a density time series. A motivating example is shown in Figure 2.1, depicting the distribution of 5-minute intraday returns of the XLK fund, which tracks the technology and telecommunication sectors within the S&P 500 index. The data we plot in Figure 2.1a covers 305 trading days, each with 78 records of 5-minute intraday return. Figure 2.1b demonstrates an alternative look at this data set by plotting returns from three selected trading days. Kokoszka et al. (2019) considered various methods for forecasting density time series, most of which produced forecasts by first applying FPCA to the densities (or transformations of these), followed by fitting a multivariate time series models to the vectors of coefficients. Finally, the density forecasts were obtained by using the forecasts of the coefficients in the FPCA basis representation. Of these different methods, a modified version of the transformation of Petersen et al. (2016a) gave superior forecasts in the majority of cases, and was also based on a sound theoretical justification in terms of explicitly controlling for the density constraints.

The main contribution of this chapter is to develop a geometric approach to density time series modeling under the Wasserstein metric. It is well-known that this geometry is intimately connected with quantile functions, and thus provides a flexible framework



(a) 9/1/2009 - 11/17/2010

(b) 9/1/2009, 9/4/2009 and 8/31/2010

Figure 2.1: Densities of XLK, the Technology Select Sector SPDR Fund 5-minute intraday returns on selected dates.

for modeling samples of densities that tend to exhibit “horizontal” variability, which can be thought of as variability of the quantiles. Examples of such variability in densities are given in Figure 2.1b. We develop theoretical foundations of autoregressive modeling in the space of densities equipped with the Wasserstein metric, followed by methodology for estimation and forecasting, including order selection. Since the Wasserstein geometry is not linear, care needs to be taken to ensure the model components and their restrictions are appropriately specified. Autoregressive models have been the backbone of time series analysis for scalar and vector-valued data for many decades, see e.g. Lütkepohl (2006), among many other excellent textbooks. Autoregression has been extensively studied in the context of linear functional time series; most papers study or use order one autoregression, see Bosq (2000) and Horváth and Kokoszka (2012). This chapter thus merges two successful approaches: the Wasserstein geometry and time series autoregression.

In a very recent preprint, Chen et al. (2020) independently proposed a similar geometric approach to regression when distributions appear as both predictors and responses.

As an extension of this formulation, they also developed an autoregressive model of order one for distribution-valued time series. Our AR(1) model proposed in Section 2.3.1 can be viewed as a special case of the model in Chen et al. (2020). However, the generalization, theory and methodology we subsequently pursue move in a completely different direction, so the two projects have little overlap. Even though we were not aware of the work of Chen et al. (2020), we did include their model, which is termed the fully functional Wasserstein autoregressive model in this work, as a one of the competing methods in our empirical analyses in Section 2.5. We also note that our focus on densities with respect to the Lebesgue measure is motivated by practical considerations, as such densities occur in applications. In particular, we formulate numerical algorithms applicable to this common setting. From the theoretical angle, our results related to existence and convergence could be extended to general probability measures. Working with densities actually introduces nontrivial complications. For example, the objects we want to predict must be densities, not general probability measures.

The remainder of the chapter is organized as follows. In Section 2.2 we provide the requisite background on Wasserstein geometry and introduce relevant definitions related to density time series. Section 2.3 is devoted to the development of the Wasserstein AR(p) model, including its estimation and forecasting, both in terms of theory and algorithms. Proofs of theorems and lemmas are also presented in the same section. Finite sample properties of our estimator are explored in Section 2.4, while Section 2.5 compares our forecasting algorithm to those currently available. We conclude the chapter with a discussion in Section 2.6.

2.2 Preliminaries

A density time series is a sequence of random densities $\{f_t, t \in \mathbb{Z}\}$. In the spirit of functional data analysis, no parametric form for the densities will be assumed. Furthermore, the models will be developed under the setting in which the densities are completely observed, although in practical situations they will need to be estimated from raw data that they generate. For example, the densities in Figure 2.1 are kernel density estimates with a Gaussian kernel.

Density time series are a special case of functional time series, so it would be natural to adapt a functional autoregressive model (see e.g. Chapter 8 of Kokoszka and Reimherr (2017b)). However, such a direct approach is only suitable if one first transforms the densities into a linear space, although this approach too comes with disadvantages. The transformations of Petersen et al. (2016a) and Hron et al. (2016) require that all densities in the sample share the same support, an assumption that is often broken in real data sets. Although Kokoszka et al. (2019) modified the method of Petersen et al. (2016a) to remove this constraint, the associated transformation is not connected with any meaningful density metric, and can suffer from noticeable boundary effects if the observed densities decay to zero near the boundaries. Still, the transformation approach remains viable and will be compared to the Wasserstein models that we propose.

2.2.1 Wasserstein Geometry and Tangent Space

We begin with a brief discussion of the necessary components of the Wasserstein geometry. Consider the space of probability measures $\mathcal{W}_2 = \{\mu : \mu$ is a probability measure on $\mathbb{R}$ and $\int x^2 \mathrm{d}\mu(x) < \infty\}$. Denoted by $\mathcal{D}$ the subset of $\mathcal{W}_2$ consisting of measures with densities with respect to Lebesgue measure, so that one may think of $\mathcal{D}$ as a collection of densities. For $f, g \in \mathcal{D}$, consider the collection $\mathbb{K}_{f,g}$ of maps $K : \mathbb{R} \to \mathbb{R}$ that transport f to g, that is, if $K \in \mathbb{K}_{f,g}$ and U is a random variable that follows the distribution characterized by f, i.e. $U \sim f$, then $K(U) \sim g$. Intuitively, f and g are close if there exists a $K \in \mathbb{K}_{f,g}$ such that $K \approx \mathrm{id}$, where $\mathrm{id}(u) = u$ denotes the identity map. This is the motivation behind the Wasserstein distance

$$d_W(f,g) = \inf_{K \in \mathbb{K}_{f,g}} \left\{ \int_{\mathbb{R}} (K(u) - u)^2 f(u) \mathrm{d}u \right\}^{1/2}. \tag{2.1}$$

That d_W is a proper metric is well-established Villani (2003), and (2.1) is indeed only one of a large class of such metrics that can in fact be defined for measures on quite general spaces. In the particular setting of univariate distributions, a surprising property is that the infimum in (2.1) is attained by the so-called optimal transport map $K^* = G^{-1} \circ F$, where F and G are the cdfs of f and g, respectively. Note that any optimal transport map must be strictly increasing, so that, by the change of variable $s = F(u)$, this leads to an alternative definition of the Wasserstein metric

$$d_W(f,g) = \left\{ \int_{\mathbb{R}} (K^*(u) - u)^2 f(u) \mathrm{d}u \right\}^{1/2} = \left\{ \int_0^1 \left(G^{-1}(s) - F^{-1}(s) \right)^2 \mathrm{d}s \right\}^{1/2}. \tag{2.2}$$

For clarity, we will use u as the input for densities and cdfs, and s as the input for quantile functions. Interestingly, even for univariate probability measures in $\mathcal{W}_2$ that do not admit a density, the Wasserstein metric remains well-defined, and both optimal transport maps and corresponding distance can be expressed in terms of their quantile functions (which always exist), as above.

Another surprising characteristic of the Wasserstein metric is that, although $(\mathcal{W}_2, d_W)$ is not a linear space, its structure is strikingly similar to that of a Riemannian manifold Ambrosio et al. (2008). As mentioned previously, a key challenge in analyzing samples of probability density functions is that these reside in a convex space where linear methods fall short. However, due to the manifold-like structure, to each $\mu \in \mathcal{W}_2$ corresponds a tangent space $\mathcal{T}_\mu$ that *is* a complete linear subspace of $L^2(\mathbb{R}, \mathrm{d}\mu)$ (see Chapter 8 of Ambrosio et al. (2008)), opening the door for development of linear models for distributional data. According to (8.5.1) in Ambrosio et al. (2008), we define the tangent space for $\mu \in \mathcal{W}_2$ by

$$\mathcal{T}_\mu = \overline{\{\lambda(T - \mathrm{id}) : T \text{ is the optimal transport from } \mu \text{ to some } \nu \in \mathcal{W}_2, \lambda > 0\}}, \qquad (2.3)$$

where the closure is with respect to $L^2(\mathbb{R}, \mathrm{d}\mu)$. With a slight abuse of notation, when μ possesses a density f, we will denote this tangent space by $\mathcal{T}_f$. The definition in (2.3) of the tangent space can be motivated by the following fact. For $\mu, \nu \in \mathcal{W}_2$ and T the optimal transport from μ to ν, define the curve (known as McCann's interpolant) $\lambda \in [0, 1] \mapsto [\mathrm{id} + \lambda(T - \mathrm{id})]_\# \mu$, where $g_\# \mu(A) = \mu(g^{-1}(A))$ for $A \in \mathcal{B}(\mathbb{R})$ denotes the pushforward measure induced by a measurable function g. For different measures ν, these are geodesic curves connecting μ to ν in $\mathcal{W}_2$ Panaretos and Zemel (2020). Thus, the extension to values $\lambda > 0$ defines a tangent cone. That $\mathcal{T}_\mu$ is indeed a linear space is not obvious from the definition, but this property can indeed be established; see, for example, Chapter 2.3 of Panaretos and Zemel (2020).

We next describe two maps that bridge the tangent space and the space of densities. Let $f, g \in \mathcal{D}$ have cdfs F and G, respectively. The map Log_f: $\mathcal{D} \to \mathcal{T}_f$ defined by

$$\mathrm{Log}_f(g) = G^{-1} \circ F - \mathrm{id} \tag{2.4}$$

is called the logarithmic map at f, and effectively lifts the space $\mathcal{D}$ to the tangent space $\mathcal{T}_f$. Intuitively, $\mathrm{Log}_f(g)$ represents the discrepancy between the optimal transport map $G^{-1} \circ F$ and the identity. In fact, (2.2) shows that $d_W^2(f, g) = \int_{\mathbb{R}} [\mathrm{Log}_f(g)(u)]^2 f(u) \mathrm{d}u$, so that the logarithmic map takes the place of the ordinary functional difference $g - f$ that is commonly used in linear spaces. The second is the exponential map $\mathrm{Exp}_f : \mathcal{T}_f \to \mathcal{W}_2$. Let $V \in \mathcal{T}_f$, and define Exp_f by

$$\mathrm{Exp}_f(V) = (V + \mathrm{id})_{\#}\mu_f, \tag{2.5}$$

where μ_f is the measure with density f and

$$(V + \mathrm{id})_{\#}\mu_f(A) = \mu_f\left((V + \mathrm{id})^{-1}(A)\right), \quad A \in \mathcal{B}(\mathbb{R}),$$

where $\mathcal{B}(\mathbb{R})$ denotes the Borel sets. Observe that, for any $f, g \in \mathcal{D}$, $\mathrm{Exp}_f(\mathrm{Log}_f(g)) = g$, but $\mathrm{Log}_f(\mathrm{Exp}_f(V)) = V$ holds if and only if $V + \mathrm{id}$ is increasing.

Looking forward to building a Wasserstein autoregressive model, the logarithmic map will be used to lift the random densities into a linear tangent space, where the autoregressive model is imposed. An important point to keep in mind is that the image of $\mathcal{D}$ under Log_f is a convex cone, and thus a nonlinear subset of $\mathcal{T}_\mu \subset L^2(\mathbb{R}, f(u)\mathrm{d}u)$. We will deal with this technicality in the development of Wasserstein autoregressive models in Section 2.3. In particular, the forecasts produced by the model in the tangent space will not be constrained to lie in the image of the logarithmic map. This poses no practical problem since the forecasted densities are obtained through the exponential map, which is defined on the entirety of the tangent space.

2.2.2 Wasserstein Mean, Variance, and Covariance

Consider a random density f, which is a measurable map that assumes values in $\mathcal{D}$ almost surely. Assume $\mathbb{E}\left[d_W^2(f,g)\right] < \infty$ for some, and thus all, $g \in \mathcal{D}$. Petersen et al. (2020) demonstrated sufficient conditions for the Wassersetin mean density of f, written as

$$\mathbb{E}_\oplus[f] = f_\oplus = \operatorname*{argmin}_{g \in \mathcal{D}} \mathbb{E}\left[d_W^2(f,g)\right], \tag{2.6}$$

to exist, which represents the Fréchet mean in the metric space $\mathcal{D}$ equipped with the Wasserstein distance. We will thus assume that $f_\oplus$ exists and is unique, and write $F_\oplus$ and $Q_\oplus$ for the cdf and quantile functions, respectively, that correspond to $f_\oplus$. Letting $T = F^{-1} \circ F_\oplus$ be the random optimal transport map from $f_\oplus$ to f, the Wasserstein variance of f is

$$\mathrm{Var}_{\oplus}(f) = \mathbb{E}\left[d_W^2(f, f_{\oplus})\right] = \mathbb{E}\left[\int_{\mathbb{R}} (T(u) - u)^2 f_{\oplus}(u) \mathrm{d}u\right]. \tag{2.7}$$

Since $\mathbb{E}\left[d_W^2(f, g)\right] < \infty$ for all $g \in \mathcal{D}$ by assumption, existence of the Wasserstein mean $f_{\oplus}$ implies that the Wasserstein variance $\mathrm{Var}_{\oplus}(f)$ is finite.

Now, suppose f_1 and f_2 are two random densities, with Wasserstein means $f_{\oplus,1}$ and $f_{\oplus,2}$, respectively. Since we will consider an autoregressive model, it is necessary to develop a suitable notion of covariance within and between these random densities. The usual approach in functional data analysis would quantify this by the crosscovariance kernel of the centered processes $f_t - f_{\oplus,t}$, $t = 1, 2$. However, as mentioned previously, this differencing operation is not suitable for nonlinear spaces, and we thus replace it with the logarithmic map in (2.4). Let $T_t = F_t^{-1} \circ F_{\oplus,t}$ be the optimal transport map from the Wasserstein mean $f_{\oplus,t}$ to the random density f_t. To make clear the parallel between the ordinary functional covariance and the Wasserstein version we will define, recall that the logarithmic map replaces the usual notion of difference between two densities, so we introduce the alternative suggestive notation

$$f_t \ominus f_{\oplus,t} = \mathrm{Log}_{f_{\oplus,t}}(f_t) = T_t - \mathrm{id} \tag{2.8}$$

for the logarithmic map. Then the Wasserstein covariance kernel is defined by

$$\begin{aligned} \mathcal{C}_{t,t'}(u, v) &= \mathrm{Cov}\left[(f_t \ominus f_{\oplus,t})(u), (f_{t'} \ominus f_{\oplus,t'})(v)\right] \qquad (2.9) \\ &= \mathrm{Cov}\left[T_t(u) - u, T_{t'}(v) - v\right], \quad t, t' = 1, 2. \end{aligned}$$

Since $\int_{\mathbb{R}} \mathbb{E}\left(f_t \ominus f_{\oplus,t}(u)\right)^2 f_{\oplus,t}(u)\mathrm{d}u < \infty$, $\mathbb{E}\left(f_t \ominus f_{\oplus,t}(u)\right)^2 < \infty$ for almost all u in the support of $f_{\oplus,t}$. This means that the Wasserstein covariance kernels $\mathcal{C}_{t,t'}(u,v)$ are defined for almost all $(u,v) \in \mathrm{supp}(f_{\oplus,t}) \times \mathrm{supp}(f_{\oplus,t'})$. To further solidify the intuition behind this definition, observe that the Wasserstein variance in (2.7) can be rewritten as

$$\mathrm{Var}_{\oplus}(f_t) = \int_{\mathbb{R}} \mathcal{C}_{t,t}(u,u) f_{\oplus,t}(u)\mathrm{d}u,$$

echoing the notion of total variance typically used for functional data. This was the motivation used in Petersen and Müller (2019) in order to define a scalar measure of Wasserstein covariance between two random densities.

2.2.3 Stationarity of Density Time Series

Stationarity plays a fundamental role in time series analysis. It is a condition generally imposed on the random part of the process that remains after removing trends, periodicity, differencing or after other transformations. It is needed to develop estimation and prediction techniques. Here we develop notions of stationarity and strict stationarity for a time series of densities $\{f_t, t \in \mathbb{Z}\}$.

Definition 2.2.1. A density time series $\{f_t, t \in \mathbb{Z}\}$ is said to be (second-order) stationary if the following two conditions hold.

1. $\mathbb{E}_{\oplus}\left[f_t\right] = f_{\oplus}$ for all $t \in \mathbb{Z}$, so the f_t share a common Wasserstein mean. Denote $\mathrm{supp}(f_{\oplus})$ by $D_{\oplus}$.

2. $\mathrm{Var}_{\oplus}(f_t) < \infty$.

3. For any $t, h \in \mathbb{Z}$, and almost all $u, v \in D_{\oplus}$, $\mathcal{C}_{t,t+h}(u,v)$ does not depend on t.

As we take the approach that focuses on the geometry of the space of densities, the above notion of stationarity is defined by the Wasserstein mean and covariance kernel, which is not equivalent to those traditional stationarity definitions of functional time series. In particular, a conventional stationarity notion for a stochastic process is understood in the following sense, see e.g. Bosq (2000).

Definition 2.2.2. A sequence $\{V_t\}$ of elements of a separable Hilbert space is said to be stationary if the following conditions hold: (i) $\mathbb{E}\left[\|V_t\|^2\right] < \infty$, (ii) $\mathbb{E}\left[V_t\right]$ does not depend on t, and (iii) the autocovariance operators defined by $\mathcal{G}_{t,t+h}(x) = \mathbb{E}\left[\langle (V_t - \mu), x\rangle (V_{t+h} - \mu)\right]$ do not depend on t ($\mu = \mathbb{E}V_0$).

Observe that Definition 2.2.2 clearly does not apply to the density time series $\{f_t, t \in \mathbb{Z}\}$ as densities do not form a vector space. The fact alone that differences $f_t - \mathbb{E}\left[f_{\oplus}\right]$ are not well-defined in a nonlinear space renders Definition 2.2.2 unsuitable for density time series. However, upon taking $V_t = \mathrm{Log}_{f_{\oplus}}(f_t)$, Definition 2.2.1 implies Definition 2.2.2, with the separable Hilbert space in the latter being the tangent space $\mathcal{T}_{f_{\oplus}}$. As has been observed elsewhere (e.g., Panaretos and Zemel (2016); Petersen et al. (2016a)), the Wasserstein mean $f_{\oplus}$ (when it exists) is characterized by being the unique solution to $\mathbb{E}\left[\mathrm{Log}_{f_{\oplus}}(f_t)(u)\right] = 0$ for almost all u in the support of $f_{\oplus}$. Hence, condition (ii) is satisfied since $\mu = \mathbb{E}\left[V_0\right] = 0$, from which condition (i) follows as $\mathbb{E}\left[\|V_t\|^2\right] = \mathrm{Var}_{\oplus}(f_t) < \infty$. Lastly, condition (iii) holds since, for any element $x \in \mathcal{T}_{f_{\oplus}}$,

$$\mathcal{G}_{t,t+h}(x) = \mathbb{E}\left[\left(\int_{D_\oplus} V_t(u)x(u)f_\oplus(u)\mathrm{d}u\right) V_{t+h}\right] = \int_{D_\oplus} \mathcal{C}_{t,t+h}(\cdot,u)x(u)f_\oplus(u)\mathrm{d}u,$$

which is independent of t. Equivalently, if Q_t is the quantile function corresponding to f_t, Definition 2.2.1 implies that the optimal transport maps $T_t = Q_t \circ F_\oplus = X_t + \mathrm{id}$ form a stationary sequence in $\mathcal{T}_{f_\oplus}$ according to Definition 2.2.2 with $\mu = \mathrm{id}$.

Definition 2.2.3. A density time series $\{f_t, t \in \mathbb{Z}\}$ is said to be strictly stationary if the joint distributions on $\mathcal{D}^k$ of $(f_{t_1}, f_{t_2}, \ldots, f_{t_k})$ and $(f_{t_1+h}, f_{t_2+h}, \ldots, f_{t_k+h})$ are the same for any $k \in \mathbb{N}$ and choices $t_1, t_2, \ldots, t_k, h \in \mathbb{Z}$.

Note that, if the densities f_t share a common Wasserstein mean $f_\oplus$ and the joint distributions of $(T_{t_1}, T_{t_2}, \ldots, T_{t_k})$ and $(T_{t_1+h}, T_{t_2+h}, \ldots, T_{t_k+h})$ are the same for any $k \in \mathbb{N}$ and choices $t_1, t_2, \ldots, t_k, h \in \mathbb{Z}$, then $\{f_t, t \in \mathbb{Z}\}$ is strictly stationary according to Definition 2.2.3. Since the existence and uniqueness of the Wasserstein mean implies that the Wasserstein variance is finite, it also follows that $\{f_t, t \in \mathbb{Z}\}$ is stationary according to Definition 2.2.1, provided the Wasserstein mean exists and is unique.

2.3 Wasserstein Autoregression

The above notions of stationarity and strict stationarity in the tangent space facilitate the development of autoregressive models in $\mathcal{T}_{f_\oplus}$ by lifting the random densities via the logarithmic map. As observed previously, the image of $\mathcal{D}$ under this map is a convex cone in $\mathcal{T}_{f_\oplus}$, so it is not immediately possible to impose onto the tangent space standard structures used for functional time series, which rely on linearity of the function space (see e.g. Chapter 8 of Kokoszka and Reimherr (2017b) and references therein). To illustrate the challenges that must be overcome, we begin with a simple model involving a single scalar autoregressive parameter, and then consider extensions. For a stationary density time series $\{f_t, t \in \mathbb{Z}\}$, with Wasserstein mean cdf and quantile functions $F_\oplus$ and $Q_\oplus$, respectively, define

$$\gamma_h(u, v) := \mathrm{Cov}\left(f_t \ominus f_\oplus(u), f_{t+h} \ominus f_\oplus(v)\right). \tag{2.10}$$

2.3.1 Wasserstein AR Model of Order 1

From Definition 2.2.1, a useful path to pursue in developing an autoregressive model for density time series is to first establish a suitable primary model for a sequence $\{V_t\}$ on a tangent space $\mathcal{T}_{f_\oplus}$, for some $f_\oplus \in \mathcal{D}$. Recall that $\mathcal{T}_{f_\oplus}$ is a separable Hilbert space. The second step is to impose conditions on $\{V_t\}$ such that

a) the measures $\mu_t = \mathrm{Exp}_{f_\oplus}(V_t)$ possess densities f_t that form a stationary density time series with Wasserstein mean $f_\oplus$, and

b) the parameters in the primary model can still be estimated given observations of the f_t.

To this end, fix $f_\oplus \in \mathcal{D}$, where we assume that the support $D_\oplus$ of $f_\oplus$ is an interval, possibly unbounded. Let $\beta \in \mathbb{R}$ be the autoregressive parameter, and $\{\epsilon_t\}$ a sequence of independent and identically distributed stochastic processes (innovations) that reside in $\mathcal{T}_{f_\oplus}$ almost surely. We assume that the ϵ_t satisfy $\mathbb{E}\left[\epsilon_t(u)\right] = 0$ for all $u \in D_\oplus$ and define the innovation covariance kernel

$$C_\epsilon(u, v) = \mathrm{Cov}\left[\epsilon_t(u), \epsilon_t(v)\right], \quad u, v \in \mathbb{R}. \tag{2.11}$$

We say that a sequence $\{V_t\}$ follows an autoregressive model of order 1 if the random elements $V_t \in \mathcal{T}_{f_\oplus}$ satisfy the equation

$$V_t = \beta V_{t-1} + \epsilon_t, \quad t \in \mathbb{Z}. \tag{2.12}$$

As will be detailed in Theorem 2.3.1, (2.12) has a unique, suitably convergent, solution $V_t = \sum_{i=0}^{\infty} \beta^i \epsilon_{t-i}$ under the following conditions:

(A1) $|\beta| < 1$,

(A2) innovations are iid elements of $\mathcal{T}_{f_\oplus}$, with mean zero and $\int_{\mathbb{R}} C_\epsilon(u, u) f_\oplus(u) \mathrm{d}u < \infty$.

To ensure that requirements a) and b) above are met, we impose the following condition.

(A3) Almost surely, V_t is differentiable, and $V_t'(u) > -1$ for all $u \in D_\oplus$.

Denote the usual Hilbert norm on $L^2(\mathbb{R}, f_\oplus(u)\mathrm{d}u)$ by $\|\cdot\|$. We now state our first result associated with model (2.12), and its consequences for the density time series induced by the exponential map. Its proof, along with those of all other theoretical results, can be found in Section 2.3.3.

Theorem 2.3.1. *If (A1) and (A2) hold, then*

$$V_t = \sum_{i=0}^{\infty} \beta^i \epsilon_{t-i} \tag{2.13}$$

defines a unique, strictly stationary solution in $\mathcal{T}_{f_\oplus}$ *to model* (2.12)*. This solution converges strongly,*

$$\lim_{n\to\infty} \left\| V_t - \sum_{i=0}^{n} \beta^i \epsilon_{t-i} \right\| = 0 \text{ almost surely,} \tag{2.14}$$

and in mean square,

$$\lim_{n\to\infty} \mathbb{E} \left\| V_t - \sum_{i=0}^{n} \beta^i \epsilon_{t-i} \right\|^2 = 0. \tag{2.15}$$

If, in addition, (A3) holds, then the measures $\mu_t = \mathrm{Exp}_{f_\oplus}(V_t)$ *possess densities that form a strictly stationary sequence* $\{f_t, t \in \mathbb{Z}\}$ *with common Wasserstein mean* $f_\oplus$*, and* $V_t = T_t - \mathrm{id}$ *almost surely.*

In light of Theorem 2.3.1, we define the Wasserstein autoregressive model of order 1, or WAR(1) model, for a density time series $\{f_t, t \in \mathbb{Z}\}$ by

$$T_t - \mathrm{id} = \beta(T_{t-1} - \mathrm{id}) + \epsilon_t. \tag{2.16}$$

Under (A1)–(A3), we now know that a unique solution $f_t \ominus f_\oplus = T_t - \mathrm{id} = \sum_{i=0}^{\infty} \beta^i \epsilon_{t-i}$ exists such that $\{f_t, t \in \mathbb{Z}\}$ is strictly stationary according to Definition 2.2.3. Since they also share a common Wasserstein mean, the sequence is also stationary according to Definition 2.2.1.

In order for the results of Theorem 2.3.1 to not be vacuous, we will establish a set of innovation examples that satisfy (A2) and (A3). Given the structure of the tangent space in (2.3), consider innovations of the form $\epsilon_t(u) = \lambda_t(S_t(u) - u)$, where $\lambda_t > 0$ and S_t is an increasing map defined on $D_\oplus$ (and is thus an optimal transport map from $f_\oplus$ to some $\nu \in \mathcal{W}_2$). Both λ_t and S_t can be random. We now list specific examples for which (A2) and (A3) hold, where $|\beta| < 1$ throughout.

Example 2.3.1. *Let η_t be iid random variables with mean zero and finite variance. The WAR(1) model admits constant innovations $\epsilon_t(u) \equiv \eta_t$, which can be identified as elements in $\mathcal{T}_{f_\oplus}$ by setting $S_t(u) = \eta_t\lambda_t^{-1} + u$ for any $\lambda_t > 0$.*

Example 2.3.2. *Let η_t be as in Example 2.3.1, and δ_t be iid random variables with mean zero such that $|\delta_t| < 1 - |\beta|$. Linear innovations $\epsilon_t(u) = \eta_t + \delta_t u$ are admissible under the WAR(1) model. The tangent space representation of $\epsilon_t(u)$ can be recovered by setting $S_t(u) = (1 + \delta_t\lambda_t^{-1})u + \eta_t\lambda_t^{-1}$ for any $\lambda_t > |\delta_t|$.*

Example 2.3.3. *Let η_t and δ_t be as in Example 2.3.2, with the additional constraint that the δ_t be symmetric about 0. The WAR(1) model admits periodic innovations $\epsilon_t(u) = \eta_t + \sin(\delta_t u)$, which can be viewed as tangent space elements by writing $S_t(u) = u + \eta_t\lambda_t^{-1} + \lambda_t^{-1}\sin(\delta_t u)$ for any $\lambda_t > |\delta_t|$.*

In Examples 2.3.1 – 2.3.3, (A2) is clearly satisfied. Moreover, we have $\epsilon_t'(u) = 0$, $\epsilon_t'(u) = \delta_t$ and $\epsilon_t'(u) = \delta_t \cos(\delta_t u)$, respectively in each example. Thus, $\sup_{u \in D_\oplus} |\epsilon_t'(u)| \leq 1 - |\beta|$, so that differentiation and summation can be interchanged, yielding

$$T_t'(u) - 1 = \sum_{i=0}^{\infty} \beta^i \epsilon_{t-i}'(u) \geq - \sum_{i=0}^{\infty} |\beta|^i \sup_{u \in \mathbb{R}} |\epsilon_{t-i}'(u)| > (|\beta| - 1) \sum_{i=0}^{\infty} |\beta|^i = -1.$$

These examples establish one way to validate the WAR(1) model, namely by imposing a deterministic bound on the supremum of the derivative ϵ_t' that is related to β. In general, (A3) may be considered a compatibility restriction between the innovation sequence and the autoregressive parameter.

Next, we express the autoregressive coefficient β in terms of the autocovariance functions γ_h defined in (2.10). Following the derivation of the Yule-Walker equations, it can be shown that

$$\beta = \frac{\int_{\mathbb{R}} \gamma_1(u, u) f_\oplus(u) \mathrm{d}u}{\int_{\mathbb{R}} \gamma_0(u, u) f_\oplus(u) \mathrm{d}u}. \tag{2.17}$$

The denominator is recognizable as the usual Wasserstein variance of each f_t, while the numerator corresponds to the lag-1 scalar measure of Wasserstein covariance defined in Petersen and Müller (2019). Thus, β can be interpreted as a lag-1 Wasserstein autocorrelation measure. This characterization of β thus resembles the autocorrelation function of an AR(1) scalar time series.

Estimation and Forecasting

For any integer $h \geq 0$, define the lag-h Wasserstein autocorrelation function by

$$\rho_h = \frac{\int_{\mathbb{R}} \gamma_h(u, u) f_\oplus(u) \mathrm{d}u}{\int_{\mathbb{R}} \gamma_0(u, u) f_\oplus(u) \mathrm{d}u} = \frac{\int_{\mathbb{R}} \eta_h(u) f_\oplus(u) \mathrm{d}u}{\int_{\mathbb{R}} \eta_0(u) f_\oplus(u) \mathrm{d}u}, \quad \eta_h(u) = \gamma_h(u, u). \tag{2.18}$$

For each fixed u, $\eta_h(u)$ is the autocovariance function of the scalar time series $\{T_t(u), t \in \mathbb{Z}\}$. First, we estimate the Wasserstein mean by

$$\hat{f}_\oplus(u) = \widehat{F}'_\oplus(u), \quad \widehat{F}_\oplus = \left(\frac{1}{n}\sum_{t=1}^{n} Q_t\right)^{-1}. \tag{2.19}$$

Defining $\widehat{T}_t = Q_t \circ \widehat{F}_\oplus$, the estimators for ρ_h and η_h, $h \in \{0, 1, \ldots, n-1\}$, are

$$\hat{\rho}_h = \frac{\int_{\mathbb{R}} \hat{\eta}_h(u)\hat{f}_\oplus(u)\mathrm{d}u}{\int_{\mathbb{R}} \hat{\eta}_0(u)\hat{f}_\oplus(u)\mathrm{d}u}, \quad \hat{\eta}_h(u) = \frac{1}{n}\sum_{t=1}^{n-h}\left\{\widehat{T}_t(u) - u\right\}\left\{\widehat{T}_{t+h}(u) - u\right\}. \tag{2.20}$$

Then the natural estimator for β in (2.16) is

$$\hat{\beta} = \hat{\rho}_1. \tag{2.21}$$

In order to establish asymptotic normality of the above estimators, we require

(A4) The innovations ϵ_t satisfy $\int_{\mathbb{R}} \mathbb{E}\left[\epsilon_t^4(u)\right] f_\oplus(u)\mathrm{d}u < \infty$.

The following result is a special case of Theorem 2.3.4 in Section 2.3.2; the proof of the more general result can be found in Section 2.3.3.

Theorem 2.3.2. *Suppose (A1)–(A4) hold. Then*

$$n^{1/2}\left(\hat{\beta} - \beta\right) \xrightarrow{D} \mathbf{N}\left(0, \sigma_\epsilon^2(1-\beta^2)\right),$$

where

$$\sigma_\epsilon^2 = \frac{\int_{\mathbb{R}^2} C_\epsilon^2(u,v) f_\oplus(u) f_\oplus(v)\mathrm{d}u\mathrm{d}v}{\left[\int_{\mathbb{R}} C_\epsilon(u,u) f_\oplus(u)\mathrm{d}u\right]^2} \tag{2.22}$$

is finite due to (A4).

With a consistent estimator of β in hand, we proceed to define a one-step ahead forecast. Given observations $f_1, \ldots, f_n$, we first obtain $\hat{\beta}$ and compute the measure forecast

$$\hat{\mu}_{n+1} = \mathrm{Exp}_{\hat{f}_\oplus}(\widehat{V}_{n+1}), \quad \widehat{V}_{n+1} = \hat{\beta}(\widehat{T}_n - \mathrm{id}),$$

where $\widehat{T}_n = Q_n \circ \widehat{F}_\oplus$. It remains to convert this measure-valued forecast into a density function. Observe that one can always compute the cdf forecast

$$\begin{aligned} \widehat{F}_{n+1}(u) &= \int_{\mathbb{R}} \mathbf{1}\left(\widehat{V}_{n+1}(v) + v \leq u\right) \hat{f}_\oplus(v)\mathrm{d}v \\ &= \int_0^1 \mathbf{1}\left(\hat{\beta}Q_n(s) + (1-\hat{\beta})\widehat{Q}_\oplus(s) \leq u\right) \mathrm{d}s, \end{aligned} \tag{2.23}$$

where the second line follows from the change of variable $s = \widehat{F}_\oplus(u)$. The cdf forecast can then be converted into a density numerically. The same procedure can be followed to produce further forecasts $\hat{f}_{n+l}$, $l \geq 2$, by using the previous forecast $\hat{f}_{n+l-1}$. In practice, densities are rarely, if ever, fully observed. Instead, one observes samples generated by the random mechanisms characterized by f_i, from which densities can be estimated, e.g., by kernel density estimation. Under certain conditions, see Petersen et al. (2016a) and Panaretos and Zemel (2016), one can systematically account for the deviation from the true densities caused by the estimation process. In our theoretical developments below, we assume that the n densities $f_1, f_2, \ldots, f_n$ are fully observed as our focus is developing the Wasserstein autoregressive model. The numerical implementation of our forecasting procedure, summarized below in Algorithm 1, assumes that the available f_t are bona fide densities, in that they are nonnegative and integrate to one. Additionally, in order to simplify the presentation of the calculations, the algorithm uses the equivalent representation of $\hat{\beta}$ obtained through the change of variable $s = \widehat{F}_{\oplus}(u)$ as

$$\hat{\beta} = \frac{\int_0^1 \hat{\lambda}_1(s)\mathrm{d}s}{\int_0^1 \hat{\lambda}_0(s)\mathrm{d}s}, \quad \hat{\lambda}_h(s) = \hat{\eta}_h(\widehat{Q}_{\oplus}(s)) = \frac{1}{n}\sum_{t=1}^{n-h}(Q_t(s) - \widehat{Q}_{\oplus}(s))(Q_{t+h}(s) - \widehat{Q}_{\oplus}(s)). \quad (2.24)$$

Since $\hat{\lambda}_h(s)$ is computed for $s \in [0,1]$, (2.24) emphasizes that the input densities f_t need not share the same support or be estimated over an identical grid, since all the critical calculations are carried out in terms of quantile functions. Only the quantile functions of the density time series need to be estimated over the same grid points, which extends the flexibility of the model.

Algorithm 1: Forecasting $\hat{f}_{n+1}$

```
1  Input: densities $f_t, t = 1, 2, \ldots, n$;  grid QSup spanning $[0,1]$
   /* Quantities in steps 2--6 are evaluated for $s \in$ QSup          */
2  Evaluate $Q_1(s), Q_2(s), \ldots, Q_n(s)$;
3  $\widehat{Q}_{\oplus}(s) \leftarrow n^{-1}\sum_{t=1}^{n} Q_t(s)$;
4  $\hat{\lambda}_h(s) \leftarrow n^{-1}\sum_{t=1}^{n-h}(Q_t(s) - \widehat{Q}_{\oplus}(s))(Q_{t+h}(s) - \widehat{Q}_{\oplus}(s))$, $h = 0, 1$;
5  $\hat{\beta} \leftarrow \int_0^1 \hat{\lambda}_1(s)\mathrm{d}s / \int_0^1 \hat{\lambda}_0(s)\mathrm{d}s$;
6  $\widehat{V}_{n+1}(\widehat{Q}_{\oplus}(s)) \leftarrow \hat{\beta}(Q_n(s) - \widehat{Q}_{\oplus}(s))$ ;
7  Generate grid dSup spanning
     $(\min_{s \in QSup} \widehat{V}_{n+1}(\widehat{Q}_{\oplus}(s)) + \widehat{Q}_{\oplus}(s), \max_{s \in QSup} \widehat{V}_{n+1}(\widehat{Q}_{\oplus}(s)) + \widehat{Q}_{\oplus}(s))$
   /* Quantities in steps 8--10 are evaluated for $u \in$ dSup         */
8  Compute $\{[a_l, b_l]\}_{l=1}^{L(u)} \leftarrow \left\{s \in [0,1] : \widehat{V}_{n+1}(\widehat{Q}_{\oplus}(s)) + \widehat{Q}_{\oplus}(s) \leq u\right\}$;
   /* $\{[a_l, b_l]\}_{l=1}^{L(u)}$ are disjoint subintervals of $[0,1]$.  */
9  $\widehat{F}(u)_{n+1} \leftarrow \sum_{l=1}^{L(u)} (b_l - a_l)$;
10 $\hat{f}(u)_{n+1} \leftarrow \widehat{F}'(u)_{n+1}$
```

The first step of the algorithm is to convert the available densities f_t into quantile functions. A simple approach to obtain these quantiles from densities is to first evaluate smooth cumulative distribution functions by integrating the estimated densities, followed by some form of numerical inversion. One such approach is readily implemented by the R function dens2quantile from package fdadensity, and this is the approach taken in our numerical experiments to achieve step 2 of the algorithm. Steps 7–9 demonstrate how to implement the exponential map defined in (2.5). From this definition, it is clear that the support of the forecasted density is given by the formula in step 7. Steps 8 and 9 then discover and evaluate the probabilities $\mathrm{Exp}_{\hat{f}_\oplus}(\widehat{V}_{n+1})((-\infty, u])$, for u in the support of the forecasted measure. Finally, step 10 can be executed by numerical integration, for example by computing differences.

2.3.2 Wasserstein AR Model of Order p

A natural way to extend the WAR(1) model is to develop a Wasserstein autoregressive model of order $p \geq 1$ defined by

$$T_t - \mathrm{id} = \sum_{j=1}^{p} \beta_j (T_{t-j} - \mathrm{id}) + \epsilon_t, \tag{2.25}$$

where $\beta_j \in \mathbb{R}, j = 1, 2, \ldots, p$, and the $\epsilon_t \in \mathcal{T}_{f_\oplus}$ are again iid with mean 0 and satisfy (A2). Define the autoregressive polynomial

$$\phi(z) = 1 - \beta_1 z - \beta_2 z^2 - \cdots - \beta_p z^p, \quad z \in \mathbb{C}.$$

The WAR(p) model in (2.25) can then be written as

$$\phi(B)\,(T_t - \mathrm{id}) = \epsilon_t, \tag{2.26}$$

where B is the backward shift operator, i.e., for a discrete stochastic process $\{X_t, t \in \mathbb{Z}\}$, $B^i X_t = X_{t-i}$, $i \in \mathbb{Z}$. For the WAR(p) to have a causal solution, we make the following assumption as a generalization of (A1) in Section 2.3.1.

(A1') The autoregressive polynomial $\phi(z) = 1 - \beta_1 z - \beta_2 z^2 - \cdots - \beta_p z^p$ has no root in the unit disk $\{z : |z| \leq 1\}$.

Under (A1'), $\frac{1}{\phi(z)} = \sum_{i=0}^{\infty} \psi_i z^i$, and the sequence $\{\psi_i\}_{i=0}^{\infty}$ satisfies $\sum_{i=0}^{\infty} |\psi_i| < \infty$. We will show that the solution to equations (2.26) can be written as

$$T_t - \mathrm{id} = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}. \tag{2.27}$$

Observe (2.27) is a strictly stationary and causal process. Similarly to the development of the WAR(1) model, $\{T_t - \mathrm{id}\}$ in (2.25) should be understood at this point as a general zero mean autoregressive process of order p in $\mathcal{T}_{f_\oplus}$. As shown below, (A1') and (A2) together imply the existence of a unique, suitably convergent, solution $T_t - \mathrm{id} = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}(u)$ that is stationary in $\mathcal{T}_{f_\oplus}$ according to Definition 2.2.2. Once again, (A3) applied to $V_t = T_t - \mathrm{id}$ ensures that the application of the exponential map to $T_t - \mathrm{id}$ produces a stationary density time series with mean $f_\oplus$, as seen in the Theorem 2.3.3 below. We also remark that Examples 2.3.1–2.3.3 can be modified directly to guarantee the viability of the WAR(p) model; essentially $1 - |\beta|$ must be replaced with $\left(\sum_{i=0}^{\infty} |\psi_i|\right)^{-1}$.

Theorem 2.3.3. *The following claims hold under Assumptions (A1') and (A2).*

(i) The series (2.27) *is a strictly stationary solution in* $\mathcal{T}_{f_\oplus}$ *to the WAR(p) equation* (2.25). *This solution converges almost surely and in mean square, i.e.,*

$$\lim_{n\to\infty}\left\|T_t - \mathrm{id} - \sum_{i=0}^{n}\psi_i\epsilon_{t-i}\right\| = 0 \quad a.s., \tag{2.28}$$

and

$$\lim_{n\to\infty}\mathbb{E}\left\|T_t - \mathrm{id} - \sum_{i=0}^{n}\psi_i\epsilon_{t-i}\right\|^2 = 0. \tag{2.29}$$

(ii) There is no other stationary solution (according to Definition 2.2.2) in $\mathcal{T}_{f_\oplus}$.

(iii) If, in addition, Assumption (A3) holds for $V_t = T_t - \mathrm{id}$*, then* T_t *is strictly increasing, almost surely, and the measures* $\mathrm{Exp}_{f_\oplus}(T_t - \mathrm{id})$ *possess densities* f_t *that form a strictly stationary sequence according to Definition 2.2.1 with common Wasserstein mean* $f_\oplus$.

Questions of the existence and uniqueness of solutions to ARMA equations are not obvious beyond the setting of scalar innovations, even though care must be exercised even in that standard case, as explained in Chapter 3 of Brockwell and Davis (1991). In the multivariate case, conditions on the spectral decomposition of the autoregressive matrices are needed, see Brockwell and Lindner (2010) and Brockwell et al. (2013) whose results were extended to Banach spaces by Spangenberg (2013). Simpler sufficient conditions in Hilbert spaces are given in Bosq (2000) (AR(p) case) and Klepsch et al. (2017) (ARMA(p,q) case). In our setting, the coefficients are scalars, but the innovations must conform to a nonlinear functional structure, so our conditions involve an interplay between the structure of the functional noise and the coefficients. The fully functional

WAR(1) considered in Chen et al. (2020) is also constructed in the tangent space, so it is also subject to similar constraints as our WAR(p) model, namely that the solution must be restricted to image of the logarithmic map with probability one. We have addressed it through our assumption (A3) and suitable examples or error sequences. Assumption (B2) in Chen et al. (2020) is general, and it is, at this point, unclear whether concrete examples of innovations can be established that satisfy it for fully functional WAR models.

Estimation and Forecasting

Recall $\hat{f}_\oplus$, η_h and $\hat{\eta}_h$ as defined in (2.19), (2.18) and (2.20), respectively. Set $\{\mathbf{H}_p(u)\}_{jk} = \eta_{|j-k|}(u)$, $j,k = 1,\ldots,p$, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^\top$, and $\boldsymbol{\eta}_p(u) = (\eta_1(u),\ldots,\eta_p(u))^\top$. Following the derivation of the Yule-Walker equations, we obtain $\mathbf{H}_p(u)\boldsymbol{\beta} = \boldsymbol{\eta}_p(u)$ as a characterization of the autoregressive parameters of the WAR(p) model, whence

$$\boldsymbol{\beta} = \left(\int_{\mathbb{R}} \mathbf{H}_p(u) f_\oplus(u)\mathrm{d}u\right)^{-1} \int_{\mathbb{R}} \boldsymbol{\eta}_p(u) f_\oplus(u)\mathrm{d}u, \tag{2.30}$$

where the integrals are taken element-wise. Plugging in our estimators $\hat{\eta}_h(u)$ to obtain $\widehat{\mathbf{H}}_p(u)$ leads to

$$\widehat{\boldsymbol{\beta}} = \left(\int_{\mathbb{R}} \widehat{\mathbf{H}}_p(u) \hat{f}_\oplus(u)\mathrm{d}u\right)^{-1} \int_{\mathbb{R}} \hat{\boldsymbol{\eta}}_p(u) \hat{f}_\oplus(u)\mathrm{d}u. \tag{2.31}$$

Set $\{\boldsymbol{\Psi}_p\}_{ij} = \sum_k \psi_k\psi_{k+|i-j|}$, $i,j = 1,\ldots,p$. The following theorem establishes the asymptotic normality of the estimator (2.31).

Theorem 2.3.4. *Suppose (A1'), (A2), (A3), and (A4) hold. Then*

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \mathbf{N}\left(0, \boldsymbol{\Sigma}\right), \tag{2.32}$$

where $\Sigma_{ij} = \sigma_\epsilon^2 \left\{\mathbf{\Psi}_p^{-1}\right\}_{ij}$, $i,j = 1,\dots,p$, *and* σ_ϵ^2 *is the same as* (2.22) *in Theorem 2.3.2.*

Indeed the above asymptotic covariance matrix is a generalization of the asymptotic variance in Theorem 2.3.2. The forecasting procedure is exactly the same as described in (2.23) with steps (4)–(5) of Algorithm 1 replaced by the above steps for estimating $\boldsymbol{\beta}$ and step (6) becoming

$$\widehat{V}_{n+1} = \sum_{i=1}^{p} \hat{\beta}_i (T_{n-i+1} - \mathrm{id}). \tag{2.33}$$

In addition to the autoregressive parameters, the autocorrelation functions are an important object in the study of time series. In our case, recall the lag-h Wasserstein autocorrelation functions are defined in (2.18). Denote $\boldsymbol{\varrho}_h = (\rho_1, \rho_2, \dots, \rho_h)^\intercal$ and $\hat{\boldsymbol{\varrho}}_h = (\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_h)^\intercal$, where $\hat{\rho}_i = \int_{\mathbb{R}} \hat{\eta}_i(u)\hat{f}_\oplus(u)\mathrm{d}u / \int_{\mathbb{R}} \hat{\eta}_0(u)\hat{f}_\oplus(u)\mathrm{d}u$, $i = 1,\dots,h$.

Theorem 2.3.5. *Suppose (A1'), (A2), (A3), and (A4) hold. Then*

$$n^{1/2}(\hat{\boldsymbol{\varrho}}_h - \boldsymbol{\varrho}_h) \xrightarrow{D} \mathbf{N}(0, \mathbf{D}\mathbf{V}\mathbf{D}^\intercal),$$

where

$$\mathbf{D} = \frac{1}{\int_{\mathbb{R}} \eta_0(u) f_\oplus(u)\mathrm{d}u} \begin{bmatrix} -\rho_1 & 1 & 0 & 0 & \dots & 0 \\ -\rho_2 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & & & & \vdots \\ -\rho_h & 0 & 0 & 0 & \dots & 1 \end{bmatrix},$$

and the entries v_{jk}, $j,k = 1,\dots,n-1$, *of* $\mathbf{V}$ *are defined in* (2.42) *and* (2.43) *in Lemma 2.3.7 in Section 2.3.3.*

2.3.3 Proofs of Theorems and Lemmas

Proofs of Theorem 2.3.1 and Theorem 2.3.3

Theorem 2.3.1 is a special case of Theorem 2.3.3 when $p = 1$. Therefore, it suffices to prove Theorem 2.3.3. We begin with a Lemma needed in the proof. It extends Proposition 3.1.2 in Brockwell and Davis (1991) and the discussion that follows that Proposition to Hilbert space valued time series.

Lemma 2.3.6. *Suppose $\{X_t\}$ is a stationary, according to Definition 2.2.2, sequence in a separable Hilbert space.*

(i) If $\sum_{j=1}^{\infty} |\psi_j| < \infty$, then the sequence $\psi(B)X_t := \sum_{j=0}^{\infty} \psi_j X_{t-j}$ is well defined and and is stationary. (The convergence is in the space of square integrable random elements.)

(ii) Consider three filters $\alpha(B), \beta(B), \gamma(B)$ which satisfy $\sum_{j=1}^{\infty} |\alpha_j| < \infty$, $\sum_{j=1}^{\infty} |\beta_j| < \infty$ and define the filter $\gamma(B)$ by setting $\gamma_k = \sum_{l=0}^{k} \alpha_l \beta_{k-l}$, $k \geq 0$. Then, $\sum_{k=1}^{\infty} |\gamma_k| < \infty$ and $\alpha(B)(\beta(B)X_t) = \gamma(B)X_t$.

Proof. We may assume that the mean $\mu = EX_0$ is zero because it adds constant terms like $\mu \sum_{j=0}^{\infty} \psi_j$ or $\mu \sum_{j=1}^{\infty} \alpha_j$ to all arguments.

The proof of claim (i) starts with the verification that $\sum_{j=0}^{n} \psi_j X_{t-j}$ is a Cauchy sequence. This holds because

$$E \left\| \sum_{j=m}^{n} \psi_j X_{t-j} \right\|^2 = E \left| \sum_{i,j=m}^{n} \psi_i \psi_j \langle X_{t-j}, X_{t-j} \rangle \right| \leq \left(\sum_{j=m}^{n} |\psi_j| \right)^2 E\|X_0\|^2.$$

Thus, the limit $\sum_{j=0}^{\infty} \psi_j X_{t-j}$ exists, and by the continuity of the norm $X \mapsto (E\|X\|^2)^{1/2}$, $E\| \sum_{j=0}^{\infty} \psi_j X_{t-j} \|^2 \leq (\sum_{j=0}^{\infty} |\psi_j|)^2 E\|X_0\|^2$. With the convergence established, it is immediate that

$$E\left[\left\langle \sum_{j=0}^{\infty} \psi_j X_{t-j}, x \right\rangle \sum_{i=0}^{\infty} \psi_i X_{t+h-i}\right] = \sum_{i,j=0}^{\infty} \psi_i \psi_j C_{0,j+h-i}(x)$$

does not depend on t.

To prove claim (ii), observe first that $\sum_{k=1}^{\infty} |\gamma_k| \leq (\sum_{j=1}^{\infty} |\alpha_j|)(\sum_{j=1}^{\infty} |\beta_j|) < \infty$. Thus, by part (i),

$$\gamma(B)X_t = \lim_{K\to\infty} \sum_{k=0}^{K} \gamma_k X_{t-k} = \lim_{K\to\infty} \sum_{k=0}^{K} \left(\sum_{l=0}^{k} \alpha_l \beta_{k-l} \right) X_{t-k}.$$

It is useful to visualize the double sum $\sum_{k=0}^{K} \sum_{l=0}^{k} \cdots$ as a sum over the indexes in the (i, j) grid. The summation then extends over a triangle bounded by the diagonal $j = K - i$. We can write

$$\sum_{k=0}^{K} \left(\sum_{l=0}^{k} \alpha_l \beta_{k-l} \right) X_{t-k} = \sum_{i=0}^{K} \alpha_i \sum_{j=0}^{K-i} \beta_j X_{t-i-j}.$$

As $K \to \infty$ and $J \to \infty$, the sum $\sum_{i=0}^{K} \alpha_i \sum_{0\leq j\leq J} \beta_j X_{t-i-j}$ converges $\alpha(B)(\beta(B)X_t)$. It is easy to check that the difference $\sum_{i=0}^{K} \alpha_i \sum_{i<j\leq J} \beta_j X_{t-i-j}$ tends to zero (of the Hilbert space) because the indices i and j are contained in the complement of the rectangle defined by $0 \leq i < K/2$ and $0 \leq j < K/2$. Such details are not provided in Brockwell and Davis (1991), but an argument like this would be needed even in the scalar case. ■

Proof of Theorem 2.3.3. Recall that we work under the setting of the separable Hilbert space $\mathcal{T}_{f_\oplus} \subset L^2(\mathbb{R}, f_\oplus(u)\mathrm{d}u)$ with the inner product $\langle h, g\rangle = \int_{\mathbb{R}} h(u)g(u)f_\oplus(u)\mathrm{d}u$ and norm $\|g\| = \langle g, g\rangle^{1/2}$. To prove claim (i), we first show that the series $\sum_{i=1}^{\infty} \psi_i \epsilon_{t-i}$ converges absolutely almost surely. Mean square convergence follows from part (i) of Lemma 2.3.6. Assumption (A2) implies that there exists some finite $L \in \mathbb{R}$ such that

$$\mathbb{E}\int_{\mathbb{R}} \epsilon_t^2(u) f_\oplus(u)\mathrm{d}u = L < \infty \quad \forall t \in \mathbb{Z}.$$

To show the solution converges almost surely, let $S_n = \sum_{i=0}^n |\psi_i| \, \|\epsilon_{t-i}\|$, $S = \sum_{i=0}^\infty |\psi_i| \, \|\epsilon_{t-i}\|$, then $0 \leq S_n \leq S_{n+1}$ and $\lim_{n\to\infty} S_n = S$. Observe that by Monotone Convergence

$$\begin{aligned}\mathbb{E}\left[S\right] = \lim_{n\to\infty} \mathbb{E}\left[S_n\right] = \lim_{n\to\infty} \sum_{i=0}^n |\psi_i| \, \mathbb{E}\|\epsilon_{t-i}\| \leq \lim_{n\to\infty} \sum_{i=0}^n |\psi_i| \left\{\mathbb{E}\|\epsilon_{t-i}\|^2\right\}^{1/2} \\ = L^{1/2} \sum_{i=0}^\infty |\psi_i| < \infty.\end{aligned}$$

Thus, $S = \sum_{i=0}^\infty |\psi_i| \, \|\epsilon_{t-i}\|$ is finite almost surely. Since S_n is monotone and bounded almost surely, S_n converges almost surely. Therefore

$$\left\|\sum_{i=m}^n \psi_i \epsilon_{t-i}\right\| \leq \sum_{i=m}^n |\psi_i| \, \|\epsilon_{t-i}\| \to 0 \text{ as } m, n \to \infty,$$

so that the sequence of partial sums $\sum_{i=0}^n \psi_i \epsilon_{t-i}$ is Cauchy and converges almost surely.

Set $V_t = \sum_{i=0}^\infty \psi_i \epsilon_{t-i}$. Due to the mean square convergence and the completeness of $\mathcal{T}_{f_\oplus}$, each V_t is an element of $\mathcal{T}_{f_\oplus}$ because, by assumption, $\epsilon_t \in \mathcal{T}_{f_\oplus}$. We must show that

$$V_t = \sum_{j=1}^p \beta_j V_{t-j} + \epsilon_t.$$

With the absolute a.s. convergence of the series defining V_t established, the verification of the above claim proceeds as in the scalar case; all countable manipulations are done for a fixed outcome in an event of probability 1. Changing the order of summation, we obtain

$$\sum_{j=1}^p \beta_j V_{t-j} = \sum_{k=1}^\infty a_k \epsilon_{t-k},$$

with the coefficients a_k defined by

$$\sum_{k=1}^{\infty} a_k z^k = \left(\sum_{j=1}^{p} \beta_j z^j\right)\left(\sum_{i=0}^{\infty} \psi_i z^i\right), \quad |z| \leq 1.$$

Since $\left(1 - \sum_{j=1}^{p} \beta_j z^j\right)\left(\sum_{i=0}^{\infty} \psi_i z^i\right) = 1$, $\psi_0 = 1$ and $\psi_k = a_k$, $k \geq 1$. Consequently,

$$\sum_{j=1}^{p} \beta_j V_{t-j} = \sum_{k=1}^{\infty} \psi_k \epsilon_{t-k} = V_t - \epsilon_t.$$

We now turn to the verification of claim (ii). Suppose $\{V_t^\star\}$ is a stationary sequence in the Hilbert space $\mathcal{T}_{f_\oplus}$ satisfying

$$V_t^\star - \sum_{j=0}^{p} \beta_j V_{t-j}^\star = \phi(B) V_t^\star = \epsilon_t.$$

Using Lemma 2.3.6 and $\phi(z)\psi(z) = 1$, we obtain

$$V_t^\star = \psi(B)(\phi(B)V_t^\star) = \psi(B)\epsilon_t = V_t,$$

proving the uniqueness.

Lastly, we verify claim (iii). By (A3), it is immediate that $(V_t(u) + u)' > 0$ implying $V_t + \mathrm{id}$ is strictly increasing almost surely. Thus, by the structure of $\mathcal{T}_{f_\oplus}$, $V_t + \mathrm{id}$ is effectively an optimal transport map from $\mu_{f_\oplus}$ to some $\mu_{f_t} \in \mathcal{W}_2$. Denote $T_t(u) = V_t(u) + u$. For $\forall a \in \mathbb{R}$, consider

$$\begin{aligned} F_t(a) &= \mathrm{Exp}_{f_\oplus}(V_t)\left((-\infty, a]\right) \\ &= \mu_{f_\oplus}\left((V_t + \mathrm{id})^{-1}(-\infty, a]\right) \\ &= F_\oplus\left(T_t^{-1}(a)\right), \end{aligned}$$

thus $f_t = F_t' = f_\oplus\left(T_t^{-1}\right)\left(T_t^{-1}\right)'$. Consequently, $V_t = \mathrm{Log}_{f_\oplus}(f_t)$ almost surely. Stationarity follows since $\mathbb{E}\left[T_t(u)\right] = u$ implies that $f_\oplus$ is the Wasserstein mean of f_t. ■

Proofs of Theorem 2.3.2 and Theorem 2.3.4

Theorem 2.3.2 is a special case of Theorem 2.3.4 when $p = 1$, hence it suffices to prove Theorem 2.3.4.

Proof of Theorem 2.3.4. The proof relies on a number of technical lemmas whose formulation requires the notation introduced in its course. For this reason, these lemmas are stated and proven after the main body of the proof.

Many manipulations become easier if one works with the two-sided moving average

$$T_t - \mathrm{id} = \sum_{i=-\infty}^{\infty} \psi_i \epsilon_{t-i} \tag{2.34}$$

because one does not have to keep track of indexes corresponding to non-zero coefficients; one must set $\psi_i = 0$ for $i < 0$. Causality is however needed for our proof to go through, see the proof of Lemma 2.3.10.

Recall that Q_t is the quantile function corresponding to f_t and that we assume that $\mathbb{E}_\oplus [f_t] = f_\oplus$ exists and is unique with $Q_\oplus$ and $F_\oplus$ being its quantile function and cdf, respectively. We denote $X_t(s) = Q_t(s) - Q_\oplus(s)$ and $\varepsilon_t(s) = \epsilon_t (Q_\oplus(s))$ throughout the proof. Note that, by the change of variable $s = F_\oplus(u)$, the WAR(p) model in (2.25) can be written as

$$Q_t(s) - Q_\oplus(s) = \sum_{j=1}^{p} \beta_j (Q_{t-j}(s) - Q_\oplus(s)) + \epsilon_t (Q_\oplus(s)), \tag{2.35}$$

Thus, in order to study the properties of $\widehat{\boldsymbol{\beta}}$, we consider the following formulation of the WAR(p) model.

$$\underbrace{\begin{bmatrix} X_0(s) & X_{-1}(s) & \dots & X_{1-p}(s) \\ X_1(s) & X_0(s) & \dots & X_{2-p}(s) \\ \vdots & & & \\ X_{n-1}(s) & X_{n-2}(s) & \dots & X_{n-p}(s) \end{bmatrix}}_{\mathbf{X}(s)} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \varepsilon_1(s) \\ \varepsilon_2(s) \\ \vdots \\ \varepsilon_n(s) \end{bmatrix}}_{\boldsymbol{\varepsilon}(s)} = \underbrace{\begin{bmatrix} X_1(s) \\ X_2(s) \\ \vdots \\ X_n(s) \end{bmatrix}}_{\mathbf{Y}(s)}. \tag{2.36}$$

Some elements of $\mathbf{X}(s)$ are not observable, but are used in our asymptotic analysis. We define the least squares estimator

$$\begin{aligned} \boldsymbol{\beta}^* &= \left\{\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s\right\}^{-1} \left\{\int_0^1 \mathbf{X}^\intercal(s)\mathbf{Y}(s)\mathrm{d}s\right\} \\ &= \left\{\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s\right\}^{-1} \left\{\int_0^1 \mathbf{X}^\intercal(s)\left[\mathbf{X}(s)\boldsymbol{\beta} + \boldsymbol{\varepsilon}(s)\right]\mathrm{d}s\right\} \\ &= \left\{\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s\right\}^{-1} \left\{\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s\boldsymbol{\beta} + \int_0^1 \mathbf{X}^\intercal(s)\boldsymbol{\varepsilon}(s)\mathrm{d}s\right\} \\ &= \boldsymbol{\beta} + \left\{\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s\right\}^{-1} \left\{\int_0^1 \mathbf{X}^\intercal(s)\boldsymbol{\varepsilon}(s)\mathrm{d}s\right\}. \end{aligned} \tag{2.37}$$

Under Assumptions (A1'), (A2), (A3) and (A4), by Lemma 2.3.10,

$$n^{1/2}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}\right) \overset{D}{\to} \mathbf{N}\left(0, \boldsymbol{\Sigma}\right),$$

where $\boldsymbol{\Sigma}$ is as defined in the statement of Theorem 2.3.4. By Lemma 2.3.11,

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = o_p(1),$$

so that

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{D}{\to} \mathbf{N}\left(0, \boldsymbol{\Sigma}\right).$$

■

Proofs of Lemmas

To simplify notation in the proofs, for population quantities in the tangent space, we define alternative versions by applying the change of variable $s = F_{\oplus}(u)$. For instance, we use $X_t(s) = Q_t(s) - Q_{\oplus}(s)$ instead of $T_t(u) - u$, and define $\varepsilon_t(s) = \epsilon_t\left(Q_{\oplus}(s)\right)$. The quantities $\mathbf{X}(s)$ and $\mathbf{Y}(s)$ are defined in (2.36). Additionally, the key parameters γ_h in (2.10) and η_h in (2.18) are replaced by

$$\begin{aligned}\widetilde{\gamma}_h(s, s') :&= \operatorname{Cov}\left(Q_t(s), Q_{t+h}(s')\right) = \operatorname{Cov}\left[T_t \circ Q_{\oplus}(s), T_{t+h} \circ Q_{\oplus}(s')\right] \\ &= \gamma_h\left(Q_{\oplus}(s), Q_{\oplus}(s')\right)\end{aligned} \tag{2.38}$$

and

$$\lambda_h(s) = \eta_h(Q_{\oplus}(s)) = \widetilde{\gamma}_h(s, s) = \operatorname{Cov}\left[Q_t(s), Q_{t+h}(s)\right], \tag{2.39}$$

respectively. Similarly, we define the sample version

$$\hat{\lambda}_s = \hat{\eta}_h \circ \widehat{Q}_{\oplus}(s) = \frac{1}{n}\sum_{t=1}^{n-h}[Q_t(s) - \widehat{Q}_{\oplus}(s)][Q_{t+h}(s) - \widehat{Q}_{\oplus}(s)]. \tag{2.40}$$

Finally, we also define

$$\boldsymbol{\lambda}_p(s) = (\lambda_1(s), \ldots, \lambda_p(s)), \quad \boldsymbol{\Gamma}_p(s) = \mathbf{H}_p(Q_{\oplus}(s)). \tag{2.41}$$

with plug-in estimates $\hat{\boldsymbol{\lambda}}_p(s)$ and $\widehat{\boldsymbol{\Gamma}}_p(s)$.

Lemma 2.3.7. *Assume (A1'), (A2), (A3), and (A4) hold. Consider the following approximation to the sample autocovariance function:*

$$\lambda_h^*(s) = \frac{1}{n}\sum_{t=1}^{n}\left[Q_t(s) - Q_{\oplus}(s)\right]\left[Q_{t+h}(s) - Q_{\oplus}(s)\right], \, h \in \mathbb{Z}.$$

For $i, j = 1, \ldots, n-1$, the following limit exists:

$$\begin{aligned} v_{ij} &:= \lim_{n\to\infty} n \operatorname{Cov}\left[\int_0^1 \lambda_i^*(s)\mathrm{d}s, \int_0^1 \lambda_j^*(s)\mathrm{d}s\right] \\ &= \sum_{r=-\infty}^{\infty} \left(S_1(r)K_2 + S_2(r)K_1 + S_3(r)K_1\right), \end{aligned} \tag{2.42}$$

where

$$\begin{aligned} S_1(r) &= \sum_{k=-\infty}^{\infty} \psi_k\psi_{k+i}\psi_{k+r}\psi_{k+r+j},\ S_2(r) = \sum_{k=-\infty}^{\infty} \psi_k\psi_{k+r} \sum_{l=-\infty}^{\infty} \psi_l\psi_{l+r+j-i}, \\ S_3(r) &= \sum_{k=-\infty}^{\infty} \psi_k\psi_{k+r+j} \sum_{l=-\infty}^{\infty} \psi_l\psi_{l+r-i},\ K_1 = \int_{\mathbb{R}^2} C_\epsilon^2(u,v) f_\oplus(u) f_\oplus(v)\mathrm{d}u\mathrm{d}v, \\ \text{and } K_2 &= \int_{\mathbb{R}^2} \left\{\mathbb{E}\left[\epsilon_t^2(u)\epsilon_t^2(v)\right] - 2C_\epsilon^2(u,v) - C_\epsilon(u,u)C_\epsilon(v,v)\right\} f_\oplus(u) f_\oplus(v)\mathrm{d}u\mathrm{d}v, \end{aligned} \tag{2.43}$$

all of which are well-defined.

Proof. First observe

$$\begin{aligned} &\operatorname{Cov}\left[\int_0^1 \lambda_i^*(s)\mathrm{d}s, \int_0^1 \lambda_j^*(s)\mathrm{d}s\right] \\ =&\operatorname{Cov}\left[\int_0^1 \frac{1}{n}\sum_{t=1}^n X_t(s)X_{t+i}(s)\mathrm{d}s, \int_0^1 \frac{1}{n}\sum_{t'=1}^n X_{t'}(s')X_{t'+j}(s')\mathrm{d}s'\right] \\ =&\frac{1}{n^2}\sum_{t=1}^n\sum_{t'=1}^n \int_0^1\int_0^1 \operatorname{Cov}\left[X_t(s)X_{t+i}(s), X_{t'}(s')X_{t'+j}(s')\right]\mathrm{d}s\mathrm{d}s'. \end{aligned} \tag{2.44}$$

Denote $\sum_i = \sum_{i=-\infty}^{\infty}$ and recall that $\varepsilon_t(s) = \epsilon_t\left(Q_\oplus(s)\right)$. For any $r \in \mathbb{Z}$, define the covariance kernel

$$\begin{aligned} &G_{ijr}(s,s') \\ =&\operatorname{Cov}\left[X_t(s)X_{t+i}(s), X_{t+r}(s')X_{t+r+j}(s')\right] \\ =&\mathbb{E}\left[X_t(s)X_{t+i}(s)X_{t+r}(s')X_{t+r+j}(s')\right] - \mathbb{E}\left[X_t(s)X_{t+i}(s)\right]\mathbb{E}\left[X_{t+r}(s')X_{t+r+j}(s')\right]. \end{aligned} \tag{2.45}$$

Set $t' = t + r$, then (2.44) can be written as

$$\mathrm{Cov}\left[\int_0^1 \lambda_i^*(s)\mathrm{d}s, \int_0^1 \lambda_j^*(s)\mathrm{d}s\right] = \frac{1}{n^2}\sum_{|r|=0}^{n-1}\sum_{t'-t=r}\int_0^1\int_0^1 G_{ijr}(s,s')\mathrm{d}s\mathrm{d}s'.$$

Notice that

$$\begin{aligned}
&\mathbb{E}\left[X_t(s)X_{t+i}(s)X_{t+r}(s')X_{t+r+j}(s')\right]\\
=&\mathbb{E}\left[\sum_k \psi_k\varepsilon_{t-k}(s)\sum_{k'}\psi_{k'}\varepsilon_{t+i-k'}(s)\sum_l \psi_l\varepsilon_{t+r-l}(s')\sum_{l'}\psi_{l'}\varepsilon_{t+r+j-l'}(s')\right]\\
=&\sum_{k,k',l,l'}\psi_k\psi_{k'+i}\psi_{l+r}\psi_{l'+r+j}\mathbb{E}\left[\varepsilon_{t-k}(s)\varepsilon_{t-k'}(s)\varepsilon_{t-l}(s')\varepsilon_{t-l'}(s')\right].
\end{aligned} \tag{2.46}$$

To further analyze (2.46), note that

$$\begin{aligned}
&\mathbb{E}\left[\varepsilon_{t_1}(s)\varepsilon_{t_2}(s)\varepsilon_{t_3}(s')\varepsilon_{t_4}(s')\right]\\
=&\begin{cases}
\mathbb{E}\left[\varepsilon_{t_1}(s)\varepsilon_{t_2}(s)\right]\mathbb{E}\left[\varepsilon_{t_3}(s')\varepsilon_{t_4}(s')\right], & t_1=t_2, t_3=t_4 \text{ and } t_1\neq t_3,\\
\mathbb{E}\left[\varepsilon_{t_1}(s)\varepsilon_{t_3}(s')\right]\mathbb{E}\left[\varepsilon_{t_2}(s)\varepsilon_{t_4}(s')\right], & t_1=t_3, t_2=t_4 \text{ and } t_1\neq t_2,\\
\mathbb{E}\left[\varepsilon_{t_1}(s)\varepsilon_{t_4}(s')\right]\mathbb{E}\left[\varepsilon_{t_2}(s)\varepsilon_{t_3}(s')\right], & t_1=t_4, t_2=t_3 \text{ and } t_1\neq t_2,\\
\mathbb{E}\left[\varepsilon_{t_1}(s)\varepsilon_{t_2}(s)\varepsilon_{t_3}(s')\varepsilon_{t_4}(s')\right], & t_1=t_2=t_3=t_4,\\
0, & \text{otherwise.}
\end{cases}
\end{aligned}$$

Hence (2.46) can be decomposed into the following cases:

$$\begin{cases} k = k', l = l' \text{ and } k \neq l, \\ k = l, k' = l' \text{ and } k \neq k', \\ k = l', k' = l \text{ and } k \neq k', \\ k = k' = l = l', \\ o.w. \end{cases}$$

Denote $C_\varepsilon(s, s') = C_\epsilon(Q_\oplus(u), Q_\oplus(v))$. Also notice that

$$\begin{aligned} \widetilde{\gamma}_h(s, s') &= \mathbb{E}\left[X_t(s)X_{t+h}(s')\right] \\ &= \mathbb{E}\left[\sum_k \psi_k \varepsilon_{t-k}(s) \sum_l \psi_l \varepsilon_{t+h-l}(s')\right] \\ &= \sum_k \psi_k \psi_{k+h} \mathbb{E}\left[\varepsilon_{t-k}(s)\varepsilon_{t-k}(s')\right] \\ &= \sum_k \psi_k \psi_{k+h} C_\varepsilon(s, s'). \end{aligned}$$

Thus, when $k = k', l = l'$ and $k \neq l$,

$$\begin{aligned} &\sum_{k,k',l,l'} \psi_k \psi_{k'+i} \psi_{l+r} \psi_{l'+r+j} \mathbb{E}\left[\varepsilon_{t-k}(s)\varepsilon_{t-k'}(s)\varepsilon_{t-l}(s')\varepsilon_{t-l'}(s')\right] \\ =&\sum_{k \neq l}\sum \psi_k \psi_{k+i} \psi_{l+r} \psi_{l+r+j} C_\varepsilon(s, s) C_\varepsilon(s', s') \\ =&\left\{\sum_k \sum_l \psi_k \psi_{k+i} \psi_{l+r} \psi_{l+r+j} - \sum_k \psi_k \psi_{k+i} \psi_{k+r} \psi_{k+r+j}\right\} C_\varepsilon(s, s) C_\varepsilon(s', s') \\ =&\lambda_i(s)\lambda_j(s') - \sum_k \psi_k \psi_{k+i} \psi_{k+r} \psi_{k+r+j} C_\varepsilon(s, s) C_\varepsilon(s', s'). \end{aligned} \tag{2.47}$$

Similarly, for $k = l, k' = l'$ and $k \neq k'$,

$$\begin{aligned} &\sum_{k,k',l,l'} \psi_k \psi_{k'+i} \psi_{l+r} \psi_{l'+r+j} \mathbb{E}\left[\varepsilon_{t-k}(s)\varepsilon_{t-k'}(s)\varepsilon_{t-l}(s')\varepsilon_{t-l'}(s')\right] \\ =&\widetilde{\gamma}_r(s, s')\widetilde{\gamma}_{r+j-i}(s, s') - \sum_k \psi_k \psi_{k+i} \psi_{k+r} \psi_{k+r+j} C_\varepsilon^2(s, s'); \end{aligned} \tag{2.48}$$

for $k = l', k' = l$ and $k \neq k'$,

$$\begin{aligned}&\sum_{k,k',l,l'} \psi_k \psi_{k'+i} \psi_{l+r} \psi_{l'+r+j} \mathbb{E}\left[\varepsilon_{t-k}(s)\varepsilon_{t-k'}(s)\varepsilon_{t-l}(s')\varepsilon_{t-l'}(s')\right] \\ =&\widetilde{\gamma}_{r+j}(s,s')\widetilde{\gamma}_{r-i}(s,s') - \sum_k \psi_k \psi_{k+i} \psi_{k+r} \psi_{k+r+j} C_\varepsilon^2(s,s');\end{aligned} \tag{2.49}$$

and for $k = k' = l = l'$,

$$\begin{aligned}&\sum_{k,k',l,l'} \psi_k \psi_{k'+i} \psi_{l+r} \psi_{l'+r+j} \mathbb{E}\left[\varepsilon_{t-k}(s)\varepsilon_{t-k'}(s)\varepsilon_{t-l}(s')\varepsilon_{t-l'}(s')\right] \\ =&\sum_k \psi_k \psi_{k+i} \psi_{k+r} \psi_{k+r+j} \mathbb{E}\left[\varepsilon_t^2(s)\varepsilon_t^2(s')\right].\end{aligned} \tag{2.50}$$

Denote $\mathbb{E}\left[\varepsilon_t^2(s)\varepsilon_t^2(s')\right] - 2C_\varepsilon^2(s,s') - C_\varepsilon(s,s)C_\varepsilon(s',s') = \mathcal{K}(s,s')$. By (2.47) - (2.50), we can rewrite the covariance kernel defined in (2.45) as

$$\begin{aligned}&G_{ijr}(s,s') \\ =&\widetilde{\gamma}_r(s,s')\widetilde{\gamma}_{r+j-i}(s,s') + \widetilde{\gamma}_{r+j}(s,s')\widetilde{\gamma}_{r-i}(s,s') + \mathcal{K}(s,s')\sum_k \psi_k \psi_{k+i} \psi_{k+r} \psi_{k+r+j}.\end{aligned} \tag{2.51}$$

By (A4), we have

$$\int_0^1\int_0^1 \mathbb{E}\left[\varepsilon_t^2(s)\varepsilon_t^2(s')\right] \mathrm{d}s\mathrm{d}s' \leq \int_0^1\int_0^1 \mathbb{E}\left[\varepsilon_t^4(s)\right]^{1/2}\mathbb{E}\left[\varepsilon_t^4(s')\right]^{1/2} \mathrm{d}s\mathrm{d}s' < \infty,$$

and

$$\int_0^1\int_0^1 C_\varepsilon^2(s,s')\mathrm{d}s\mathrm{d}s' \leq \int_0^1\int_0^1 \mathbb{E}\left[\varepsilon_t^2(s)\right]\mathbb{E}\left[\varepsilon_t^2(s')\right] \mathrm{d}s\mathrm{d}s' < \infty.$$

Since $\{\psi_k\}$ is absolutely summable, we have

$$\sum_k |\psi_k \psi_{k+i} \psi_{k+r} \psi_{k+r+j}| \leq \sum_k |\psi_k| \sum_{k'} |\psi_{k'}| \sum_l |\psi_l| \sum_{l'} |\psi_{l'}| < \infty.$$

Hence $\int_0^1 \int_0^1 \{\mathcal{K}(s,s') \sum_k \psi_k \psi_{k+i} \psi_{k+r} \psi_{k+i+j}\} \mathrm{d}s\mathrm{d}s' < \infty$. Note that $\widetilde{\gamma}_r(s,s')\widetilde{\gamma}_{r+j-i}(s,s')$ and $\widetilde{\gamma}_{r+j}(s,s')\widetilde{\gamma}_{r-i}(s,s')$ can be bounded in a similar way. Therefore, denoted by τ_r, the double integral of $G_{ijr}(s,s')$ over the unit square is finite, i.e.

$$\tau_r = \int_0^1 \int_0^1 G_{ijr}(s,s')\mathrm{d}s\mathrm{d}s' < \infty.$$

Next, we will show τ_r is absolutely summable in r. Notice that the components of the covariance kernel are absolutely summable in r,

$$\begin{aligned}
&\sum_r |\widetilde{\gamma}_r(s,s')\widetilde{\gamma}_{r+j-i}(s,s')| \\
=&\sum_r \left|\sum_k \psi_k \psi_{k+r} C_\varepsilon(s,s')\right| \left|\sum_l \psi_l \psi_{l+r+j-i} C_\varepsilon(s,s')\right| \\
\leq&\sum_r \sum_k \sum_l |\psi_k \psi_{k+r} \psi_l \psi_{l+r+j-i}| \, C_\varepsilon^2(s,s') \\
\leq&\sum_k |\psi_k| \sum_{k'} |\psi_{k'}| \sum_l |\psi_l| \sum_{l'} |\psi_{l'}| \, C_\varepsilon^2(s,s') < \infty.
\end{aligned} \tag{2.52}$$

Similarly, we have

$$\begin{aligned}
&\sum_r \left|\mathcal{K}(s,s) \sum_k \psi_k \psi_{k+i} \psi_{k+r} \psi_{k+r+j}\right| < \infty, \text{ and} \\
&\sum_r |\widetilde{\gamma}_{r+j}(s,s')\widetilde{\gamma}_{r-i}(s,s')| < \infty.
\end{aligned} \tag{2.53}$$

By (2.52) and (2.53), we have $\sum_r |\tau_r| < \infty$. Hence by the dominated convergence theorem

$$
\begin{aligned}
&\lim_{n\to\infty} n\,\mathrm{Cov}\left[\int_0^1 \lambda_i^*(s)\mathrm{d}s, \int_0^1 \lambda_j^*(s)\mathrm{d}s\right] \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{|r|=0}^{n-1} \sum_{t'-t=r} \int_0^1 \int_0^1 G_{ijr}(s,s')\mathrm{d}s\mathrm{d}s' \\
&= \lim_{n\to\infty} \frac{\left\{\tau_{-(n-1)} + 2\tau_{-(n-2)} + \cdots + (n-1)\tau_{-1} + n\tau_0 + (n-1)\tau_1 + \cdots + \tau_{(n-1)}\right\}}{n} \\
&= \lim_{n\to\infty} \sum_{|r|<n} \left(1 - n^{-1}\,|r|\right)\tau_r \\
&= \sum_{r=-\infty}^{\infty} \tau_r < \infty.
\end{aligned}
$$

It follows that

$$
\lim_{n\to\infty} n\,\mathrm{Cov}\left[\int_0^1 \lambda_i^*(s)\mathrm{d}s, \int_0^1 \lambda_j^*(s)\mathrm{d}s\right] = \sum_{r=-\infty}^{\infty} \left(S_1(r)K_2 + S_2(r)K_1 + S_3(r)K_1\right). \quad (2.54)
$$

■

Lemma 2.3.8. *Assume (A1'), (A2), (A3), and (A4) hold. Then*

$$
\frac{1}{n}\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s \xrightarrow{P} \int_0^1 \boldsymbol{\Gamma}_p(s)\mathrm{d}s, \quad (2.55)
$$

where the convergence holds element-wise.

Proof. Note the ij^{th} element of $\frac{1}{n}\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s$ is

$$
\frac{1}{n}\int_0^1 \sum_{t=1}^{n} X_{t-i}(s)X_{t-j}(s)\mathrm{d}s = \frac{1}{n}\int_0^1 \sum_{t=1-i}^{n-i} X_t(s)X_{t+i-j}(s)\mathrm{d}s = \int_0^1 \lambda_{|i-j|}^*(s)\mathrm{d}s.
$$

By stationarity, $\mathbb{E}\int_0^1 \lambda_{|i-j|}^*(s)\mathrm{d}s = \int_0^1 \lambda_{|i-j|}(s)\mathrm{d}s$. Hence it suffices to show for $i, j =$

$1, \ldots, p$,

$$\lim_{n\to\infty} \mathrm{Var}\left[\int_0^1 \lambda^*_{|i-j|}(s)\mathrm{d}s\right] = 0. \tag{2.56}$$

By Lemma 2.3.7, the variance of $\int_0^1 \lambda^*_{|i-j|}(s)\mathrm{d}s$ converges at rate $O(n^{-1})$, i.e.

$$\lim_{n\to\infty} n\,\mathrm{Var}\left[\int_0^1 \lambda^*_{|i-j|}(s)\mathrm{d}s\right] < \infty. \tag{2.57}$$

Therefore, (2.56) holds and the result follows.

■

Lemma 2.3.9. *Assume (A1'), (A2), (A3), and (A4) hold. Then*

$$\frac{1}{n}\int_0^1 \mathbf{X}^\intercal(s)\mathbf{Y}(s)\mathrm{d}s \xrightarrow{P} \int_0^1 \boldsymbol{\lambda}_p(s)\mathrm{d}s, \tag{2.58}$$

where the convergence holds element-wise.

Proof. The proof is a small modification of the proof of Lemma 2.3.8, so it is omitted. ■

Lemma 2.3.10. *Assume (A1'), (A2), (A3), and (A4) hold. Then*

$$n^{1/2}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta}\right) \xrightarrow{D} \mathbf{N}\left(0, \boldsymbol{\Sigma}\right),$$

where the matrix $\boldsymbol{\Sigma}$ is the same as in Theorem 2.3.4.

Proof. By (2.37),

$$n^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) = n\left\{\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s\right\}^{-1}\left\{n^{-1/2}\int_0^1 \mathbf{X}^\intercal(s)\boldsymbol{\varepsilon}(s)\mathrm{d}s\right\}. \tag{2.59}$$

To further analyze the second factor in (2.59), we set $\mathbf{U}_t(s) = [X_{t-1}(s), \ldots, X_{t-p}(s)]^\intercal \varepsilon_t(s)$, $t \geq 1$. Then

$$n^{-1/2} \int_0^1 \mathbf{X}^\intercal(s) \boldsymbol{\varepsilon}(s) \mathrm{d}s = n^{-1/2} \int_0^1 \sum_{t=1}^n \mathbf{U}_t(s) \mathrm{d}s.$$

The sequence $X_t(s)$ is causal under (A1'), hence it is easy to check $\mathbb{E} \int_0^1 \mathbf{U}_t(s) \mathrm{d}s = 0$ and for $i, j = 1, 2, \ldots, p$,

$$\begin{aligned} &\mathbb{E}\left[\int_0^1 \mathbf{U}_t(s) \mathrm{d}s \int_0^1 \mathbf{U}_t^\intercal(s') \mathrm{d}s'\right]_{ij} \\ =& \int_0^1 \int_0^1 \mathbb{E}\left[X_{t-i}(s) \epsilon_t(s) X_{t-j}(s') \epsilon_t(s')\right] \mathrm{d}s \mathrm{d}s' \\ =& \int_0^1 \int_0^1 \mathbb{E}\left[X_{t-i}(s) X_{t-j}(s')\right] \mathbb{E}\left[\epsilon_t(s) \epsilon_t(s')\right] \mathrm{d}s \mathrm{d}s' \quad \text{(by causality)} \\ =& \int_0^1 \int_0^1 \sum_k \psi_k \psi_{k+|i-j|} C_\varepsilon^2(s, s') \mathrm{d}s \mathrm{d}s' < \infty. \end{aligned} \tag{2.60}$$

Moreover, $\mathbb{E}\left[\int_0^1 \mathbf{U}_t(s) \mathrm{d}s \int_0^1 \mathbf{U}_{t+h}^\intercal(s') \mathrm{d}s'\right]_{ij} = 0$ for $h \neq 0$.

Recall the notation in (2.34), i.e., $X_t(s) = \sum_{k=-\infty}^\infty \psi_k \varepsilon_{t-k}(s)$. For some $m \in \mathbb{Z}^+$, we define the process $X_t^m(s) = \sum_{k=-m}^m \psi_k \varepsilon_{t-k}(s)$ and $\mathbf{U}_t^m(s) = [X_{t-1}^m(s), \ldots, X_{t-p}^m(s)]^\intercal \varepsilon_t(s)$. By (2.60), for $i, j = 1, 2, \ldots, p$, the following expected values exist:

$$\mathbb{E}\left[\int_0^1 \mathbf{U}_t^m(s) \mathrm{d}s \int_0^1 \mathbf{U}^m{}_t^\intercal(s') \mathrm{d}s'\right]_{ij}.$$

For any $\mathbf{a} \in \mathbb{R}^p$ such that $\mathbf{a}^\intercal \mathbb{E}\left[\int_0^1 \mathbf{U}_t^m(s) \mathrm{d}s \int_0^1 \mathbf{U}^m{}_t^\intercal(s') \mathrm{d}s'\right] \mathbf{a} > 0$, $\int_0^1 \mathbf{a}^\intercal \mathbf{U}_t^m(s) \mathrm{d}s$ is an $(m+p)$-dependent process, hence by the Central Limit Theorem for m-dependent processes,

$$n^{-1/2}\sum_{t=1}^{n}\int_0^1 \mathbf{a}^{\intercal}\mathbf{U}_t^m(s)\mathrm{d}s \xrightarrow{D} Z_m, \tag{2.61}$$

where $Z_m \sim \mathbf{N}\left(0, \mathbf{a}^{\intercal}\mathbb{E}\left[\int_0^1 \mathbf{U}_t^m(s)\mathrm{d}s \int_0^1 \mathbf{U}^m{}_t^{\intercal}(s')\mathrm{d}s'\right]\mathbf{a}\right)$.

Clearly $\mathbb{E}\left[\int_0^1 \mathbf{U}_t^m(s)\mathrm{d}s \int_0^1 \mathbf{U}^m{}_t^{\intercal}(s')\mathrm{d}s'\right]_{ij} \to \mathbb{E}\left[\int_0^1 \mathbf{U}_t(s)\mathrm{d}s \int_0^1 \mathbf{U}_t^{\intercal}(s')\mathrm{d}s'\right]_{ij}$ as $m \to \infty$, hence

$$Z_m \xrightarrow{D} Z, \tag{2.62}$$

where $Z \sim \mathbf{N}\left(0, \mathbf{a}^{\intercal}\mathbb{E}\left[\int_0^1 \mathbf{U}_t(s)\mathrm{d}s \int_0^1 \mathbf{U}_t^{\intercal}(s')\mathrm{d}s'\right]\mathbf{a}\right)$.

Moreover, for $\forall n$,

$$\begin{aligned} &n^{-1}\operatorname{Var}\left[\mathbf{a}^{\intercal}\sum_{t=1}^{n}\int_0^1 \left(\mathbf{U}_t^m(s) - \mathbf{U}_t(s)\right)\mathrm{d}s\right] \\ =&\mathbf{a}^{\intercal}\int_0^1\int_0^1 \mathbb{E}\left[\left(\mathbf{U}_t^m(s) - \mathbf{U}_t(s)\right)\left(\mathbf{U}_t^m(s') - \mathbf{U}_t(s')^{\intercal}\right)\right]\mathrm{d}s\mathrm{d}s'\mathbf{a} \to \quad 0 \text{ as } m \to \infty. \end{aligned} \tag{2.63}$$

According to (2.61) through (2.63), by a well-known result used to establish weak convergence via truncation (see Proposition 6.3.9 in Brockwell and Davis (1991)), and the Cramér-Wold device, we have

$$n^{-1/2}\int_0^1 \mathbf{X}^{\intercal}(s)\boldsymbol{\varepsilon}(s)\mathrm{d}s \xrightarrow{D} \mathbf{N}\left(0, \mathbb{E}\left[\int_0^1 \mathbf{U}_t(s)\mathrm{d}s\int_0^1 \mathbf{U}_t^{\intercal}(s')\mathrm{d}s'\right]\right). \tag{2.64}$$

Denote $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}_p^{-1}\mathbb{E}\left[\int_0^1 \mathbf{U}_t(s)\mathrm{d}s \int_0^1 \mathbf{U}_t^{\intercal}(s')\mathrm{d}s'\right]\boldsymbol{\Gamma}_p^{-1}$. By Lemma 2.3.8 and (2.64),

$$n^{1/2}\left(\boldsymbol{\beta}^{*} - \boldsymbol{\beta}\right) \xrightarrow{D} \mathbf{N}\left(0, \boldsymbol{\Sigma}\right).$$

Also by (2.60), we can verify the ij^{th} element of $\mathbf{\Sigma}$ defined in Theorem 2.3.4 is

$$\Sigma_{ij} = \frac{\int_{\mathbb{R}^2} C_\epsilon^2(u,v) f_\oplus(u) f_\oplus(v) \mathrm{d}u\mathrm{d}v}{\left[\int_{\mathbb{R}} C_\epsilon(u) f_\oplus(u) \mathrm{d}u\right]^2} \{\mathbf{\Psi}_p^{-1}\}_{ij} = \sigma_\epsilon^2 \{\mathbf{\Psi}_p^{-1}\}_{ij}, \quad i,j = 1,2,\ldots,p.$$

■

Lemma 2.3.11. *Assume (A1'), (A2), (A3) and (A4) hold. Then*

$$n^{1/2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = o_P(1).$$

Therefore, $n^{1/2}\widehat{\boldsymbol{\beta}}$ and $n^{1/2}\boldsymbol{\beta}^*$ share the same weak limit provided the weak limit exists.

Proof. Note that

$$\begin{aligned}
& n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\
=& n^{1/2}\left[\left(\int_0^1 \widehat{\mathbf{\Gamma}}_p(s)\mathrm{d}s\right)^{-1} \int_0^1 \hat{\boldsymbol{\lambda}}_p(s)\mathrm{d}s - \left(\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s\right)^{-1} \int_0^1 \mathbf{X}^\intercal(s)\mathbf{Y}(s)\mathrm{d}s\right] \\
=& n^{1/2}\left(\int_0^1 \widehat{\mathbf{\Gamma}}_p(s)\mathrm{d}s\right)^{-1}\left(\int_0^1 \hat{\boldsymbol{\lambda}}_p(s)\mathrm{d}s - n^{-1}\int_0^1 \mathbf{X}^\intercal(s)\mathbf{Y}(s)\mathrm{d}s\right) \quad (\star) \\
& + n^{1/2}\left[\left(\int_0^1 \widehat{\mathbf{\Gamma}}_p(s)\mathrm{d}s\right)^{-1} - n\left(\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s\right)^{-1}\right] n^{-1}\int_0^1 \mathbf{X}^\intercal(s)\mathbf{Y}(s)\mathrm{d}s. \quad (\star\star)
\end{aligned}$$

To analyze $(\star)$, first observe

$$
\begin{aligned}
& n^{1/2}\int_0^1 \mathbb{E}\bar{X}^2(s)\mathrm{d}s \\
=& n^{1/2}\int_0^1 \mathbb{E}\left\{\frac{1}{n}\sum_{j=1}^{n} X_j(s)\right\}^2 \mathrm{d}s \\
=& n^{1/2}\int_0^1 n^{-2}\left\{n\lambda_0(s)+2(n-1)\lambda_1(s)+\ldots 2(2)\lambda_{n-2}(s)+2\lambda_{n-1}(s)\right\}\mathrm{d}s \\
=& n^{1/2}\int_0^1 n^{-1}\sum_{h=-n+1}^{n-1}\left(1-\frac{|h|}{n}\right)\lambda_h(s)\mathrm{d}s \\
=& n^{-1/2}\sum_{h=-n+1}^{n-1}\left(1-\frac{|h|}{n}\right)\left\{\sum_{k=-\infty}^{\infty}\psi_k\psi_{k+h}\int_0^1 C_\varepsilon(s,s)\mathrm{d}s\right\} \\
\leq& n^{-1/2}\sum_{h=-n+1}^{n-1}\sum_{k=-\infty}^{\infty}|\psi_k\psi_{k+h}|\left\{\int_0^1 C_\varepsilon(s,s)\mathrm{d}s\right\} \\
=& n^{-1/2}\underbrace{\sum_{k=-\infty}^{\infty}|\psi_k|\left(\sum_{h=-n+1}^{n-1}|\psi_{k+h}|\right)}_{<\infty \text{ as } n\to\infty}\left\{\int_0^1 C_\varepsilon(s,s)\mathrm{d}s\right\} \\
& \to 0 \text{ as } n\to\infty.
\end{aligned}
$$

Therefore $n^{1/2}\int \bar{X}^2(s)\mathrm{d}s \xrightarrow{L1} 0$ implying $n^{1/2}\int \bar{X}^2(s)\mathrm{d}s = o_P(1)$. Hence for $i = 1,2,\ldots,p$, $p<n$,

$$
C_1 := n^{1/2}\int_0^1\left\{\bar{X}(s)\left((1-i/n)\bar{X}(s)-n^{-1}\sum_{j=1}^{n-i}(X_{j+i}(s)+X_j(s))\right)\right\}\mathrm{d}s = o_P(1). \tag{2.65}
$$

It is clear that

$$
C_2 := n^{-1/2}\int_0^1\left\{-\sum_{j=1-i}^{0}X_j(s)X_{j+i}(s)\right\}\mathrm{d}s \to 0 \text{ as } n\to\infty. \tag{2.66}
$$

By (2.65) and (2.66), for $i = 1, 2, \ldots, p$, we have

$$
\begin{aligned}
& n^{1/2}\left(\int_0^1 \hat{\lambda}_i(s)\mathrm{d}s - n^{-1}\int_0^1 \sum_{j=1}^{n} X_{j-i}(s)X_j(s)\mathrm{d}s\right) \\
=& n^{-1/2}\int_0^1 \left\{\sum_{j=1}^{n-i}(X_j(s)-\bar{X}(s))(X_{j+i}(s)-\bar{X}(s)) - \sum_{j=1-i}^{n-i} X_j(s)X_{j+i}(s)\right\}\mathrm{d}s \\
=& n^{-1/2}\int_0^1 \left\{-\sum_{j=1-i}^{0} X_j(s)X_{j+i}(s) - \sum_{j=1}^{n-i}\bar{X}(s)X_{j+i}(s) - \sum_{j=1}^{n-i} X_j(s)\bar{X}(s) + (n-i)\bar{X}^2(s)\right\}\mathrm{d}s \\
=& n^{-1/2}\int_0^1 \left\{-\sum_{j=1-i}^{0} X_j(s)X_{j+i}(s)\right\}\mathrm{d}s \\
& + n^{1/2}\int_0^1 \left\{\bar{X}(s)\left(\left(1-\frac{i}{n}\right)\bar{X}(s) - n^{-1}\sum_{j=1}^{n-i}(X_{j+i}(s)+X_j(s))\right)\right\}\mathrm{d}s \\
=& C_1 + C_2 \\
=& o_P(1).
\end{aligned}
\tag{2.67}
$$

Therefore

$$
n^{1/2}\left(\int_0^1 \hat{\boldsymbol{\lambda}}_p(s)\mathrm{d}s - n^{-1}\int_0^1 \mathbf{X}^{\intercal}(s)\mathbf{Y}(s)\mathrm{d}s\right) = o_P(1). \tag{2.68}
$$

Moreover, we can also conclude from (2.67) that

$$
n^{1/2}\left(\int_0^1 \widehat{\boldsymbol{\Gamma}}_p(s)\mathrm{d}s - n^{-1}\int_0^1 \mathbf{X}^{\intercal}(s)\mathbf{X}(s)\mathrm{d}s\right) = o_P(1). \tag{2.69}
$$

Hence, we can conclude $(\star) = o_P(1)$.

To analyze ($\star\star$), let $\|\cdot\|_F$ be the Frobenius norm, we have

$$
\begin{aligned}
&n^{1/2}\left\|\left(\int_0^1 \widehat{\mathbf{\Gamma}}_p(s)\mathrm{d}s\right)^{-1} - n\left(\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s\right)^{-1}\right\|_F \\
=&n^{1/2}\left\|\left(\int_0^1 \widehat{\mathbf{\Gamma}}_p(s)\mathrm{d}s\right)^{-1}\left(n^{-1}\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s - \int_0^1 \widehat{\mathbf{\Gamma}}_p(s)\mathrm{d}s\right) n\left(\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s\right)^{-1}\right\|_F \\
\leq&n^{1/2}\left\|\left(\int_0^1 \widehat{\mathbf{\Gamma}}_p(s)\mathrm{d}s\right)^{-1}\right\|_F \left\|n^{-1}\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s - \int_0^1 \widehat{\mathbf{\Gamma}}_p(s)\mathrm{d}s\right\|_F \left\|n\left(\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s\right)^{-1}\right\|_F \\
&= o_P(1),
\end{aligned}
$$

since $n\left(\int_0^1 \mathbf{X}^\intercal(s)\mathbf{X}(s)\mathrm{d}s\right)^{-1} \xrightarrow{P} \mathbf{\Gamma}_p^{-1}$, and $\left(\int_0^1 \widehat{\mathbf{\Gamma}}_p(s)\mathrm{d}s\right)^{-1} \xrightarrow{P} \mathbf{\Gamma}_p^{-1}$. Then with Lemma 2.3.9, we can conclude ($\star\star$) $= o_P(1)$. Therefore the claim $n^{1/2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = o_P(1)$ follows. ■

Proof of Theorem 2.3.5

Proof. For some integer $h \in \{0, 1, \ldots, n-1\}$, define the vector

$$
\begin{aligned}
\mathbf{\Lambda}_h^* &= \left[\int_0^1 \lambda_0^*(s)\mathrm{d}s, \int_0^1 \lambda_1^*(s)\mathrm{d}s, \ldots, \int_0^1 \lambda_h^*(s)\mathrm{d}s\right]^\intercal, \\
\mathbf{\Lambda}_h &= \left[\int_0^1 \lambda_0^*(s)\mathrm{d}s, \boldsymbol{\lambda}_h^\intercal\right]^\intercal, \text{ and } \hat{\mathbf{\Lambda}}_h = \left[\int_0^1 \hat{\lambda}_0(s)\mathrm{d}s, \hat{\boldsymbol{\lambda}}_h^\intercal\right]^\intercal.
\end{aligned}
$$

Similarly to the proof of Lemma 2.3.10,

$$
n^{1/2}\left(\mathbf{\Lambda}_h^* - \mathbf{\Lambda}_h\right) \xrightarrow{D} \mathbf{N}\left(0, \mathbf{V}\right),
$$

where $\mathbf{V}$ is the covrariance matrix whose ij^{th} elements v_{ij} are defined in (2.42) and (2.43) in Lemma 2.3.7.

By a similar argument as in the proof of Lemma 2.3.11, we have

$$n^{1/2}\left(\hat{\mathbf{\Lambda}}_h - \mathbf{\Lambda}_h^*\right) = o_P(1).$$

Thus

$$n^{1/2}\left(\hat{\mathbf{\Lambda}}_h - \mathbf{\Lambda}_h\right) \overset{D}{\to} \mathbf{N}\left(0, \mathbf{V}\right).$$

The result follows from an application of the delta method. ■

2.4 Finite Sample Properties of Autoregressive Parameter Estimators

Simulations of the WAR(p) model were conducted to show that the autoregressive coefficients β_j can be accurately estimated, and to explore the normality of the estimators in finite samples. The simulation parameters included the Wasserstein mean density $f_\oplus$ and quantile function $Q_\oplus$, the autoregressive parameters β_j, and a generative process for the innovations ϵ_t. The relation $Q_t(s) = T_t \circ Q_\oplus$ was used to obtain the quantile functions Q_t for use in our algorithms. Simulations were conducted using different Wasserstein mean densities and innovation processes to probe the sensitivity of estimators. In this section, results are presented for a setting in which the Wasserstein mean density corresponds to the uniform distribution on the unit interval, i.e., $Q_\oplus(s) = s$, for $s \in [0, 1]$. Results under more complicated settings can be found in Section 2.4.1.

The optimal transport maps T_t were generated from a WAR(3) model specified by

$$T_t - \mathrm{id} = \beta_1 \left(T_{t-1} - \mathrm{id}\right) + \beta_2 \left(T_{t-2} - \mathrm{id}\right) + \beta_3 \left(T_{t-3} - \mathrm{id}\right) + \epsilon_t, \tag{2.70}$$

with autoregressive coefficients $\beta_1 = 0.825$, $\beta_2 = -0.1875$, $\beta_3 = 0.0125$, and innovations

$$\epsilon_t(u) = \eta_t + \sin\left(\delta_t u\right) \text{ with } \eta_t \overset{iid}{\sim} \mathrm{N}(0,1),\ \delta_t \overset{iid}{\sim} \mathrm{Uniform}[-0.2, 0.2], \quad \eta_t \perp \delta_t,\ u \in [0,1].$$

To begin, it is necessary to generate the initial maps T_1, T_2, and T_3. There exists a unique, stationary and causal solution to (2.70) in the form of (2.27). Hence, one can generate the initial signals purely based on past innovations. A burn-in period of $m = 1000$ was used to stabilize the simulated signals. Given a sequence of m burn-in innovations $\{\epsilon_{1-m}, \epsilon_{2-m}, \ldots, \epsilon_{-1}, \epsilon_0\}$ generated as above, based on (2.27), define

$$\begin{cases} T_{1-m} = \mathrm{id} + \epsilon_{1-m}, \\ T_{2-m} = \mathrm{id} + \epsilon_{2-m} + \beta_1(T_{1-m} - \mathrm{id}), \\ T_{3-m} = \mathrm{id} + \epsilon_{3-m} + \beta_1(T_{2-m} - \mathrm{id}) + \beta_2(T_{1-m} - \mathrm{id}). \end{cases}$$

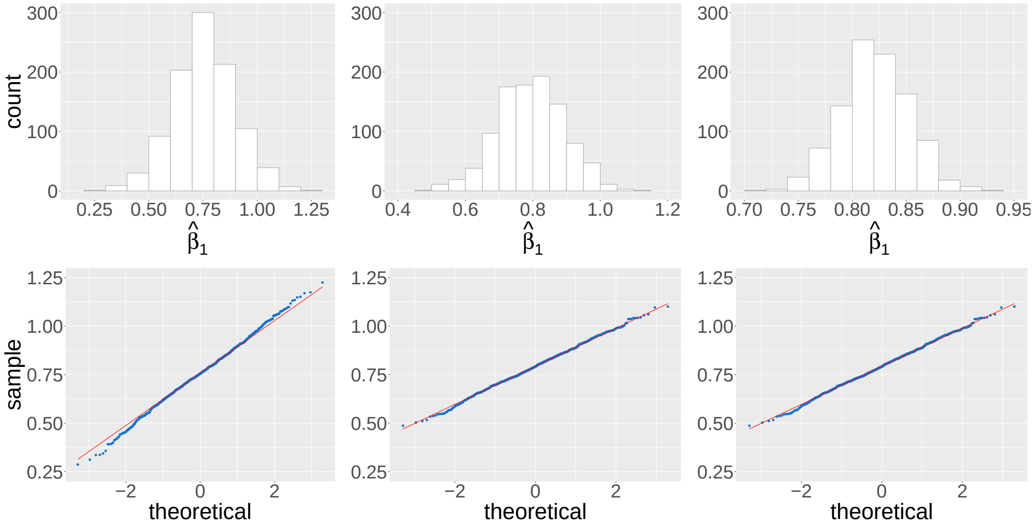
Then (2.70) can be applied recursively until $T_1 - \mathrm{id}$ through $T_3 - \mathrm{id}$ are obtained. One can then generate a time series of desired lengths with $T_1 - \mathrm{id}$ through $T_3 - \mathrm{id}$ and (2.70). This approach is equivalent to truncating the infinite sum in (2.27) but avoids the calculation of the coefficients ψ_i. In our numerical implementation, an equally spaced grid of length 100 on $[0,1]$ was used for both u and s arguments, since the support of the Wasserstein mean and that of the quantile functions are both $[0,1]$ in this setting. The autoregressive parameter estimates in (2.31) were computed using numerical integration.

The simulation was repeated 1000 times with sample sizes $n = 50, 100, 500, 1000, 2000$. The bias, standard deviation and root mean-square error (RMSE) are summarized in

Table 2.1: Bias, standard deviation and RMSE of $\hat{\beta}_i$, $i = 1, 2, 3$.

Sample Size	**Bias**			**SD**			**RMSE**		
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
50	-0.0686	0.0028	-0.0297	0.1432	0.1605	0.1313	0.1588	0.1606	0.1347
100	-0.0319	0.0062	-0.0186	0.0996	0.1171	0.0948	0.1045	0.1172	0.0967
500	-0.0073	0.0022	-0.0028	0.0458	0.0566	0.0453	0.0464	0.0567	0.0454
1000	-0.0043	0.0017	-0.0012	0.0317	0.0406	0.0319	0.0320	0.0406	0.0320
2000	-0.0011	0.0003	-0.0004	0.0227	0.0285	0.0225	0.0228	0.0285	0.0225

Table 2.1, from which we can observe that they all trail off as sample size increases. For the purpose of demonstration, we only display histograms and QQ-plots for $n = 50, 100$ and 1000. The graphical evidence of the asymptotic marginal normality of the estimators $\hat{\beta}_i$, $i = 1, 2, 3$, is presented in Figures 2.2–2.4.

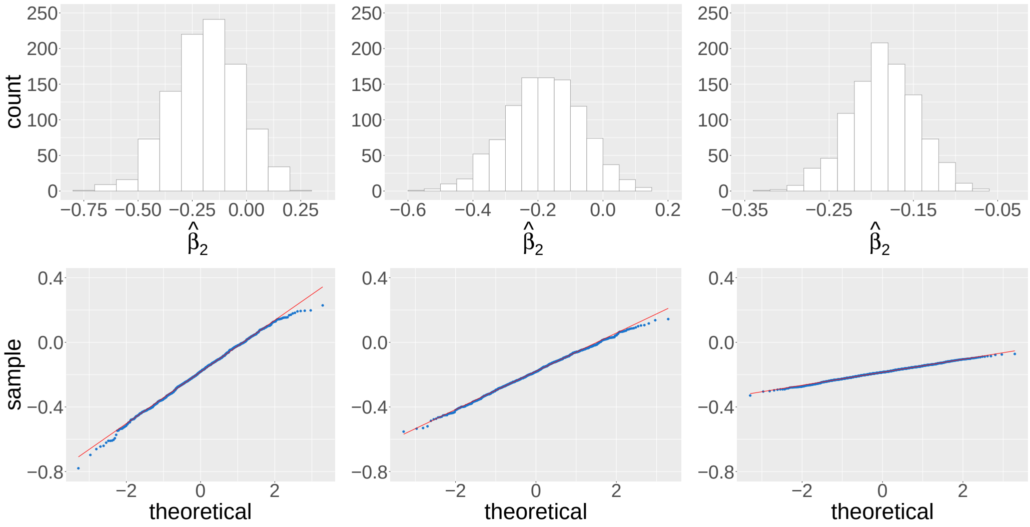


(a) $n = 50$

(b) $n = 100$

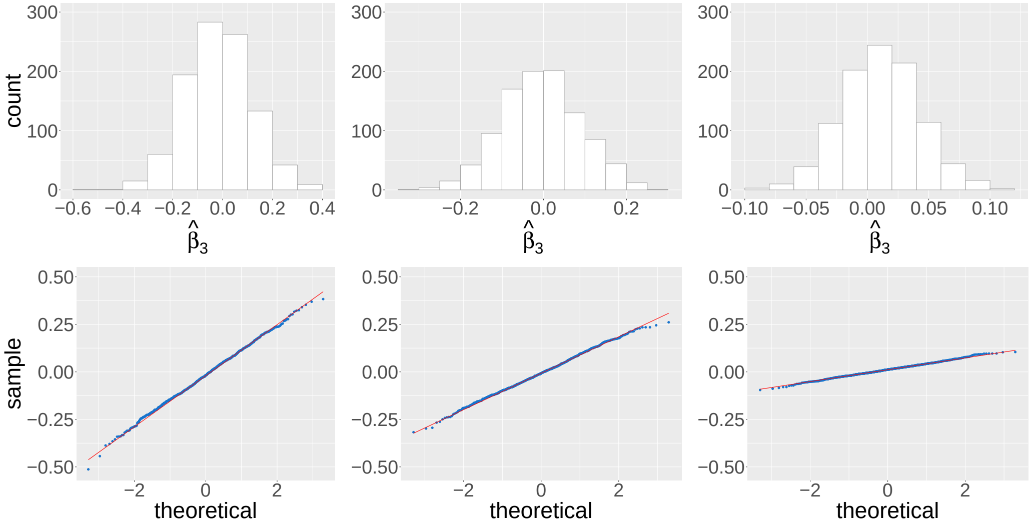
(c) $n = 1000$

Figure 2.2: QQ plots and histograms of $\hat{\beta}_1$



(a) $n = 50$ (b) $n = 100$ (c) $n = 1000$

Figure 2.3: QQ plots and histogram of $\hat{\beta}_2$



(a) $n = 50$ (b) $n = 100$ (c) $n = 1000$

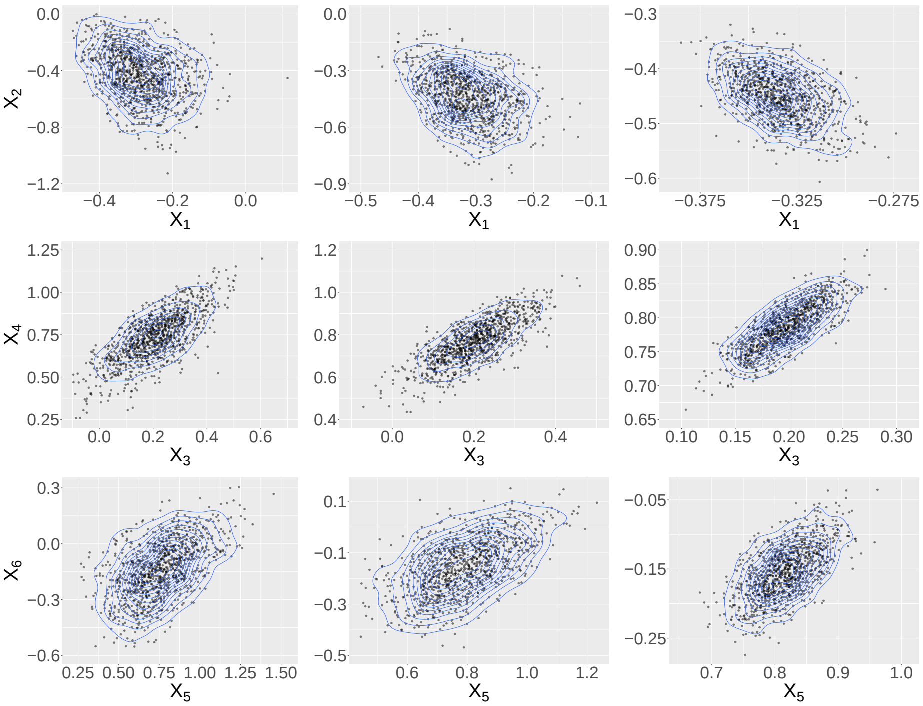
Figure 2.4: QQ plots and histograms of $\hat{\beta}_3$

To investigate the joint normality, denote $\widehat{\boldsymbol{\beta}}_j = [\hat{\beta}_{1j}, \hat{\beta}_{2j}, \hat{\beta}_{3j}]^\intercal$, where $j = 1, 2, \ldots, 1000$ denotes the number of replicates. We randomly generate three pairs of 3×1, linearly independent unit vectors (v_1, v_2), (v_3, v_4) and (v_5, v_6). Calculate $X_{ij} = v_{ij}^\intercal \widehat{\boldsymbol{\beta}}_j$, $i = 1, 2, \ldots, 6$,

$j = 1, 2, \ldots, 1000$. Scatter plots of X_{ij} v.s. $X_{(i+1)j}$, $i = 1, 3, 5$, are shown in Figure 2.5. The idea is that if a vector $[\beta_1, \beta_2, \beta_3]^\intercal$ is normal, then for any coefficients, the vectors $\sum_{j=1}^{3} v_{1j}\beta_j$ and $\sum_{j=1}^{3} v_{2j}\beta_j$ have a joint bivariate normal distribution, which can be approximately verified by visual examination of scatter plots, if replications of $[\beta_1, \beta_2, \beta_3]^\intercal$ are available. As before, we only display the cases where $n = 50, 100$ and 1000 for demonstration. The elliptical patterns in Figure 2.5 suggest bivariate Gaussian distribution, which is what we expect. Moreover, for each n, we calculate $\widehat{\boldsymbol{\Sigma}}$, the sample covariance matrix of $\{\widehat{\boldsymbol{\beta}}_j, j = 1, 2, \ldots, 1000\}$, which is an estimator of the theoretical covariance matrix $\boldsymbol{\Sigma}$ in (2.32). Let $\|\cdot\|_F$ be the Frobenius norm, we use the relative Frobenius norm, $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F / \|\boldsymbol{\Sigma}\|_F$ to measure the differences between the sample covariance matrix and the theoretical asymptotic covariance matrix based on equation (2.32). Figure 2.6 shows that the relative difference approaches zero as sample size increases. All the aforementioned evidence supports the result of Theorem 2.3.4.

2.4.1 Additional Simulation Results

In order to explore the impact of a more complicated Wasserstein mean and noisy innovations on our estimators, we present an additional simulation that assumes all the same settings as in Section 2.4, except that the Wasserstein mean is set to Beta(2,5), and the innovation components $\eta_t \overset{iid}{\sim} \mathrm{N}(0, 25)$, $\delta_t \overset{iid}{\sim} \mathrm{Uniform}[-0.3, 0.3]$. We conduct this simulation for sample sizes $n = 50, 100, 500, 1000, 2000$. Since the theoretical sampling distributions of the estimators are well approximated by their finite sample versions at



(a) $n = 50$ (b) $n = 100$ (c) $n = 1000$

Figure 2.5: Scatter Plots of X_i v.s. X_{i+1}, $i = 1, 3, 5$.

$n = 1000$, we will only present related plots for cases $n = 50, 100, 1000$ for demonstration purpose. The bias, standard deviation and RMSE will be reported for all the sampling sizes.

Figures 2.7–2.9 demonstrate the visual evidence of marginal asymptotic normality of $\hat{\beta}_i, i = 1, 2, 3$. The graphical evidence of asymptotic joint-normality is presented in Figure 2.10. The reader is referred to Section 2.4 for more details of the interpretation of Figure 2.10.

We use the relative Frobenius norm as in Section 2.4 to measure the differences between the sample covariance matrix and the theoretical asymptotic covariance matrix based on equation (2.32) and present the result in Figure 2.11. The difference tends to

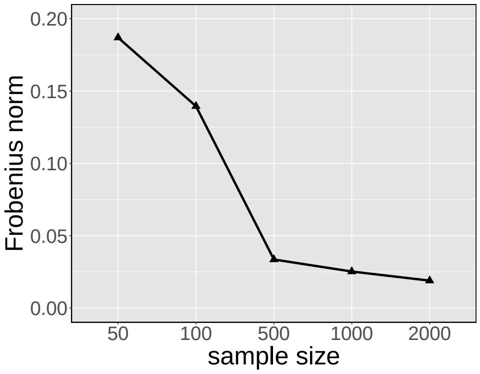
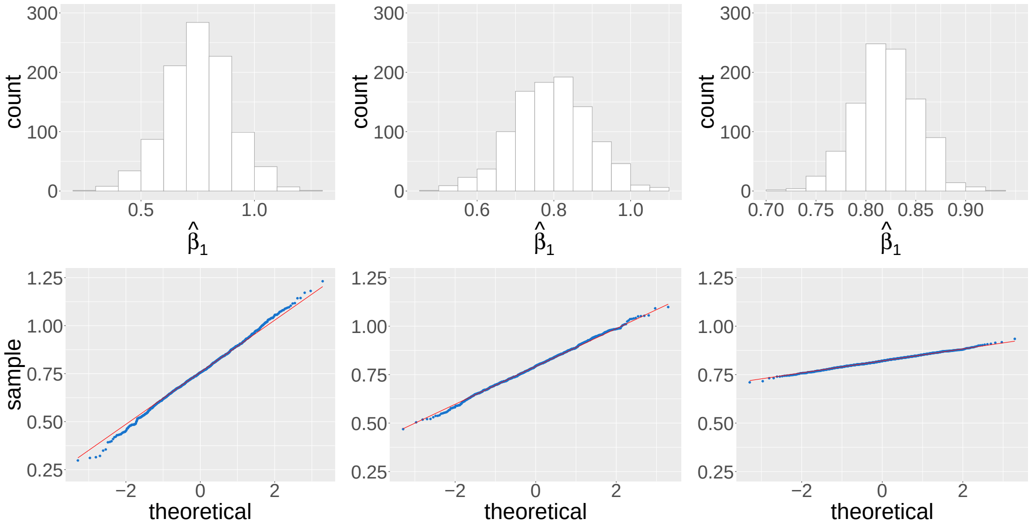


Figure 2.6: Difference between sample and theoretical covariance matrices

Table 2.2: Simulation Beta$(2,5)$: Bias, standard deviation and RMSE of $\hat{\beta}_i$, $i = 1, 2, 3$.

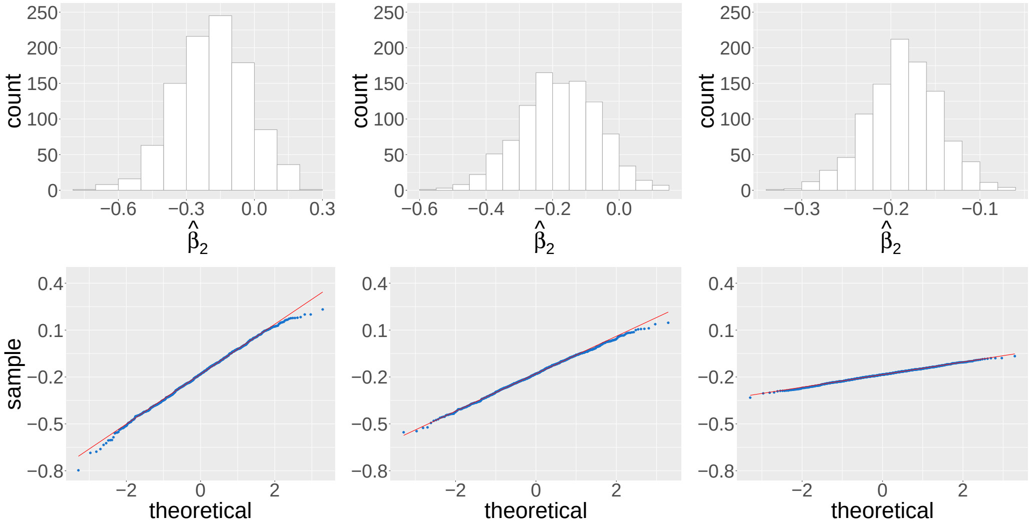
Sample Size	**Bias**			**SD**			**RMSE**		
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
50	-0.0691	0.0032	-0.0294	0.1437	0.1602	0.1311	0.1594	0.1602	0.1342
100	-0.0321	0.0066	-0.0187	0.1001	0.1176	0.0950	0.1051	0.1178	0.0968
500	-0.0074	0.0023	-0.0028	0.0458	0.0565	0.0452	0.0464	0.0566	0.0452
1000	-0.0043	0.0017	-0.0013	0.0318	0.0407	0.0320	0.0321	0.0407	0.0321
2000	-0.0012	0.0003	-0.0005	0.0227	0.0285	0.0225	0.0227	0.0285	0.0225

zero as the sample size gets larger. In particular, a drastic decrease in the difference can be observed as we increase the sample size from 100 to 500. The asymptotic properties remain robust to the choice of a more complicated Wasserstein mean. Lastly, according to Table 2.2, despite that the bias, standard deviation and RMSE of $\hat{\beta}_i, i = 1, 2, 3$ are slightly larger in most cases than those presented in Section 2.4, they all decay consistently as the sample size increases. The theoretical properties of our estimators are well supported in this simulation with Wasserstein mean Beta$(2, 5)$ and increased noise level.



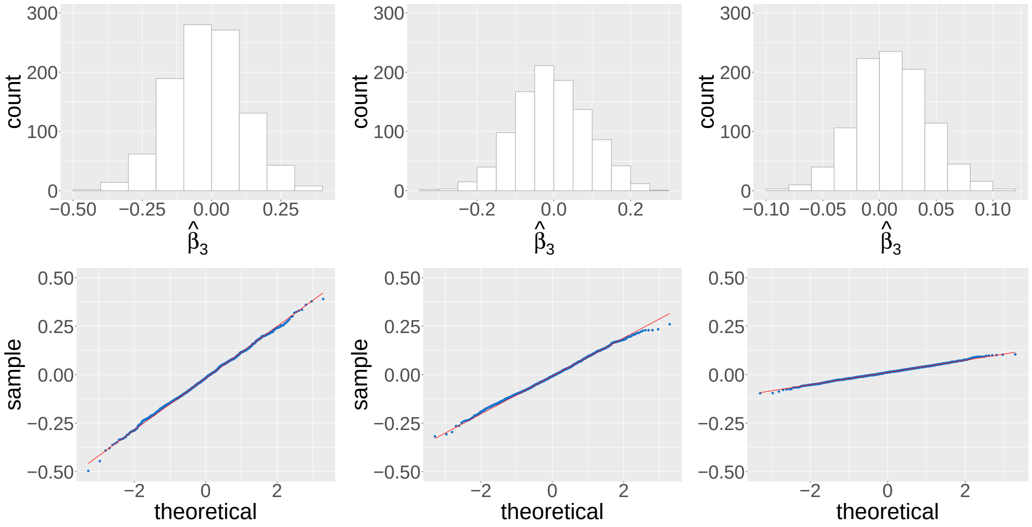
(a) $n = 50$ (b) $n = 100$ (c) $n = 1000$

Figure 2.7: Simulation Beta$(2, 5)$: QQ plots and histograms of $\hat{\beta}_1$



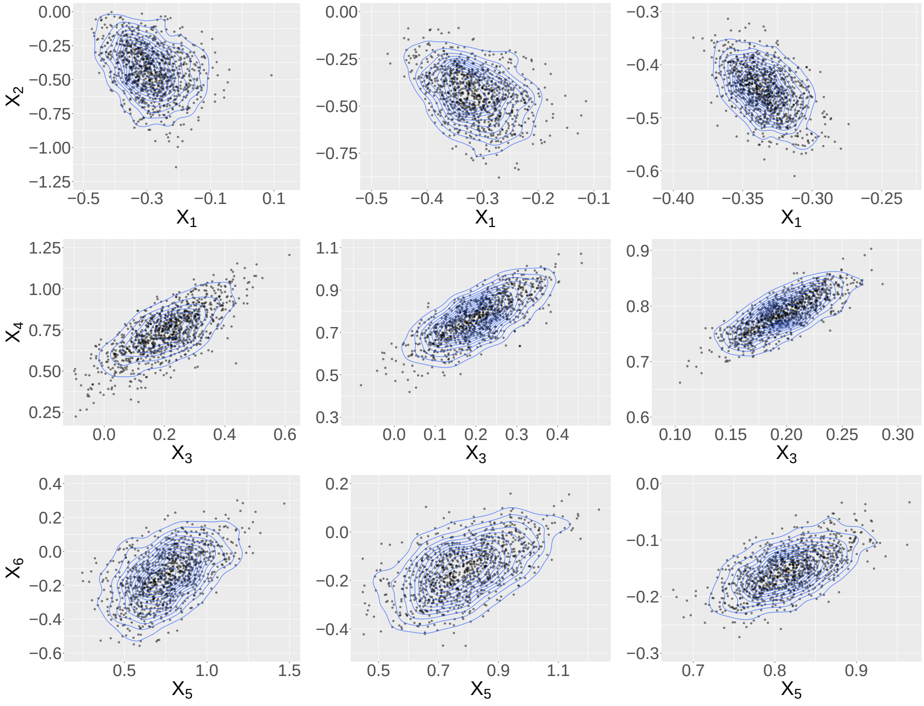
(a) $n = 50$ (b) $n = 100$ (c) $n = 1000$

Figure 2.8: Simulation Beta$(2, 5)$: QQ plots and histogram of $\hat{\beta}_2$



(a) $n = 50$ (b) $n = 100$ (c) $n = 1000$

Figure 2.9: Simulation Beta(2, 5): QQ plots and histograms of $\hat{\beta}_3$



(a) $n = 50$ (b) $n = 100$ (c) $n = 1000$

Figure 2.10: Simulation Beta(2, 5): Scatter Plots of X_i v.s. X_{i+1}, $i = 1, 3, 5$.

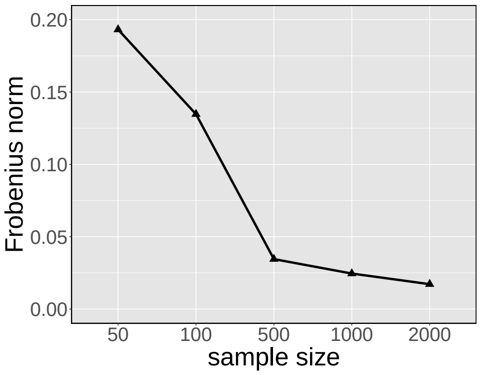


Figure 2.11: Simulation Beta$(2, 5)$: Difference between sample and theoretical covariance matrices

2.5 Comparison with Other Forecasting Methods

We proceed to applying our WAR(1) model to real data sets and comparing its forecasting performance with that of four other density time series forecasting approaches, studied in Kokoszka et al. (2019), where they are introduced in great detail.

2.5.1 Benchmark Methods

We consider the following existing methods.

Compositional Data Analysis. The general methodology of Compositional Data Analysis has been used in various contexts for about four decades, see Pawlowsky-Glahn et al. (2015) for a comprehensive account. Inspired by the similarity between density observations and compositional data, Kokoszka et al. (2019) proposed to remove the constrains on f_t by applying a centered log-ratio transformation. The forecast is produced by first applying FPCA to the output of these transformations, then fitting a time series model to the coefficient vectors.

Log Quantile Density Transformation. This approach is based on the work of Petersen et al. (2016a) and modified by Kokoszka et al. (2019). It transforms the density f_t to a Hilbert space where multiple FDA tools can be applied to forecast the transformed density, then apply the inverse transformation to get the forecast density back. Specifically, a modified log quantile density(LQD) transformation was applied to get the density forecasts.

Dynamic Functional Principal Component Regression. This method was implemented exactly the same way as in Horta and Ziegelmann (2018). Essentially it applies FPCA with a specific kernel, then forecasts the scores with a vector autoregressive(VAR) model. Predictions are produced by reconstructing densities with predicted scores. Negative predictions are replaced by zero and the reconstructed densities are standardized.

Skewed t Distribution. Proposed by Wang (2012), this method fits a skewed t density to data at each time point. Predictions are made by fitting a VAR model to the MLEs of the coefficients of the t distribution.

2.5.2 Data Sets and Performence Metrics

The data sets we use are monthly Dow Jones cross-sectional returns from April 2004 to December 2017, monthly S&P 500 cross-sectional returns from April 2004 to December 2017, Bovespa 5-minute intraday returns that cover 305 trading days from September 1, 2009, to November 6, 2010, and XLK, the Technology Select Sector SPDR Fund returns sampled at the same time intervals as the Bovespa data. The DJIA, S&P 500 and XLK data used in this research are publicly available from the CRSP database (Center for

Research in Security Prices, crsp.org). They are available as supplementary files. The Bovespa data were provided by Capse Investimentos (capse.com.br), and can be requested from that company.

To measure the accuracy of forecast results, we consider the following metrics

1. The discrete version of Kullback-Leibler divergence (KLD; see Kullback and Leibler (1951))

2. The square root of the Jensen-Shannon divergence (JSD; see Shannon (1948))

3. L_1 norm.

Again, we refer to Kokoszka et al. (2019) for more details on the data sets and these metrics as we carry out the comparison exactly the same way as in their paper to keep the comparison consistent.

2.5.3 WAR(p) Models

We implement a data-driven procedure to select the order p and the size of training window K. Denote by n the present time. We use K samples in the time interval $[n-K+1,n]$ to predict f_{n+1}. For each $t \in [n-K+1,n]$ we compute the prediction $\hat{f}_{t,p}$ based on the WAR(p) model and samples in the interval $[t-K,t-1]$. Let ρ be a performance metric, I_p and I_K be some sets for possible choices of p, K, respectively. We evaluate

$$R_p(n,K) = \sum_{t\in[n-K+1,n]} \rho\left(\hat{f}_{t,p}, f_t\right), \quad p \in I_p \text{ and } K \in I_K.$$

Denote by $\hat{p}(n)$ and $\widehat{K}(n)$, the value of p and K which minimizes $R_p(n, K)$, we use WAR($\hat{p}(n)$) and the training window $[n - \widehat{K}(n) + 1, n]$ to predict f_{n+1}. One way to implement this data-driven procedure is to select K and p simultaneously, which entails $|I_p| \times |I_K|$ runs of the forecasting algorithm. In our numerical experiments in this section, we observed that the choice of K has greater impact on the forecasting accuracy than the choice of p. In addition, within the data sets we investigated, the choice of K is relatively robust to the choice of p as the number of window sizes are small, i.e., $|I_K| = 2$ for intraday data sets and $|I_K| = 3$ for cross-sectional data sets (see Section 2.5.5). Therefore, in order to reduce the computational cost, we first use the WAR(1) model to determine K. After choosing training windows for each day, we then determine the order p.

2.5.4 Fully Functional WAR(p) Models

Similar to the idea of the WAR(p) model, one can build a fully functional model in the tangent space to forecast and use the exponential map to recover the forecast density. As mentioned in the introduction, in a recent preprint, Chen et al. (2020) investigated this approach in the case $p = 1$. We specify the general order p model as follows. The fully functional WAR(p) model is defined by

$$T_t(u) - u = \sum_{j=1}^{p} \int_{\mathbb{R}} \phi_j(u, v)(T_{t-j}(v) - v) f_{\oplus}(v) \mathrm{d}v + \epsilon_t(u), \tag{2.71}$$

where ϕ_j are the autoregressive parametric functions to be recovered. Thus, the key difference between the WAR(p) model proposed in this chapter and that of Chen et al.

(2020) is in how the quantities $T_{t-j} - \mathrm{id}$ from previous timepoints are mapped to the tangent space prior to adding the innovations. In the WAR(p) model, these are simply multiplied by the autoregressive coefficients β_j. In contrast, the fully function WAR(p) applies an integral operator with kernel ϕ_j to these quantities. Note that, technically, the WAR(p) model is not a special case of the fully functional version, since the operation of multiplying by β_j is not compact, whereas the integral operators in (2.71) are compact. The estimation procedure follows by fitting the usual functional AR(p) model (see, for example, Bosq (2000)) to the observed quantile functions Q_t, yielding estimates $\hat{\varphi}_j$ of the kernels $\varphi(s, s') = \phi_j(Q_\oplus(s), Q_\oplus(s'))$. In the case $p = 1$, this matches the estimation of Chen et al. (2020). Similarly to the WAR(p) model, forecasts are then constructed in the tangent space using the plug-in estimates $\hat{\phi}_j(u, v) = \hat{\varphi}_j(\hat{F}_\oplus(u), \hat{F}_\oplus(v))$, followed by application of the exponential map (2.5). Thus, in the presentation of our results, the method labeled "Fully Functional WAR(p)" can be considered as an extension of the model of Chen et al. (2020) to include orders $p \geq 1$.

In particular, we implement the same data-adaptable procedure as described in Section 2.5.3 with one additional component. The method used to fit the functional AR(p) model to the quantile functions performs functional principal component analysis as a first, which requires one to specify the number of components to retain. We thus introduce an additional tuning parameter R that represents proportion of variance required by the FPCA. Specifically, in the forecasting procedure, we reconstruct $\widehat{T}_t - \mathrm{id}$ with the smallest number of PCs that explain R percent of variance; see, for example, Section 3.3

of Horváth and Kokoszka (2012). We incorporate R into the data-driven procedure to determine its value for forecasting. Specifically, we compute

$$R_p(n, K, R) = \sum_{t \in [n-K+1, n]} \rho\left(\hat{f}_{t,p}, f_t\right),$$

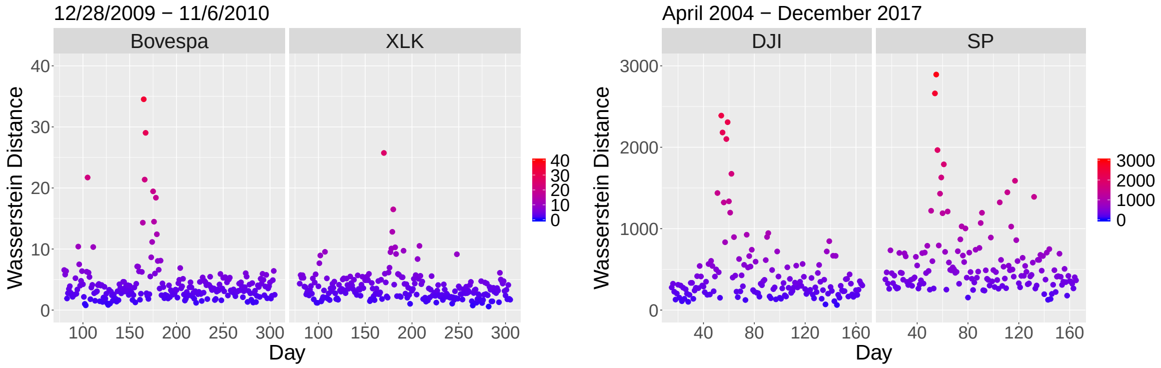
where $p \in I_p, R \in I_R$ and $K \in I_K$. For each n, we use the optimal $\hat{p}(n)$, $\widehat{K}(n)$ and $\widehat{R}(n)$ to predict $\hat{f}_{n+1}$. Within the fully functional WAR(p) model, some initial results show that the case $p = 1$ outperforms higher order cases across all different settings of K and R, hence to simplify the procedure, we fix $p = 1$ and implement the procedure to choose R and K.

2.5.5 Results

The WAR(p) model was tuned with both Kullback-Leibler divergence and Wasserstein distance under the data-adaptable procedure with $I_p = \{1, 2, \ldots, 10\}$, while the fully functional WAR(p) model was only tuned with the former one for demonstration purpose with $I_R = \{0.4, 0.5, \ldots, 0.8\}$. For both approaches, we use $I_K = \{20, 62\}$ for the intra-day data sets and $I_K = \{12, 24, 48\}$ for the monthly cross-sectional data sets. These choices correspond approximately to monthly and quarterly data (20, 62) and to 1, 2, and 4 years (12, 24, 48) for the monthly data. They are often used for financial and economic data, but there is no profound statistical reason for choosing them. Our method could be elaborated on by using a data driven maximum value of K, some form of an approach advocated in Chen et al. (2010), but the simple choices we propose work well and do not lead to an excessive computational burden.

From Tables 2.3–2.6, we can see both WAR(p) and fully functional WAR(p) models produce excellent predictions in the XLK and DJI data sets. (In 19 out of 20 cases the WAR(p) performs better than the fully functional WAR(p).) Indeed, the WAR(p) model is the top performer in these two data sets. In the XLK data set, the WAR(p) model tuned by KL divergence topped under three performance metrics, and ranked second under the rest two metrics with small margins to the top performer LQDT. In the DJI data set, the WAR(p) model topped under two metrics, and again, with narrow margins to the top performers under the rest of the metrics. Specifically, we can see in DJI data set, the average rank of forecasting performance of WAR(p) model (tuned by KL divergence) is 1.6, while the two contenders LQDT and CoDa (no standardization) scored 2.8 and 1.6, respectively, which put the WAR(p) model in tie with the CoDa method as the top performers.

The performance of WAR(p) model in the Bovespa and S&P500 data sets is not as competitive. Since our models rely on stationarity, we informally investigate the stationarity condition for each data set. In Figure 2.12, we plot the Wasserstein distance from all densities used in forecasting to their sample Wasserstein mean. These distances are larger in the Bovespa and S&P500 data sets, compared to those in XLK and DJI data sets. Indeed, the average Wassertein distance from these plots in Figure 2.12 are XLK: 4.045, Bovespa: 4.255, DJI: 421.25 and S&P500: 571.63. Hence stationarity could be a potential cause for a weaker performance of the WAR(p) model in the Bovespa and S&P500 data sets. Generally, no prediction method can be expected to be uniformly superior across all data sets and all time periods and according to all metrics. In our empirical study, The WAR(p) methods performs best for some data sets, and the LQDT and CoDa methods perform better for others.



(a) Intraday Returns

(b) Monthly Returns

Figure 2.12: Wasserstein Distance Between Sample Points To Their Wasserstein Mean

Table 2.3: Forecast accuracies of five methods, XLK intraday returns

Method	KLdiv	JSdiv	JSdiv.geo	L1	Wasserstein
Horta-Zieglman	0.2831	1.5095	4.2909	11257.47	3.97×10^{-4}
LQDT	0.3831	**1.3411**	5.2559	**10891.16**	3.97×10^{-4}
CoDa(standardization)	0.3231	2.6076	4.9518	14689.67	4.04×10^{-4}
CoDa(no standardization)	0.3579	2.8919	5.2173	15053.57	4.11×10^{-4}
Skewed-t	0.2666	1.7418	3.8736	13701.89	4.16×10^{-4}
WAR(p) (KL)	**0.1761**	1.4408	**2.7569**	11214.40	$\mathbf{3.32 \times 10^{-4}}$
WAR(p) (WD)	0.1827	1.4713	2.8730	11418.83	3.38×10^{-4}
Fully Functional WAR(p) (KL)	0.1837	1.4753	2.8821	11576.42	3.36×10^{-4}

Table 2.4: Forecast accuracies of five methods, Bovespa intraday returns

Method	KLdiv	JSdiv	JSdiv.geo	L1	Wasserstein
Horta-Ziegelman	0.4009	1.9098	6.1713	16993.19	4.47×10^{-4}
LQDT	0.4258	**1.6634**	6.0687	**16313.87**	3.09×10^{-4}
CoDa(standardization)	**0.2271**	1.7360	**3.7000**	16351.17	$\mathbf{3.08 \times 10^{-4}}$
CoDa(no standardization)	0.2278	1.7448	3.7038	16391.76	3.10×10^{-4}
Skewed-t	0.2750	1.9909	3.9774	19261.90	4.13×10^{-4}
WAR(p) (KL)	0.2534	1.8769	4.1364	17153.26	3.92×10^{-4}
WAR(p) (WD)	0.2383	1.8065	3.8622	16878.16	3.86×10^{-4}
Fully Functional WAR(p) (KL)	0.2550	1.8963	4.1478	17226.79	3.79×10^{-4}

Table 2.5: Forecast accuracies of five methods, Dow-Jones cross-sectional returns

Method	KLdiv	JSdiv	JSdiv.geo	L1	Wasserstein
Horta-Ziegelman	1.3070	3.5986	9.4038	1039.36	3.99×10^{-2}
LQDT	1.0421	**3.0129**	6.9443	948.77	2.61×10^{-2}
CoDa(standardization)	0.6658	3.2359	5.1780	953.42	2.63×10^{-2}
CoDa(no standardization)	0.6510	3.1785	**5.0572**	**943.62**	$\mathbf{2.59 \times 10^{-2}}$
Skewed-t	1.3590	5.2532	10.4784	1324.97	3.82×10^{-2}
WAR(p) (KL)	**0.6448**	3.0407	5.0965	947.0983	$\mathbf{2.59 \times 10^{-2}}$
WAR(p) (WD)	0.6616	3.1838	5.1538	975.3546	2.63×10^{-2}
Fully Functional WAR(p) (KL)	0.6480	3.0821	5.0993	952.4613	2.61×10^{-2}

Table 2.6: Forecast accuracies of five methods, S&P 500 cross-sectional returns

Method	KLdiv	JSdiv	JSdiv.geo	L1	Wasserstein
Horta-Ziegelman	0.5315	1.9986	3.1032	222.62	6.94×10^{-2}
LQDT	0.4252	1.8165	2.5232	213.10	$\mathbf{4.78\times10^{-2}}$
CoDa(standardization)	**0.3156**	**1.7994**	**2.3023**	**208.71**	6.45×10^{-2}
CoDa(no standardization)	0.3233	1.8465	2.3550	211.29	6.50×10^{-2}
Skewed-t	0.5560	3.0961	3.6383	286.04	6.67×10^{-2}
WAR(p) (KL)	0.4454	1.9578	2.7626	213.2848	7.37×10^{-2}
WAR(p) (WD)	0.4349	1.9166	2.7163	216.4794	7.23×10^{-2}
Fully Functional WAR(p) (KL)	0.4762	2.1384	2.8143	223.7424	7.91×10^{-2}

2.6 Discussion

The WAR(p) model provides an interpretable approach to model density time series by representing each density through its optimal transport map from the Wasserstein mean. Under this representation, stationarity of a density time series, whose elements reside in a nonlinear space, is defined according to the usual stationarity of the random transport maps in the tangent space, which is a separable Hilbert space. This chapter demonstrates how autoregressive models, built on the tangent space corresponding to the Wasserstein mean, possess stationary solutions that, in turn, define a stationary density time series. This link is not automatic, however, due to the fact that the logarithmic map lifting the densities to the tangent space is not surjective, and constraints are necessary to ensure the viability of the model. In our empirical analysis, the proposed WAR(p) model emerged as a competitive forecasting method for financial return densities when compared to various existing methods and using several different metrics for forecasting accuracy. The option of selecting the order p to suit a specific purpose is a useful feature of the model. We proposed a data-driven procedure that targets optimal prediction in terms of a specific metric, but other objectives, including a model fit in terms of information criteria could be used as well.

There are several research directions that emerge from our work. It can be expected that the theory for more general ARMA(p, q) processes can be developed by extending the arguments we used. However, as we discussed, even scalar ARMA processes are theoretically more complex than pure AR(p) models and ARMA processes in function spaces must be approached with particular care. The extension thus appears to be not trivial, but may turn out to be useful for some purposes. In the case of scalar, but not necessarily vector, observations, ARMA processes provide more parsimonious models, but their predictive performance is not necessarily better that that of AR(p) models. ARMA predictors are constructed through the Durbin-Levinson or innovations algorithms, but truncated predictors, effectively equivalent to order selected AR(p) models, generally perform better, see e.g. Section 3.5 of Shumway and Stoffer (2018).

We explored empirically the fully functional WAR(p) model, but we did not pursue its theoretical underpinnings because its predictive performance was not competitive; simpler models often provide better predictions. The theory of fully functional WAR(1) model was developed, independently and in parallel with our research, in Chen et al.

(2020). It is also a model constructed in the tangent space and so it is subject to similar constraints as our WAR(p) models, namely that the solution must be restricted to image of the logarithmic map with probability one (see assumption (A3) in this chapter, and assumption (B2) in Chen et al. (2020)). It is unclear whether concrete examples of innovations can be established that satisfy this constraint for fully functional WAR(p) models, whereas we have established several concrete examples for WAR(p) models in this chapter. Still, fully functional WARMA(p, q) models might be useful in some settings, and their theory might then be developed.

We have seen that, as for any time series models, assumptions of stationarity are key to establishing theoretical properties, such as the asymptotic normality of the WAR parameters and Wasserstein autocorrelations, and to good forecasting performance. Research on testing stationarity and detecting possible change points may be facilitated by our work. Research of this type has been done for linear functional time series, see e.g. Berkes et al. (2009), Horváth et al. (2014), Zhang and Shao (2015), but not for density times series. In general, it is hoped that this chapter not only provides a set of theoretical and practical tools, but also lays out a framework within which questions of inference for density time series can be addressed.

Chapter 3

Quantifying Brain Functional Connectivity with Noisy Voxel Level Signals

3.1 Introduction

In recent years, with the rapid advancement and increasing accessibility of neuroimaging techniques, data sets that record brain activities, such as electroencephalogram (EEG) scans and functional magnetic resonance imaging (fMRI) time series, are becoming widely available to scientists and medical practitioners. The rich volume of previously unattainable data sets has consequently catalyzed a wide range of interests in modeling and estimating functional brain connectivity, which is of paramount importance in shedding lights on the evolution of pathologies such as neurodegenerative diseases or consciousness disorders.

However, challenges arise in preprocessing and conducting robust and reproducible analyses with the massive amount of data. Signals collected from human brains are frequently modeled as realizations of random fields that traverse across both the spatial and temporal domains (Achard et al., 2011; Achard and Gannaz, 2019). These signals are usually of noisy nature due to the inherent properties of random fields and (or) the measuring instrument errors (Chaimow et al., 2018). Various methods have been proposed to overcome these challenges. Achard et al. (2006) applied a wavelet transformation to obtain brain networks based on frequency-dependent correlations between regional fMRI time series. Machine learning techniques, in particular, feature embedding and clustering are investigated to classify and characterize changes in brain dynamics due to pathology or cognitive state changes (Richiardi et al., 2013). Termenon et al. (2016) studied the relationship between the reproducibility of brain networks and the subject count and fMRI scan length using the large test-retest (TRT) resting-state fMRI data set from the Human Connectome Project (HCP). In addition, Petersen et al. (2016b) proposed to view the path lengths of brain networks as functions of network density, so that functional principal component analysis (fPCA) are used and covariates are regressed on functional principle component scores to study the association between connectivity and subjects' age, episodic memory, and executive function.

Functional connectivity quantification and estimation is a primary goal of these neuroscience inquiries. Following the characterization of Van Den Heuvel and Pol (2010), functional connectivity is the temporal dependency of neuronal activation patterns of anatomically separated brain regions. Consequently, the vast majority of existing studies aggregate signals within certain regions instead of directly modeling voxel level signals. This definition is also known as the *inter-regional*, or sometimes referred to as the *long-range* connectivity. Brain regions usually refer to the disjoint anatomical or functional parcellation of the brain. The choice of the parcellations is still an open question and is attracting active investigations. In this chapter, we will use a parcellation of rat brains that contains 51 regions, and with our model, we aim to provide a way to quantify functional connectivity and construct brain networks on the individual level. Another measurement of connectivity is termed the *intra-regional* connectivity, which measures the connectivity within a particular brain region. We will emphasize on modeling the inter-regional connectivity as it is of primal interest of a large body of neuroimaging studies (Eickhoff et al., 2018; Moghimi et al., 2021).

A brain can be considered as a network. Thus, it is not surprising that network-based approaches are prevalent in inter-regional connectivity modeling (Van Den Heuvel and Pol, 2010; Meskaldji et al., 2011; Kaiser, 2011). The nodes and edges in a brain network represent brain regions and connections, respectively. An edge is identified by a high degree of some similarity metric for the brain activity signals, which is usually quantified by the Pearson correlation (Achard et al., 2006; Zalesky et al., 2012; Becq et al., 2020). While this estimator provides a consistent estimation in general, it does not take the spatiotemporal noise that is inherent to BOLD signals into account (Achard and Gannaz, 2019). More details on this matter will be discussed in Section 3.2 as we develop our model. In addition, While some of these existing methods can localize connectivity to the subject level (Petersen et al., 2016b), the analysis phase usually requires group level information.

In this chapter, we focus on the analysis of resting-state fMRI data on a subject level using signals at the voxel level. The main contribution of this paper is twofold. First, we propose a novel spatiotemporal statistical model for the BOLD signals at the voxel level. These signals are considered as realizations of a Gaussian process with a carefully designed covariance structure that explicitly models the spatiotemporal dependency of BOLD signals. The inter-regional connectivity between two arbitrary regions is quantified by a primal model parameter, which needs to be estimated along with a collection of auxiliary parameters that characterize the statistical properties of the model. One should keep in mind that our model has the flexibility to be extended to relatively general spatiotemporal processes. However, we will restrict our discussion within the scope of brain connectivity quantification using BOLD signals in this chapter as it is the intended application at the current stage. We also derive the large sample properties of the model parameter estimators, propose methods to construct brain networks using the estimated parameters, and conduct simulations to investigate robustness and reproducibility of the analysis. Secondly, we devise an efficient two-stage strategy to estimate model parameters as computations associated with Gaussian processes are usually prohibitive. We also document the detailed implementation of our method, which involves large scale parallel computing with GPU accelerated matrix operations.

The remainder of the chapter is arranged as follows. In Section 3.2, we present model construction, parameterization, and estimation. In particular, we propose the two-stage estimation procedure in Section 3.2.3. Assumptions and results on the asymptotic behaviour of our model parameter estimators are presented in Section 3.3. Section 3.4 provides a set of simulation studies that demonstrates the empirical performance of our estimators. In Section 3.5, our model is applied to real rat data sets and brain connectivity is quantified by the estimated model parameters. We further construct brain networks with the estimated parameters by utilizing their large sample properties. Section 3.6 discusses results and the potential extension to this chapter.

3.2 Model

Our main goal is to model the BOLD signals at the voxel level such that spatiotemporal features are sufficiently addressed and brain connectivity, hence a network, can be recovered from the estimated model parameters. The model is developed in a data-inspired manner, that is, we propose our model with a clear goal to model the properties of BOLD signals, even though such properties may not be exclusive to BOLD signals only and the resulting model can be applied to more general spatiotemporal processes. Then we proceed with the investigation of model estimation and large sample properties.

3.2.1 A Spatiotemporal Model for BOLD Signals

A brain volume $\mathcal{B}$ is divided into spatially disjoint and contiguous regions $\mathcal{R}_j$, that is, $\mathcal{B} = \cup_{j=1}^{J} \mathcal{R}_j$ and $\mathcal{R}_j \cap \mathcal{R}_{j'} = \emptyset$, $j \neq j'$. In principle, brain activity is a continuous process, though we do not observe it directly or cleanly. So, we begin by building a model for the unobserved (thus latent) process which will be the foundation for the data model in (3.2). To that end, the latent BOLD signals from voxel $v \in \mathcal{R}_j$, $j = 1, \ldots, J$, at time $t \in [0, \mathcal{T}]$, where $\mathcal{T}$ is the duration of fMRI scan, is modeled by

$$Y_j(v, t) = \mu_j + \eta_j(t) + \gamma_j(v, t), \tag{3.1}$$

where μ_j is the deterministic mean signal assumed to be constant in time throughout. The process η_j is a common signal that is shared by all voxels in a given region, and represents the idea that signals within each region are, in some sense, homogeneous. Note that this is consistent with the characterization of connectivity by Van Den Heuvel and Pol (2010). The spatiotemporal process γ_j represents the voxel-specific variations from the regional signals that are correlated across both spatial (within region only) and temporal domains. By the decomposition in (3.1), it is clear that η_j contains the inter-regional connectivity information, while the random perturbations from γ_j can be considered as voxel-level spatiotemporal noise. Therefore, we propose covariance structures with η_j representing inter-regional while γ_j representing intra-regional connectivity.

First, we assume the collection of regional signals $\{\eta_j\}$ is uncorrelated with the voxel-specific fluctuations $\{\gamma_j\}$, $j = 1, \ldots, J$. Furthermore, the random fields γ_j and $\gamma_{j'}$ are uncorrelated whenever $j \neq j'$. The spatiotemporal covariance structures for η_j (inter-regional dependence) and γ_j (intra-regional dependence) are given by

1. $\mathrm{Cov}\,(\eta_j(t), \eta_{j'}(t')) = \rho_{jj'} A_\eta(t, t')$, where $\rho_{jj} = 1$, $\boldsymbol{R} = \{\rho_{jj'}\}_{j,j'=1}^{J}$ is a $J \times J$ positive-definite (p.d.) correlation matrix, and A_η is a p.d. covariance kernel on $[0, \mathcal{T}]^2$. $\rho_{jj'}$, $1 \leq j \neq j' \leq J$, are the primal parameters of interest that we use to quantify brain functional connectivity, which will be referred to as the *inter correlation* throughout.

2. $\mathrm{Cov}\,(\gamma_j(v, t), \gamma_j(v', t')) = B_j(t, t')C_j(v, v')$, where B_j and C_j are p.d. covariance and correlation kernels, respectively, on $[0, \mathcal{T}]^2$ and $\mathcal{R}_j \times \mathcal{R}_j$.

It is clear the latent BOLD signal model (3.1) can be viewed as spatial functional data, which is a rapidly developing subject of functional data analysis (FDA). In spatial FDA, classical problems of mean and covariance estimations are extensively studied. For example, see Hörmann and Kokoszka (2011) and Gromenko et al. (2012). The principled approaches to predict the spatial functional processes at unknown locations are nonparametric regression and functional kriging. Nonparametric regression mainly uses kernel type techniques to model the conditional mean of a process at an unknown location. Mixing conditions are often invoked to make convergence problems tractable (Dabo-Niang and Yao, 2007). Functional kriging is developed analogously to the kriging method of real-valued spatial processes, which mainly models the variogram to obtain an optimal linear predictor of the process at unknown locations (Giraldo et al., 2010, 2011). Functional kriging is best suited in the situations where the deterministic mean component of the functional spatial processes contain the most important information and the random component is mainly independent random errors (Delicado et al., 2010). These spatial FDA methods usually assume a compact domain of spatial functional data. Furthermore, the asymptotic regimes of spatial functional data, either the *infill domain* or *increasing domain* sampling schemes, assume increasing spatial sampling locations, which is not an appropriate assumption for our BOLD signals. Indeed, our setting is more relevant to spatiotemporal modeling (Christakos, 2000) which usually assumes that rich information is contained in data's spatiotemporal dependency and models the covariance structures of data processes directly. For example, Bel et al. (2008) used spatiotemporal kriging to align data atoms on irregular grids with the assumption of separable radial basis function (RBF) kernel based spatiotemporal covariance structures. Therefore, we will take the spatiotemporal modeling point of view and adopt a mixed-effects model approach, which is a primary tool in spatiotemporal modeling, to carry out our connectivity study.

3.2.2 Mixed-Effects Models

In practice, one observes the signals Y_j only at a discrete number of voxels $v_{jl} \in \mathcal{R}_j$, $l = 1, \ldots, L_j$, and at time points t_m, $m = 1, \ldots, M$, where we assume the latter to be equispaced with $t_1 = 0$ and $t_M = \mathcal{T}$. The observed BOLD signals are modeled as

$$X_{jlm} = \mu_j + \eta_{jm} + \gamma_{jlm} + \epsilon_{jlm}, \tag{3.2}$$

where $\eta_{jm} = \eta_j(t_m)$, $\gamma_{jlm} = \gamma_j(v_{jl}, t_m)$, and $\epsilon_{jlm} \sim \mathbf{N}(0, \sigma^2)$ are independently, identically distributed (i.i.d.) across all indices.

Model (3.2) constitutes a linear mixed-effects model, where μ_j is an overall signal level related to the neurophysiological behavior during rest, η_{jm} are random effects that encapsulate the long-range connectivity that is the target of brain functional connectivity studies, γ_{jlm} represent local variations in blood flow that induce correlated behavior between signals that are measured closely in time and space, and ϵ_{jlm} represent additional noise inherent to the scanner or other external sources.

To stabilize estimation of this mixed-effects model, we will assume that the random effects have joint Gaussian distributions with the following parameterized spatial and temporal covariance structures.

1. $\{\boldsymbol{A}\}_{mm'} = A_\eta(t_m, t_{m'}) = k_\eta \exp\{-\tau_\eta^2 \frac{(t_m - t_{m'})^2}{2}\} + \sigma_\eta^2 \delta_{mm'}$, where $\delta_{mm'}$ is the Kronecker delta and σ_η^2 represents the nugget effect that models the short-scale variability of the signals.

2. $\{\boldsymbol{B}_j\}_{mm'} = B_j(t_m, t_{m'}) = k_{\gamma_j} \exp\{-\tau_{\gamma_j}^2 \frac{(t_m - t_{m'})^2}{2}\}$.

3. $\{\boldsymbol{C}_j\}_{ll'} = C_j(v_{jl}, v_{jl'}) = \Big(1 + \phi_{\gamma_j}\sqrt{5}d + \frac{5}{3}\phi_{\gamma_j}^2 d^2\Big) \exp\{-\sqrt{5}\phi_{\gamma_j} d\}$, where $d = \|v_{jl} - v_{jl'}\|_2$.

We use the RBF and Matérn-5/2 kernels to model the temporal and spatial dependencies, respectively. Indeed, the choices for the spatiotemporal covariance structures $\boldsymbol{A}, \boldsymbol{B}_j$, and $\boldsymbol{C}_j$ could be potentially quite flexible provided that the model parameters can be consistently estimated. We will postpone the discussion of these conditions after the development of our model. We choose RBF and Matérn-5/2 as they are flexible and popular choices for spatiotemporal modeling that have been extensively investigated. For example, see (Stein, 1999; Flaxman et al., 2015).

With model (3.2), it is also possible to investigate our observation that the commonly used Pearson correlation estimator could be heavily biased due to the spatiotemporal noise. For a pair of regions, the common approach is to first average voxel level signals across space to obtain a mean signal for each region given by

$$\bar{X}_{jm} = L_j^{-1} \sum_{l=1}^{L_j} X_{jlm}. \tag{3.3}$$

Then, letting $\tilde{\mu}_j$ and $\tilde{\zeta}_j^2$ be the empirical mean and variance of $\bar{X}_{jm}$ across m (time), the usual Pearson correlation type estimator of connectivity is given by

$$\hat{\rho}_{jj'}^{\mathrm{CA}} = \frac{\sum_{m=1}^{M} \left(\bar{X}_{jm} - \tilde{\mu}_j\right) \left(\bar{X}_{j'm} - \tilde{\mu}_{j'}\right)}{\tilde{\zeta}_j \tilde{\zeta}_{j'}}. \tag{3.4}$$

We term the estimator in (3.4) the "correlation of averages" estimator as it averages signals by regions prior to the calculation of the Pearson correlation. In view of this estimator, one could get a consistent estimator for $\operatorname{Corr}\left(X_j(t), X_{j'}(t)\right)$ with mild conditions on the decaying rate of spatiotemporal dependency. However, in view of model (3.2),

$$\operatorname{Corr}\left(X_j(v,t), X_{j'}(w,t)\right) = \frac{\rho_{jj'} A(t,t)}{\sqrt{A(t,t) + B_j(t,t) + \sigma^2}\sqrt{A(t,t) + B_{j'}(t,t) + \sigma^2}}, \tag{3.5}$$

which indicates that $\hat{\rho}_{jj'}^{\mathrm{CA}}$ tend to underestimate the $\rho_{jj'}$ that we use to quantify the functional connectivity. The bias would disappear provided that both the measurement errors and spatiotemporal noise are correlated across regions with correlation strength of the same magnitude as those of the signals, which is a rather counter intuitive assumption to make. Indeed, it is easy to show that in general, the Pearson correlation can be corrupted by uncorrelated additive errors. Nevertheless, we note that when spatiotemporal dependencies are weak and region sizes are large, one should expect that $\hat{\rho}_{jj'}^{\text{CA}}$ provides an easy and reasonable estimator to quantify functional connectivity.

3.2.3 Model Estimation

An obvious approach to estimate the parameters of model (3.2) is to assume a Gaussian likelihood, which would arise if $\{\eta_j\}_{j=1}^J$ is a multivariate Gaussian process and the $\{\gamma_j\}_{j=1}^J$ are independent Gaussian spatiotemporal random fields. However, as Gaussian likelihood usually requires, with n being the sample size, $O(n^3)$ computing time and $O(n^2)$ memory, the large size of the data set usually makes a straight shot at likelihood evaluation prohibitive. Therefore, we will endeavor to break down the computation into smaller, simpler pieces.

In short, we will restrict our attention to pairs of regions so that each time, we will get one estimate of the inter correlation parameter for one pair of regions. More importantly, we propose a two-stage estimation approach. In the first step, one isolates data for each region in order to estimate the covariance parameters associated with the intra-regional spatiotemporal noise γ_j. In the second step, one isolates each pair of regions in order to estimate the remaining parameters, which include the inter correlation. One should bear in mind that the sole primary parameter of interest is the inter correlation matrix $\boldsymbol{R}$, whereas all others serve the secondary, but crucial roles, of improving estimation of $\boldsymbol{R}$ by adequately modeling the inherent spatiotemporal dependencies in the data.

Stage 1: Estimating Region-Specific Parameters

In the first step, one considers all data for each region $\mathcal{R}_j$ individually. A benefit of this approach is that all regional parameters

$$\boldsymbol{\theta}_j = \{\phi_{\gamma_j}, k_{\gamma_j}, \tau_{\gamma_j}\} \tag{3.6}$$

can be estimated in parallel. It is important to note that at this stage, all signals within a same region have a common component signal that is a single realization of η_j, which should be effectively treated as a fixed effect. Letting $\nu_{jm} = \mu_j + \eta_{jm}$, we define the *intra-regional* model of the observed BOLD signals by conditioning on η_j, yielding

$$X_{jlm} \mid \{\eta_{jm}\}_{m=1}^{M} = \nu_{jm} + \gamma_{jlm} + \epsilon_{jlm}. \tag{3.7}$$

Note that, since the spatiotemporal noise γ_j and measurement errors are assumed to be independent of η_j, their variance components are not affected by conditioning. Thus, the sole effect of conditioning on η_j is to treat it as a fixed effect in the first step of estimation.

Because the number of fixed effects now scales with the number time points M, we reduce the dimension of the problem by using a basis expansion. Specifically, suppose $\{\psi_k\}_{k=1}^{K}$ is a cubic B-spline basis on a given set of interior knots $\{s_u\}_{u=1}^{K-4}$. Then we approximate the fixed effects by $\nu_{jm} \approx \sum_{k=1}^{K} v_{jk}\psi_k(t_m)$. Let $\boldsymbol{v}_j = [v_{j1}, \ldots, v_{jK}]^\intercal$. An initial estimate of the fixed effect represented by $\boldsymbol{v}_j$ can be obtained by regressing B-spline basis on the sample pairs $\{t_m, X_{jlm}\}$, $l = 1, \ldots, M$, $m = 1, \ldots, M$, which yields $\hat{\boldsymbol{v}}_j$. This estimate can either be fixed throughout the minimization of the negative likelihood function, or simply be used as a starting point that is updated along with the variance components.

For $j = 1, \dots, J$, denote the observed signals for region $\mathcal{R}_j$ by

$$\boldsymbol{X}_j = [X_{j11}, X_{j12}, \dots, X_{j1M}, X_{j21}, \dots, X_{jL_jM}]^\intercal, \tag{3.8}$$

which is a column vector of stacked BOLD signals. Similarly, set

$$\boldsymbol{\eta}_j =[\eta_{j1}, \dots, \eta_{jM}]^\intercal, \tag{3.9}$$

$$\boldsymbol{\gamma}_j =[\gamma_{j11}, \gamma_{j12}, \dots, \gamma_{j1M}, \gamma_{j21}, \dots, \gamma_{jL_jM}]^\intercal, \text{ and} \tag{3.10}$$

$$\boldsymbol{\epsilon}_j =[\epsilon_{j11}, \epsilon_{j12}, \dots, \epsilon_{j1M}, \epsilon_{j21}, \dots, \epsilon_{jL_jM}]^\intercal. \tag{3.11}$$

Define the matrix

$$\tilde{\boldsymbol{G}} = \begin{bmatrix} \psi_1(t_1) & \psi_2(t_1) & \dots & \psi_K(t_1) \\ \psi_1(t_2) & \psi_2(t_2) & \dots & \psi_K(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(t_M) & \psi_2(t_M) & \dots & \psi_K(t_M) \end{bmatrix}, \tag{3.12}$$

and $\boldsymbol{G} = \mathbf{1}_{L_j} \otimes \tilde{\boldsymbol{G}}$, where $\otimes$ denotes the Kronecker product and $\mathbf{1}_{L_j}$ is the column vector of 1's with length L_j. Then the matrix form of (3.7) is

$$\boldsymbol{X}_j \mid \boldsymbol{\nu}_j = \boldsymbol{G}_j \boldsymbol{v}_j + \boldsymbol{\gamma}_j + \boldsymbol{\epsilon}_j. \tag{3.13}$$

Looking ahead, in order to estimate the regional parameters (3.6) in this model, we construct and solve for restricted maximum likelihood (ReML) problem, or equivalently, minimize the negative value of the restricted log likelihood function of $\boldsymbol{X}_j|\boldsymbol{\nu}_j$. The reason that we prefer ReML over the usual maximum likelihood (ML) is because ReML yields unbiased estimators for the covariance components by fitting likelihood based on a set of contrasts that eliminates the fixed effects (Harville, 1974; Jennrich and Schluchter, 1986; Lindstrom and Bates, 1988; Pinheiro and Bates, 2006). We note that it is feasible and reasonable to recover the fixed effects from ReML estimates. However, the caveat is that the ReML estimated fixed effects are not invariant under change of basis of data, which is a desired property possessed by the usual ML estimators (see Pinheiro and Bates, 2006, page 76). It is clear that the ReML function includes the variance of measurement errors σ^2, which enters the covariance matrix diagonally in an additive manner. If one factors out σ^2 from the entire covariance structure of $\boldsymbol{X}_j|\boldsymbol{\nu}_j$, σ^2 can be analytically represented by the other parts of the likelihood function, which makes the computation process faster and more accurate; see (3.17) and (3.25).

This implies that the model parameter estimates are invariant to, provided that one uses the analytical solution for σ^2, whether one estimates k_{γ_j}, σ^2, the variance of γ_j and $\epsilon_{j\cdot\cdot}$ explicitly, or estimates the variance ratio between signals and measurement errors, k_{γ_j}/σ^2, as a whole. We adopt the latter approach for improved computation efficiency. Indeed, one can observe that the estimation of the primal parameter only requires estimating the variance ratios of the model components. Therefore, let $\tilde{k}_{\gamma_j} = k_{\gamma_j}/\sigma^2$, $\{\tilde{\boldsymbol{B}}_j\}_{t_m t_{m'}} = \tilde{B}_j(t_m, t_{m'}) = \tilde{k}_{\gamma_j} \exp\{-\tau^2_{\gamma_j}(t_m - t_{m'})^2/2\}$, and $\boldsymbol{I}_L$ be the $L \times L$ identity matrix. The marginal model of (3.7) is

$$\boldsymbol{X}_j \sim \mathrm{N}(\boldsymbol{G}_j \boldsymbol{v}_j, \sigma^2 \boldsymbol{V}_j(\boldsymbol{\theta}_j)), \tag{3.14}$$

where

$$\boldsymbol{V}_j(\boldsymbol{\theta}_j) = \boldsymbol{C}_j \otimes \tilde{\boldsymbol{B}}_j + \boldsymbol{I}_{ML_j} \tag{3.15}$$

and the corresponding negative ReML function is (notational dependency on $\boldsymbol{\theta}_j$ is suppressed when there is no ambiguity of the context)

$$\begin{aligned}\mathcal{L}_j^{\text{ReML}}(\boldsymbol{v}_j, \boldsymbol{\theta}_j \mid \boldsymbol{X}_j, \boldsymbol{\eta}_j) =& \frac{1}{2}\log|\boldsymbol{V}_j| \\ &+ \frac{1}{2}\log\left|\boldsymbol{G}_j^{\intercal}\boldsymbol{V}_j^{-1}\boldsymbol{G}_j\right| \\ &+ \frac{1}{2}(ML_j - K)\log\left(\boldsymbol{X}_j - \boldsymbol{G}_j\boldsymbol{v}_j\right)^{\intercal}\boldsymbol{V}_j^{-1}\left(\boldsymbol{X}_j - \boldsymbol{G}_j\boldsymbol{v}_j\right),\end{aligned} \tag{3.16}$$

where we profiled

$$\sigma^2 = \frac{1}{L_jM - K}\left(\boldsymbol{X}_j - \boldsymbol{G}_j\boldsymbol{v}_j\right)^{\intercal}\boldsymbol{V}_j^{-1}\left(\boldsymbol{X}_j - \boldsymbol{G}_j\boldsymbol{v}_j\right). \tag{3.17}$$

Estimating the regional covariance components $\boldsymbol{\theta}_j$ by minimizing the negative ReML function in (3.16) is referred to as *Stage* 1 estimation in our two-stage approach. In particular, $\hat{\boldsymbol{\theta}}_j$ will be plugged into the full ReML function at *Stage* 2, where it can be further updated or held fixed to reduce computation cost. The estimated parameters for fixed effects $\hat{\boldsymbol{v}}_j$ obtained by B-Spline regression are used as the initial values of the optimization. In the case where one updates these estimates, the analytical generalized least square (GLS) estimator should be used rather than running the optimizer on $\boldsymbol{v}_j$ directly, that is, $\hat{\boldsymbol{v}}_j = (\boldsymbol{G}_j^{\intercal}\hat{\boldsymbol{V}}_j^{-1}\boldsymbol{G}_j)^{-1}\boldsymbol{G}_j^{\intercal}\hat{\boldsymbol{V}}_j^{-1}\boldsymbol{X}_j$.

A modified version of the correlation of averages estimator might also be considered at the end of Stage 1. Recall that one obtains the fixed-effects estimates at Stage 1

$$\hat{\nu}_{jm} = \sum_{k=1}^{K}\hat{v}_{jk}\psi_k(t_m),\ j = 1, \dots, L_j, m = 1, \dots, M. \tag{3.18}$$

Then, letting $\check{\mu}_j$ and $\check{\zeta}_j^2$ be the empirical mean and variance of $\hat{\nu}_{jm}$ across m (time), this modified estimator of $\rho_{jj'}$ is

$$\hat{\rho}^{\text{CAb}}_{jj'} = \frac{\sum_{m=1}^{M} (\hat{\nu}_{jm} - \check{\mu}_j)(\hat{\nu}_{j'm} - \check{\mu}_{j'})}{\check{\zeta}_j \check{\zeta}_{j'}}. \tag{3.19}$$

This estimator will be similar to the correlation of averages when the number of knots is large (close to M) and the estimate of $\boldsymbol{v}_j$ is computed only using the initial spline computation, without updating it in the likelihood maximization step.

Stage 2: Estimating Global and Inter-Regional Parameters

Having obtained estimates for the regional variance components in Stage 1, we advance to pairwise regional data. Without loss of generality, consider regions $\mathcal{R}_1$, $\mathcal{R}_2$. Let $\boldsymbol{X} = [\boldsymbol{X}_1^\intercal, \boldsymbol{X}_2^\intercal]^\intercal$, $\boldsymbol{\mu} = [\mu_1, \mu_2]^\intercal$, $\boldsymbol{\eta} = [\boldsymbol{\eta}_1^\intercal, \boldsymbol{\eta}_2^\intercal]^\intercal$, $\boldsymbol{\gamma}^\intercal = [\boldsymbol{\gamma}_1^\intercal, \boldsymbol{\gamma}_2]^\intercal$, $\boldsymbol{\alpha} = [\boldsymbol{\eta}^\intercal, \boldsymbol{\gamma}^\intercal]^\intercal$, and $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1^\intercal, \boldsymbol{\epsilon}_2^\intercal]^\intercal$. Furthermore, define the design matrices

$$\boldsymbol{Z} = \begin{bmatrix} \mathbf{1}_{ML_1} & \mathbf{0}_{ML_1} \\ \mathbf{0}_{ML_2} & \mathbf{1}_{ML_2} \end{bmatrix},$$

and $\boldsymbol{U} = [\boldsymbol{Z}, \boldsymbol{I}_{M(L_1+L_2)}]$, where $\mathbf{0}_L$ is the column vector of 0's of length L. Then the pairwise *inter-regional* model has the matrix form

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{\alpha} + \boldsymbol{\epsilon}. \tag{3.20}$$

Following the discussion under (3.7), define $\tilde{k}_\eta = k_\eta/\sigma^2$, $\tilde{\sigma}^2_\eta = \sigma^2_\eta/\sigma^2$, and $\{\tilde{\boldsymbol{A}}\}_{t_m t_{m'}} = \tilde{A}_\eta(t_m, t_{m'}) = \tilde{k}_\eta \exp\{-\tau^2_\eta (t_m - t_{m'})^2/2\} + \tilde{\sigma}^2 \delta_{mm'}$. Let $\boldsymbol{\theta}$ be the vector of parameters that characterizes the covariance structure of $\boldsymbol{X}$, which is a superset of the regional parameter vector $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, given by

$$\boldsymbol{\theta} = [\tau_\eta, k_\eta, \phi_{\gamma_1}, \phi_{\gamma_2}, \tilde{\tau}_{\gamma_1}, \tilde{k}_{\gamma_1}, \tilde{\tau}_{\gamma_2}, \tilde{k}_{\gamma_2}, \rho_{12}, \tilde{\sigma}^2_\eta]^\intercal. \tag{3.21}$$

Then, the corresponding marginal model is

$$\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{Z\mu}, \sigma^2 \boldsymbol{V}(\boldsymbol{\theta})), \tag{3.22}$$

where

$$\boldsymbol{V}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{C}_1 \otimes \tilde{\boldsymbol{B}}_1 + \boldsymbol{J}_{L_1,L_1} \otimes \tilde{\boldsymbol{A}} & \rho_{12} \boldsymbol{J}_{L_1,L_2} \otimes \tilde{\boldsymbol{A}} \\ \rho_{12} \boldsymbol{J}_{L_2,L_1} \otimes \tilde{\boldsymbol{A}} & \boldsymbol{C}_2 \otimes \tilde{\boldsymbol{B}}_2 + \boldsymbol{J}_{L_2,L_2} \otimes \tilde{\boldsymbol{A}} \end{bmatrix} + \boldsymbol{I}_{M(L_1+L_2)} \tag{3.23}$$

and $\boldsymbol{J}_{L,L'}$ is the $L \times L'$ matrix of 1's. To simplify the notation, we will reduce $\boldsymbol{V}(\boldsymbol{\theta})$ to $\boldsymbol{V}$. One can estimate the model parameters $\boldsymbol{\theta}$, $\boldsymbol{\mu}$ by minimizing the negative value of the ReML function for inter-regional model, given by

$$\begin{aligned} \mathcal{L}^{\mathrm{ReML}}(\boldsymbol{\mu}, \boldsymbol{\theta} \mid \boldsymbol{X}) = & \frac{1}{2} \log |\boldsymbol{V}| \\ & + \frac{1}{2} \log \left| \boldsymbol{Z}^{\intercal} \boldsymbol{V}^{-1} \boldsymbol{Z} \right| \\ & + \frac{1}{2} (M(L_1 + L_2) - 2) \log \left(\boldsymbol{X} - \boldsymbol{Z\mu} \right)^{\intercal} \boldsymbol{V}^{-1} \left(\boldsymbol{X} - \boldsymbol{Z\mu} \right). \end{aligned} \tag{3.24}$$

As aforementioned, the estimates from Stage 1 can either be fixed throughout the optimization iterations, or be used as initial values for the full optimization procedure and updated during the optimization. Similar to Stage 1, $\hat{\boldsymbol{\mu}}$ is obtained by GLS its estimator. In addition, we profiled

$$\sigma^2 = \frac{1}{M(L_1 + L_2) - 2} \left(\boldsymbol{X} - \boldsymbol{Z\mu} \right)^{\intercal} \boldsymbol{V}^{-1} \left(\boldsymbol{X} - \boldsymbol{Z\mu} \right). \tag{3.25}$$

We refer the estimation of $\boldsymbol{\theta}$ as *Stage* 2 of our two-stage approach, and denote the estimated primal parameter as $\hat{\rho}_{12}^{\mathrm{ReML}}$.

3.3 Large Sample Properties

Large sample properties of model parameters are critical to uncertainty quantification. The classical asymptotic distributions of ML estimators from i.i.d. data are well established under certain regularity conditions. For spatiotemporal data, with our model (3.22) being a special case of the more general class of spatial regression models, one essentially have a single observation from a multivariate distribution whose dimension grows along with sample size. Therefore, it is not immediately obvious that the estimated model parameters are consistent and asymptotically normal as one would generally expect in the i.i.d. case. To that end, Sweeting (1980) established general asymptotic normality of maximum likelihood estimators which requires increasing, convergent, and smooth information. Mardia and Marshall (1984) proposed analogous regularity conditions such that the ML estimators of spatial regression model parameters, which include those of regressors and covariance matrix of residuals, converge weakly to a Gaussian random vector. The asymptotic properties for the ReML estimators under the spatial regression setting was investigated by Cressie and Lahiri (1993) in which the main focus was on the parameters of the residual covariance structure. Cressie and Lahiri (1996) further proposed practically verifiable sufficient conditions for the asymptotic results of ReML estimators to hold under various common spatial regression settings. We will follow the discussions therein to propose the sufficient conditions to establish asymptotic distribution of $\hat{\rho}^{\text{ReML}}$ in our model.

Since we always consider the pairwise model (3.20), the sample size is $n = M(L_j + L_{j'})$

and the model parameters of interest is $\boldsymbol{\theta}$ with $\hat{\rho}^{\mathrm{ReML}}$ being the primal parameter. Despite the fact that we specify the covariance structure $\boldsymbol{V}(\boldsymbol{\theta})$ and its parameter vector $\boldsymbol{\theta}$ in section 3.2.1 and 3.2.2, we note that the sufficient conditions for the asymptotic normality depend on the decaying rate of dependency rather than the specific form the $\boldsymbol{V}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}$. That being said, it is possible to extend the results in this section to other instances of covariance matrices that adequately model the spatiotemporal signals of interest and the parameter vectors, say $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$, that characterizes such covariance structures. To that end, we will assume $\boldsymbol{\theta}$ is a p-vector throughout. Theorem 2.1 of Cressie and Lahiri (1996) proposed three sufficient conditions for the ReML estimators to be asymptotically normal, one of which is replaced by their Theorem 2.2 which imposes bounds on the spectrum of the covariance structure. This condition is further replace by their Theorem 3.1 by assuming the data process is on a rectangular lattice and is covariance stationary. We note that while our model shares similarities with the case discussed in their Theorem 3.1, there are striking differences between the two. First of all, our data is collected over regularly spaced time domain, and the spatial domain consists of subsets that are enclosed within a 3D rectangular lattice. Depending on the choice of pairs of regions, such spatial domain may not be contiguous. More importantly, it is clear that our model covariance $\boldsymbol{V}(\boldsymbol{\theta})$ is only stationary within regional blocks but nonstationary across blocks. However, we note that Theorem 2.2 strikes a good balance between generality and easy verification. Therefore, we will proceed with stating Theorem 2.1 and 2.2 of Cressie and Lahiri (1996) in our context, for the sake of completeness, as Proposition 3.3.1 and 3.3.2 and then verify that our model satisfies these sufficient conditions.

Let $\boldsymbol{V}_i(\boldsymbol{\theta}) = \partial \boldsymbol{V}(\boldsymbol{\theta})/\partial\theta_i$, where θ_i is the ith element of $\boldsymbol{\theta}$, $i = 1, \ldots, p$. Also define $\boldsymbol{\Pi}(\boldsymbol{\theta}) = \boldsymbol{V}^{-1}(\boldsymbol{\theta}) - \boldsymbol{V}^{-1}(\boldsymbol{\theta})\boldsymbol{Z}\left(\boldsymbol{Z}^\intercal\boldsymbol{V}^{-1}(\boldsymbol{\theta})\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^\intercal\boldsymbol{V}^{-1}(\boldsymbol{\theta})$, which is $\boldsymbol{V}^{-1}(\boldsymbol{\theta})$ projected onto the complementary column space of the design matrix $\boldsymbol{Z}$. Given the nature of the ReML estimation, it is not surprising to see this projection. Let $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ be the $p \times p$ matrix of second-order partial derivatives of the negative ReML function, that is, $\{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\}_{ij} = \partial^2\mathcal{L}^{\text{ReML}}(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j, 1 \leq i, j \leq p$. It is straightforward to show that the ijth element of Fisher information matrix is given by

$$\mathbb{E}_{\boldsymbol{\theta}}(\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}))_{ij} = \frac{1}{2}\operatorname{Tr}\left\{\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{V}_i(\boldsymbol{\theta})\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{V}_j(\boldsymbol{\theta})\right\}, \quad 1 \leq i, j \leq p. \tag{3.26}$$

Further define $p \times p$ matrices of parameter vectors $\boldsymbol{\theta}_0^{\text{M}} = (\theta_1^0, \ldots, \theta_p^0)$, $\theta_i^0 \in \boldsymbol{\theta}$, $i = 1, \ldots, p$. We write $\{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0^{\text{M}})\}_{ij} = \partial^2\mathcal{L}^{\text{ReML}}(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j \mid_{\boldsymbol{\theta}=\theta_i^0}$, that is, the matrix of the second-order partial derivatives of $\mathcal{L}^{\text{ReML}}$ where each row is evaluated at the corresponding row of $\boldsymbol{\theta}_0^{\text{M}\intercal}$. Denote the Frobenius norm of a matrix by $\|E\|$ and the spectral norm by $\|E\|_s$.

Proposition 3.3.1. *Assume that*

A.1 $\boldsymbol{V}(\boldsymbol{\theta})$ *is twice continuously differentiable on* $\boldsymbol{\Theta}$*.*

A.2 for any $\boldsymbol{\theta}$*, there exists nonrandom, positive-definite (p.d.)* $p \times p$ *matrices* $\boldsymbol{W}(\boldsymbol{\theta})$ *and a sequence of matrices* $\{\boldsymbol{E}_n(\boldsymbol{\theta}), n \geq 1\}$*, continuous in* $\boldsymbol{\theta}$*, such that* $\|\boldsymbol{E}_n^{-1}(\boldsymbol{\theta})\| \to 0$ *and* $\boldsymbol{E}_n^{-1}(\boldsymbol{\theta})\boldsymbol{\mathcal{I}}_n(\boldsymbol{\theta})(\boldsymbol{E}_n^{-1}(\boldsymbol{\theta}))^\intercal \rightsquigarrow \boldsymbol{W}(\boldsymbol{\theta})$*, both uniformly over any compact subset* $\boldsymbol{F} \subset \boldsymbol{\Theta}$*.*

A.3 for all $c > 0$*,* $\xi > 0$*,*

(i) $\sup\{\|\boldsymbol{E}_n^{-1}(\boldsymbol{\theta})\boldsymbol{E}_n^{-1}(\boldsymbol{\theta}^0) - \boldsymbol{I}_p\| : \left\|(\boldsymbol{E}_n^{-1}(\boldsymbol{\theta}))^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}^0)\right\| \leq c; \boldsymbol{\theta}, \boldsymbol{\theta}^0 \in \boldsymbol{F}\} \to 0$ *uniformly over any compact subset* $\boldsymbol{F} \subset \boldsymbol{\Theta}$ *and*

(ii) for any fixed matrix $\boldsymbol{\theta}_0^M$*, let* $\boldsymbol{D}(\boldsymbol{\theta}_0^M, \boldsymbol{\theta}) = \big(\boldsymbol{\mathcal{I}}_n(\boldsymbol{\theta}_0^M) - \boldsymbol{\mathcal{I}}_n(\boldsymbol{\theta})\big)$,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}}(\, &\sup\{\left\|\boldsymbol{E}_n^{-1}(\boldsymbol{\theta})\boldsymbol{D}(\boldsymbol{\theta}_0^M, \boldsymbol{\theta})(\boldsymbol{E}_n^{-1}(\boldsymbol{\theta}))^\intercal\right\| : \\ &\left\|(\boldsymbol{E}_n^{-1}(\boldsymbol{\theta}))^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}_i^0)\right\| \leq c,\, 1 \leq i \leq p; \boldsymbol{\theta}, \boldsymbol{\theta}^0 \in \boldsymbol{F}\} > \xi) \to 0,\end{aligned}$$

uniformly over any compact subset $\boldsymbol{F} \subset \boldsymbol{\Theta}$*.*

Then,

$$(\boldsymbol{E}_n(\boldsymbol{\theta}))^\intercal(\hat{\boldsymbol{\theta}}^{ReML} - \boldsymbol{\theta}) \rightsquigarrow \mathbf{N}(0, \boldsymbol{W}^{-1}(\boldsymbol{\theta})). \tag{3.27}$$

As observed in Cressie and Lahiri (1993, 1996), the main challenge in verifying these conditions is A.2. Condition A.1 is easily verified for a specific covariance structure. Along with a properly chosen normalizing sequence $\{\boldsymbol{E}_n(\boldsymbol{\theta})\}$ that is sufficiently smooth, A.3 is also easily verified. Indeed, Cressie and Lahiri (1996) suggested a normalizing sequence

$$\boldsymbol{E}_n(\boldsymbol{\theta}) = \operatorname{diag}\{\|\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{V}_1(\boldsymbol{\theta})\|, \ldots, \|\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{V}_p(\boldsymbol{\theta})\|\} \tag{3.28}$$

and proposed a set of sufficient conditions in Theorem 2.2 for A.2 to hold. It is straight forward to see that if A.2 holds, then by (3.23) and (3.28), A.1, A.3 also hold. Thus, we will verify that their Theorem 2.2 holds for our model.

Denoted by $\mathcal{S} \subset \mathbb{R}^d$ the lattice of data sites. Let $\mathcal{R} = \{l_1, l_2, l_3\} \subset \mathbb{Z}_+^3$ and $T = \{t_m\} \subset \mathbb{Z}_+$ We work under the assumption that our data is collected over $\mathcal{S} = \mathcal{R} \times T$. The asymptotic regime we consider is increasing time domain and fixed spatial domain, that is, $|T| \to \infty$, which is a reasonable assumption for BOLD signals. Furthermore, we introduce the normalized information matrix $\boldsymbol{Q}_n(\boldsymbol{\theta})$ whose ijth element is given by

$$\{\boldsymbol{Q}_n(\boldsymbol{\theta})\}_{ij} = \frac{\operatorname{Tr}(\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{V}_i(\boldsymbol{\theta})\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{V}_j(\boldsymbol{\theta}))}{\|\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{V}_i(\boldsymbol{\theta})\| \, \|\boldsymbol{\Pi}(\boldsymbol{\theta})\boldsymbol{V}_j(\boldsymbol{\theta})\|}, \quad 1 \leq i, j \leq p. \tag{3.29}$$

Also, let $|\lambda_{1n}| \leq \cdots \leq |\lambda_{nn}|$, $|\lambda_{1n}^i| \leq \cdots \leq |\lambda_{nn}^i|$, and $\left|\lambda_{1n}^{ij}\right| \leq \cdots \leq |\lambda_{nn}^{ij}|$ be the ordered absolute values of eigenvalues of $\boldsymbol{V}(\boldsymbol{\theta}), \boldsymbol{V}_i(\boldsymbol{\theta})$, and $\boldsymbol{V}_{ij}(\boldsymbol{\theta})$, $1 \leq i, j \leq p$, respectively. In addition, define the a sequence $\{r_n, n \geq 1\}$ such that $\limsup_{n\to\infty} r_n/n \leq 1 - \delta$, for some $\delta \in (0, 1)$. To simplify the notation, We will suppress the dependence on $\boldsymbol{\theta}$ and n when there is no ambiguity of context. The sufficient condition for Assumption A.2 is

Proposition 3.3.2. *Assume that Assumptions A.*1 *and A.*3 *hold and that there exists a p.d. matrix* $\boldsymbol{W}(\boldsymbol{\theta})$*, continuous in* $\boldsymbol{\theta}$*, such that* $\boldsymbol{Q}_n(\boldsymbol{\theta}) \rightsquigarrow \boldsymbol{W}(\boldsymbol{\theta})$ *uniformly. Furthermore, for any compact subset* $\boldsymbol{F} \subset \boldsymbol{\Theta}$*, suppose there exist constants* $0 < h(\boldsymbol{F}) < \infty$ *and* $0 < g(\boldsymbol{F})$ *such that*

$$\limsup_{n\to\infty} \max\{|\lambda_n|, \left|\lambda_n^i\right|, \left|\lambda_n^{ij}\right| : 1 \leq i, j \leq p\} < h(\boldsymbol{F}) \tag{3.30}$$

and

$$\liminf_{n\to\infty} \min\{|\lambda_1|, \left|\lambda_{r_n}^i\right| : 1 \leq i \leq p\} > g(\boldsymbol{F}), \tag{3.31}$$

uniformly in $\boldsymbol{F}$*. Then*

$$\{\mathbb{E}\left[\boldsymbol{\mathcal{I}}_n(\boldsymbol{\theta})\right]\}^{1/2} (\hat{\boldsymbol{\theta}}^{ReML} - \boldsymbol{\theta}) \rightsquigarrow \mathbf{N}(\mathbf{0}, \boldsymbol{I}_p). \tag{3.32}$$

Next, we state two lemmas that will be used to verify that our model satisfies Proposition 3.3.2. The proofs of these two lemmas are omitted as they follow standard matrix algebra.

Lemma 3.3.3. *Let* $\boldsymbol{\mathcal{D}}$ *be a nonsingular matrix and* $\boldsymbol{\mathcal{C}}$ *be a square matrix, both of size*

$n \times n$ such that

$$\|\boldsymbol{\mathcal{D}} - \boldsymbol{\mathcal{C}}\|_s \left\|\boldsymbol{\mathcal{D}}^{-1}\right\|_s < 1.$$

Then, $\boldsymbol{\mathcal{C}}^{-1}$ exists and

$$\left\|\boldsymbol{\mathcal{C}}^{-1}\right\|_s \leq \frac{\left\|\boldsymbol{\mathcal{D}}^{-1}\right\|_s}{1 - \left\|\boldsymbol{I}_n - \boldsymbol{\mathcal{D}}^{-1}\boldsymbol{\mathcal{C}}\right\|_s}.$$

Lemma 3.3.4. *Let $\boldsymbol{\mathcal{D}}$ be an $n \times n$ matrix with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$. Then, for any $0 \leq r \leq n-2$, and any $1 \leq i_1 < \cdots < i_r \leq n$,*

$$\lambda_{r+1} \geq \min\{\lambda' : (\boldsymbol{\mathcal{D}}_{i_1,\ldots,i_r} - \lambda' \boldsymbol{I}) = \boldsymbol{0}\},$$

where $\boldsymbol{\mathcal{D}}_{i_1,\ldots,i_r}$ is the matrix obtained by deleting the i_1th, $\ldots, i_r$th columns and rows of $\boldsymbol{\mathcal{D}}$.

Theorem 3.3.5. *The model* (3.22) *satisfies Propositions 3.3.2. Therefore,*

$$\{\mathbb{E}\left[\boldsymbol{\mathcal{I}}_n(\boldsymbol{\theta})\right]\}^{1/2}\,(\hat{\boldsymbol{\theta}}^{ReML} - \boldsymbol{\theta}) \rightsquigarrow \mathbf{N}(\boldsymbol{0}, \boldsymbol{I}_p). \tag{3.33}$$

Proof. Denoted by ${\boldsymbol{V}_{(\cdot)}}^{uv}$ the uvth element of $\boldsymbol{V}_{(\cdot)} = \boldsymbol{V}, \boldsymbol{V}_i, \boldsymbol{V}_{ij}$, $1 \leq i, j \leq p$. Let $\lambda_n^{(\cdot)}$ be the largest eigenvalues of $\boldsymbol{V}_{(\cdot)}$. Then, by Gershgorin circle theorem, it is straight forward that

$$|\lambda_n^{(\cdot)}| \leq \max_{1 \leq u \leq n} \left\{ \sum_{v=1}^{n} {\boldsymbol{V}_{(\cdot)}}^{uv} \right\} \leq h(K). \tag{3.34}$$

The existence of such bound $h(K)$ is guaranteed due to a) the exponentially decaying $\tilde{\boldsymbol{B}}_1, \tilde{\boldsymbol{B}}_2$, and $\tilde{\boldsymbol{A}}$ along the temporal domain and b) the fixed spatial grids, hence the fixed $\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{J}_{L_1,L_2}$, and $\boldsymbol{J}_{L_2,L_1}$ in (3.23). The uniform property comes from the smoothness of $\boldsymbol{V}_{(\cdot)}$ in $\boldsymbol{\theta}$ and the fact that $\boldsymbol{F}$ is compact.

To verify the lower bounds in (3.31), we first check $|\lambda_1|$. Let $\boldsymbol{\mathcal{D}} = \text{diag}\{\boldsymbol{V}\}$. Then,

$$\|\boldsymbol{\mathcal{D}} - \boldsymbol{V}\|_s \leq \sum_{1 \leq u \neq v \leq n} |\boldsymbol{V}^{uv}| \leq g_1(\boldsymbol{F}) \max_{1 \leq u \leq n} \{\boldsymbol{V}^{uu}\}, \tag{3.35}$$

where the existence of $g_1(\boldsymbol{F})$ is guaranteed by the exponential decaying elements of $\boldsymbol{V}$. By Lemma 3.3.3,

$$\lambda_1 \geq \frac{1 - g_1(\boldsymbol{F})}{\max_{1 \leq u \leq n}\{\boldsymbol{V}^{uu}\}}. \tag{3.36}$$

Next, we proceed to obtain the lower bound for $\lambda^i_{r_n}$. Recall that $\boldsymbol{\mathcal{S}} = \{(l_1, l_2, l_3, t_m)\}$, where l_i, $i = 1, 2, 3$, encode signals' spatial locations, whose size are considered to be fixed. $t_m = 1, 2, \ldots$ are the regularly spaced measurement points in time, and is assumed to grow to infinity. Denoted by $\tilde{\boldsymbol{\mathcal{S}}} = \boldsymbol{\mathcal{S}} \setminus \{t_m = 2\mathbb{N}+1\}$, the index set obtained by deleting the odd number time points. Define $r_n = n - |\tilde{\boldsymbol{\mathcal{S}}}|$, then

$$\limsup_{n \to \infty} \frac{r_n}{n} \leq \frac{1}{2}.$$

Next, let $\tilde{\boldsymbol{V}}_i$ be the matrices obtained by deleting the elements of $\boldsymbol{V}_i$ that correspond to even time points. Now, set $\boldsymbol{\mathcal{D}} = \text{diag}\left(\tilde{\boldsymbol{V}}_i^2\right)\boldsymbol{I}$ and $\boldsymbol{\mathcal{C}} = \tilde{\boldsymbol{V}}_i^2$, $1 \leq i \leq p$. Then,

$$\|\boldsymbol{\mathcal{D}} - \boldsymbol{\mathcal{C}}\|_s \leq \sum_{1 \leq u \neq v \leq n} \left|(\tilde{\boldsymbol{V}}_i^2)^{uv}\right| \leq g_2(\boldsymbol{F}) \max_{1 \leq u \leq n} \{(\tilde{\boldsymbol{V}}_i^2)^{uu}\}, 1 \leq i \leq p. \tag{3.37}$$

By Lemma 3.3.3 and 3.3.4, one obtains

$$|\lambda^i_{r_n}|^2 \geq \frac{1 - g_2(\boldsymbol{F})}{\max_{1 \leq u \leq n}\{(\tilde{\boldsymbol{V}}_i^2)^{uu}\}}. \tag{3.38}$$

The verification is complete by choosing $g = \max\{g_1, g_2\}$. ■

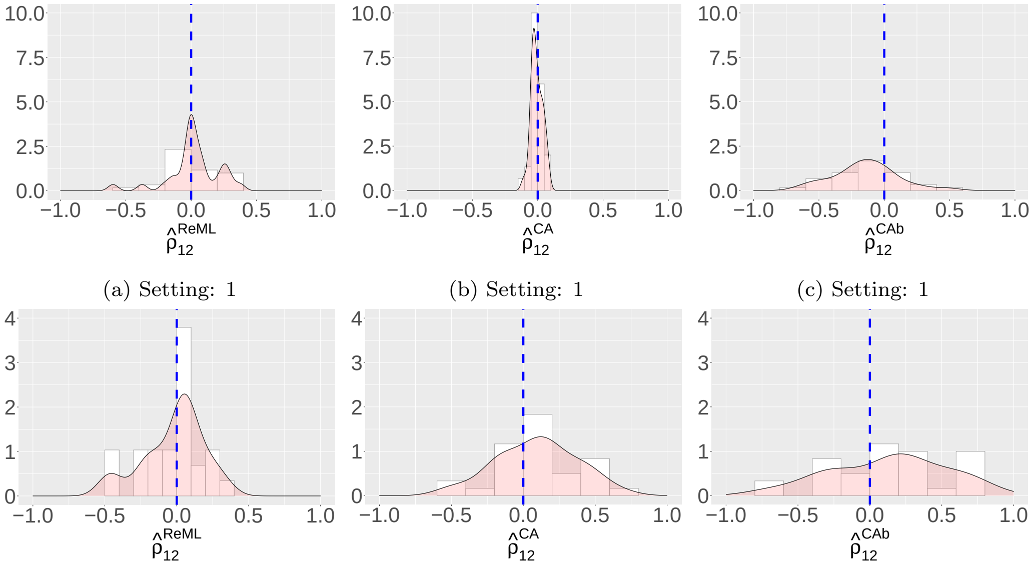
3.4 Simulation Study

To understand the performance of our model, we proceed with a simulation study. BOLD signals are simulated by an $M(L_1 + L_2)$ Gaussian vector according to (3.20) with model parameters (3.21). Various values of $\boldsymbol{\theta}$ are used to simulate different settings. For each setting, we ran $Q = 30$ simulations and across all simulation settings, we set $L_1 = L_2 = 30$ and $M = 50$. One of the main goals is to study the performance of our estimator $\hat{\rho}_{12}^{\text{ReML}}$ under different settings of signal strength relative to spatiotemporal noise, which is mainly quantified by the values of $\tilde{k}_\eta$ relative to those of $\tilde{k}_{\gamma_1}$ and $\tilde{k}_{\gamma_2}$. Therefore, we fix $\sigma^2 = 1$, $\tau_\eta = 1/25$, and $\tilde{\sigma}_\eta^2 = 0.1$ across all simulations. Table 3.1 summarizes the simulation settings. We compare three estimators, $\hat{\rho}_{12}^{\text{ReML}}$, $\hat{\rho}_{12}^{\text{CA}}$, and $\hat{\rho}_{12}^{\text{CAb}}$, which are the ReML estimator of ρ_{12}, the conventional Pearson correlation of averages, and the correlation of averages on B-spline fitted signals, respectively. The calculation of the two correlation of averages type estimators are detailed in (3.19) and (3.4). The empirical distributions of these estimators are plotted in Figures 3.1 and 3.2 and the RMSE are presented in Table 3.2.

Settings	$\tilde{k}_\eta$	ϕ_1	ϕ_2	τ_{γ_1}	$\tilde{k}_{\gamma_1}$	τ_{γ_2}	$\tilde{k}_{\gamma_2}$	ρ_{12}
1	0	1	$\sqrt{2/3}$	$\sqrt{2}/3$	2	$1/\sqrt{2}$	4	0
2	1	$1/\sqrt{2}$	$\sqrt{2}/3$	$\sqrt{2}/3$	2	$1/\sqrt{2}$	4	0
3	1	1	$\sqrt{2/3}$	$\sqrt{2}/3$	2	$1/\sqrt{2}$	4	0.5
4	0.1	$1/\sqrt{2}$	$\sqrt{2}/3$	$\sqrt{2}/6$	0.2	$\sqrt{2}/7$	0.4	0.5
5	10	$1/\sqrt{2}$	$\sqrt{2}/3$	$\sqrt{2}/6$	2	$\sqrt{2}/7$	4	0.5
6	0.1	$1/\sqrt{2}$	$\sqrt{2}/3$	1	0.2	$\sqrt{2/3}$	0.4	0.5

Table 3.1: Simulation settings.

Settings 1 and 2 represent the case where a pair of regions are not connected. In



(a) Setting: 1 (b) Setting: 1 (c) Setting: 1

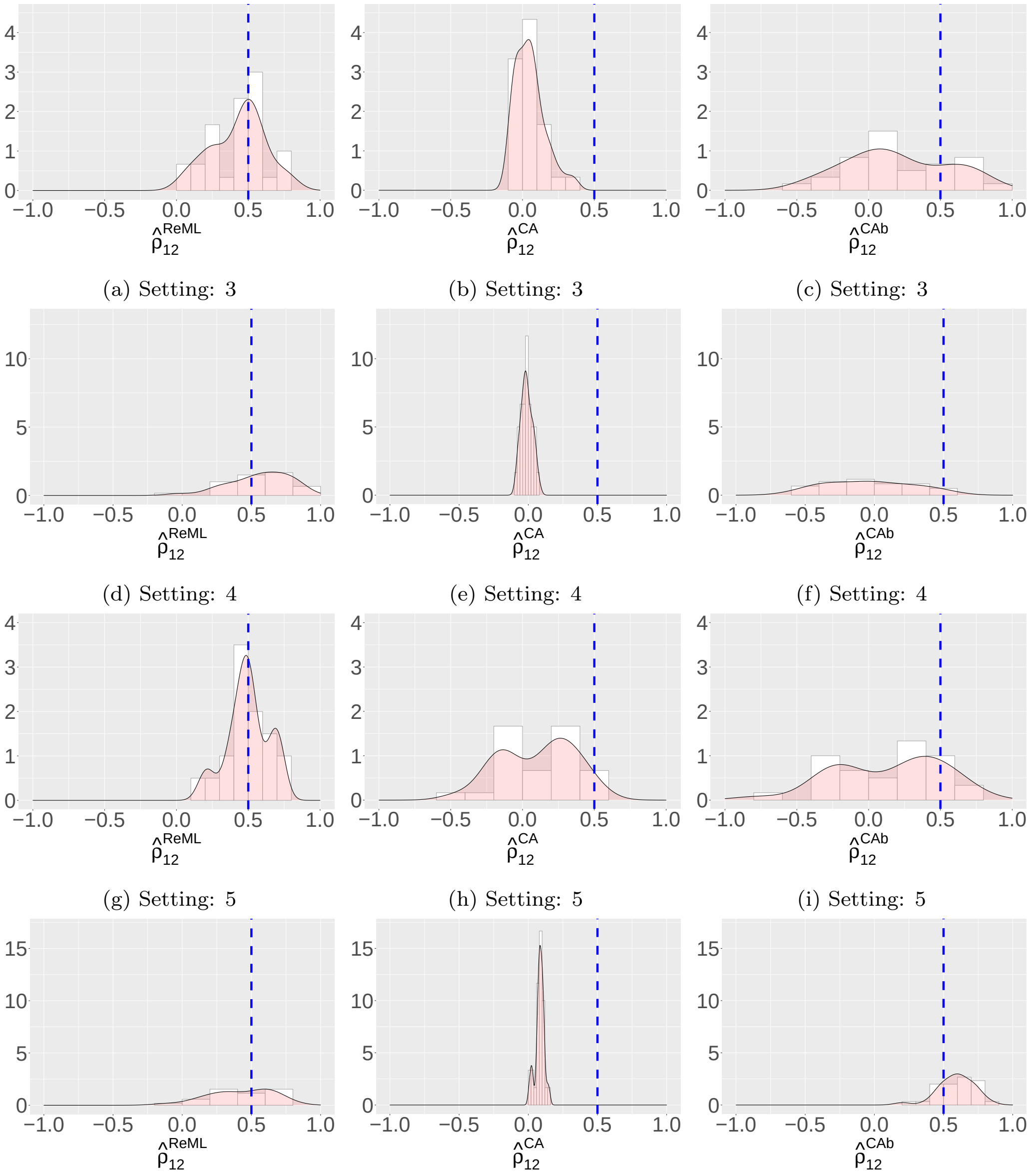
(d) Setting: 2 (e) Setting: 2 (f) Setting: 2

Figure 3.1: Simulation settings 1 and 2: disconnected regions.

Settings	$\hat{\rho}_{12}^{\text{ReML}}$	$\hat{\rho}_{12}^{\text{CA}}$	$\hat{\rho}_{12}^{\text{CAb}}$
1	0.1982	**0.0020**	0.0785
2	0.2139	**0.0843**	0.1691
3	**0.1987**	0.2104	0.2116
4	**0.2240**	0.2626	0.3855
5	**0.1457**	0.2248	0.2689
6	0.2400	0.1750	**0.0246**

Table 3.2: RMSE of $\hat{\rho}_{12}^{\text{ReML}}$, $\hat{\rho}_{12}^{\text{CA}}$, and $\hat{\rho}_{12}^{\text{CAb}}$.

particular, setting 1 is emulating the case of a dead brain, as the variation of signals $\tilde{k}_\eta = 0$. We can see from (a) - (c) in Figure 3.1 that in setting 1, all estimators are concentrated around the true parameter value 0, with $\hat{\rho}_{12}^{\text{CA}}$ having the smallest RMSE. $\hat{\rho}_{12}^{\text{ReML}}$ has better concentration around zero than $\hat{\rho}_{12}^{\text{CAb}}$ which is slightly more biased toward the negative values. Setting 2 aims to simulate two disconnected regions in a live brain. In this setting, the simulation results tell a similar story as in setting 1. Overall, in these two settings of disconnected regions, all three estimators perform reasonably



(a) Setting: 3

(b) Setting: 3

(c) Setting: 3

(d) Setting: 4

(e) Setting: 4

(f) Setting: 4

(g) Setting: 5

(h) Setting: 5

(i) Setting: 5

(j) Setting: 6

(k) Setting: 6

(l) Setting: 6

Figure 3.2: Simulation 3 to 6: connected regions.

well with $\hat{\rho}_{12}^{\text{CA}}$ having the lowest RMSE. One should not be surprised by this result as we have seen that this estimator tend to be biased towards 0 with the presence of large spatiotemporal noise. The rest 4 settings represent the cases where regions are actively connected but with different levels of relative signal strength. In particular, setting 3 has stronger signal and weaker spatiotemporal noise compared to setting 4 in the sense that the spatiotemporal noise decays slower in setting 4. While $\hat{\rho}_{12}^{\text{ReML}}$ performs best in both cases, we can clearly see the increase in RMSE in setting 4. In Figure 3.2 (a) and (d), one can see the increased spread of $\hat{\rho}_{12}^{\text{ReML}}$ due to higher spatiotemporal noise. The interesting observation is that $\hat{\rho}_{12}^{\text{CA}}$ is heavily biased towards 0 in these two cases, which verifies our claim. $\hat{\rho}_{12}^{\text{CAb}}$ is less biased than $\hat{\rho}_{12}^{\text{CA}}$ but suffers from large spread. In setting 5 where the signal strength dominates the spatiotemporal noise, $\hat{\rho}_{12}^{\text{ReML}}$ gains further advantage in terms of RMSE. Setting 6 has the same setting as setting 4 except that the noise decays faster on the time domain. We can see $\hat{\rho}_{12}^{\text{ReML}}$ is relatively robust with faster tapering temporal dependence in the noise. However, $\hat{\rho}_{12}^{\text{CAb}}$ performs surprisingly well in this case, which suggests that $\hat{\rho}_{12}^{\text{CAb}}$ could be a decent estimator when noise decays fast enough along the temporal domain. This is intuitive as correlation of average type of estimators is ideal for i.i.d. data. Overall, we can observe that $\hat{\rho}_{12}^{\text{ReML}}$ is the most robust estimator across all different simulation settings.

3.5 Data Application

With the robust empirical performance of our estimator $\hat{\rho}^{\mathrm{ReML}}$ observed in the simulation studies, we proceed to real data application. Our model is applied to one dead and one anesthetized live long Evans rat. The data set we use is the same as the one studied in Becq et al. (2020). In particular, the live rat we work on is labeled as rat 20160616_145220 whose BOLD signals are collected with anesthetic (Etomidate) administrated. The dead rat we use is labeled as rat 0160524_153000. For both rats, the BOLD signals are collected at the frequency of 2 Hz for the duration of 30 minutes. Preprocessing steps include motion removal and wavelet transformation of the original signals. Becq et al. (2020) found that the level 4 wavelet coefficients tend to decorrelate brain connectivity from systematic variable such as heart rat, body temperature, and so on, which were known factors to perturb functional connectivity. We will adopt this procedure and apply our model to the level 4 wavelet coefficients. For more details on these preprocessing steps, we refer readers to Becq et al. (2020).

We focus on 21 out of 51 regions in rat brains. First, we obtain $\hat{\rho}_{jj'}^{\mathrm{ReML}}$ for $1 \leq j < l' \leq 21$, then using the asymptotic distribution derived in Theorem 3.3.5 to calculate the z-scores of these estimators. We further adopt the Benjamini–Yekutieli (BY) procedure (Benjamini and Yekutieli, 2001) to test for significant pairs of regions. The BY procedure is used to control the false discovery rate (FDR) under arbitrary dependence assumptions, which is suitable for our use case. Furthermore, since we are expecting to make discoveries for connected brain regions, controlling FDR means we do not penalize single false discovery as harsh as controlling the family-wise error rate, which makes it less conservative and more reasonable for our purpose.

Looking at the network for the dead rat in Figure 3.3, we discover zero connected regions, which is the expected result for dead rats. Furthermore, by comparing the result in Figure 3.4 and 3.5 with the network of the same rat in Becq et al. (2020). We found that our result is mostly consistent with the previous studies. For example, regions such as ACC, M1, M2, AU are known to be highly connected. In our result, we also observed a high connectivity level for these regions. On the other hand, for Ent, RSC, these commonly unconnected regions, a low level of connectivity is also observed in our result. However, our result does show more discoveries compared to the results in the previous study. For example, Apir. This is not fully unexpected since the work done by Becq et al. (2020) is based on Pearson correlation of averages, which we have shown to be biased towards 0 in the presence of spatiotemporal noises. We also constructed node degree (ND) and functional connectivity strength (FCS) plots by regions in Figure 3.6 to further summarize the recovered network. The minor degree of asymmetry in both ND and FCS metrics between left and right regions are around the similar scale as those observed in Becq et al. (2020).

Furthermore, despite that our two-stage approach effectively reduces the dimension of parameter space, likelihood evaluation remains expensive due to the large dimensionality of our covariance matrix. To speed up the computation, we utilize parallel computing to run estimations simultaneously. Also, we delegate level 3 Basic Linear Algebra Subpro-

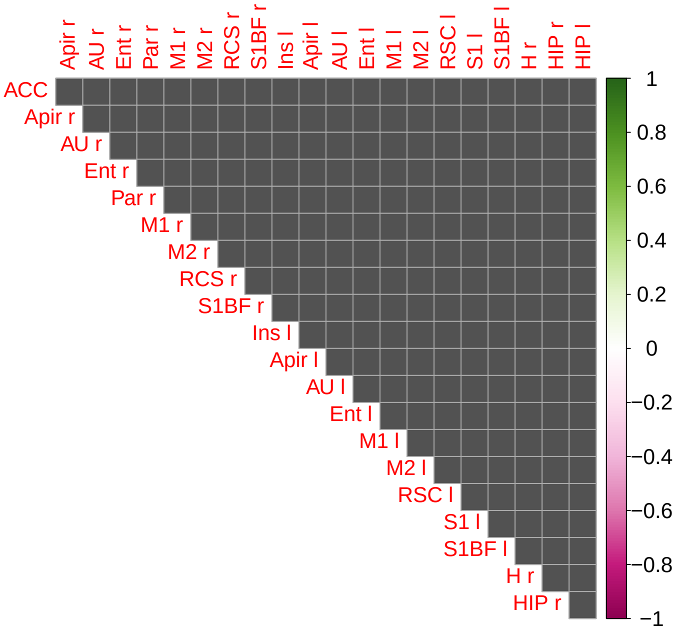


Figure 3.3: Brain network of dead rat 20160524_153000.

grams (BLAS) matrix operations to GPU using NVIDIA's CUDA framework to accelerate matrix calculations. These operations are complex enough, for example, matrix-matrix product, solving triangular matrix equations, so that the speed gain of using GPU outshines the overhead of transferring data between CPU and GPU. To demonstrate this point, we benchmarked matrix inversion and log determinant calculations, which are commonly encountered in log likelihood evaluation, using CPU and GPU for $n \times n$ matrices, $n = 1000, 2000, \ldots, 8000$. Figure 3.7 shows a decisive advantage of GPU calculation when matrix size grows over a certain level (4000×4000 in our experiment). In our data application, the covariance matrix could easily go beyond 10000 rows (columns) so using GPU acceleration could greatly improve computation efficiency.

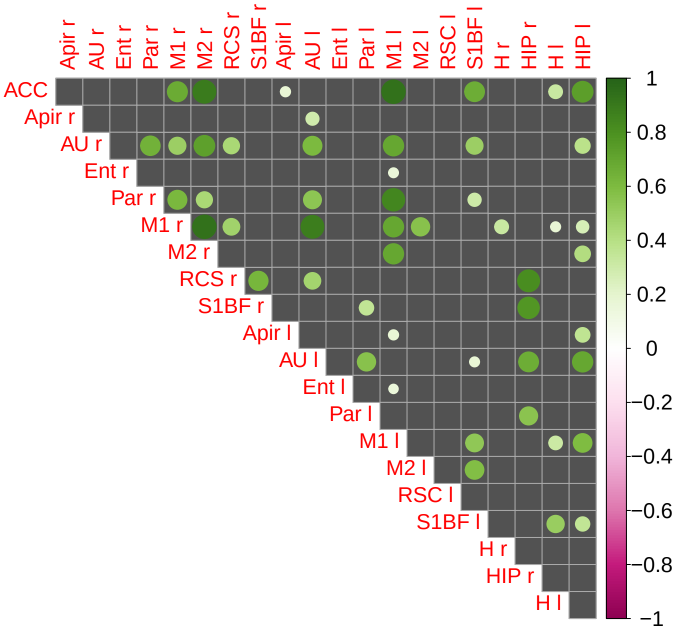


Figure 3.4: Brain network of anesthetized rat 20160616_145220, Eto-L.

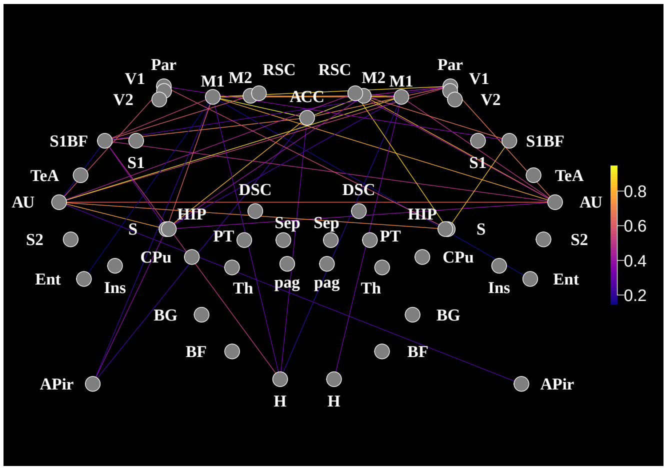


Figure 3.5: Brain network graph of anesthetized rat 20160616_145220, Eto-L.

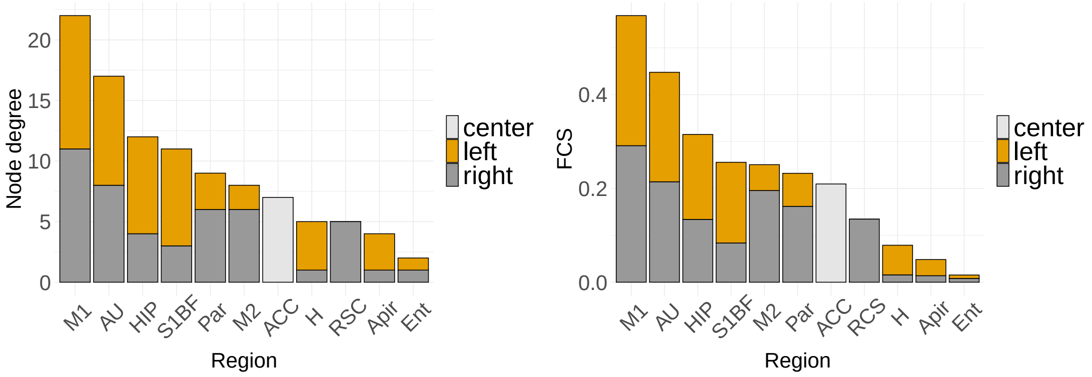


Figure 3.6: Node degree (a) and FCS (b) of anesthetized rat 20160616_145220, Eto-L.

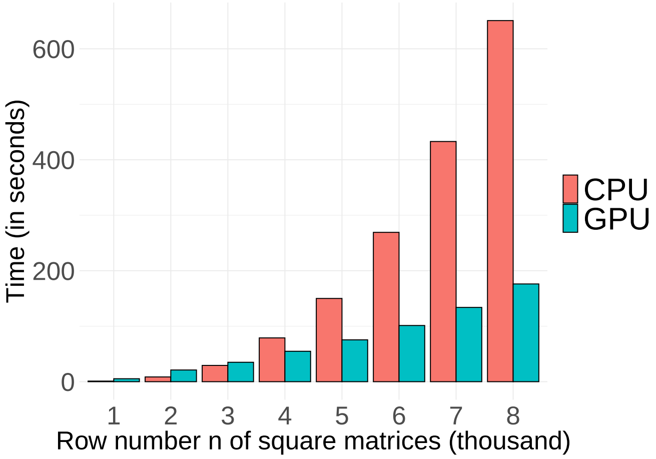


Figure 3.7: Comparison between time complexity of matrix operations, inversion and log determinant, using CPU and GPU.

3.6 Discussion

We proposed a spatiotemporal model to model voxel level BOLD signals that are tampered by spatiotemporal noises. By fully addressing the spatiotemporal dependency structure, we were able to estimate the correlation coefficient in our model and use it to quantify brain connectivity. Large sample properties were established with mild conditions on the smoothness of covariance structure and the decaying rate of dependency. Simulation studies showed that our model parameters can be accurately estimated and is advantageous to the conventional Pearson correlation of averages in the presence of spatiotemporal noises. Furthermore, the application on rat data showed that our findings are mostly consistent with previously established results, which indicates that our model can produce reasonable results while being superior in certain noisy settings. Further development includes more flexible covariance structure construction. Indeed, section 3.3 indicates that the covariance structure could be rather flexible. Consequently, our model can be extended to more general processes that share the similar noisy spatiotemporal features.

Bibliography

Achard, S., Coeurjolly, J.-F., Marcillaud, R., and Richiardi, J. (2011). fmri functional connectivity estimators robust to region size bias. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 813–816. IEEE.

Achard, S. and Gannaz, I. (2019). Wavelet-based and fourier-based multivariate whittle estimation: multiwave. *Journal of Statistical Software*, 89(6).

Achard, S., Salvador, R., Whitcher, B., Suckling, J., and Bullmore, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *Journal of Neuroscience*, 26(1):63–72.

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient Flows in Metric Spaces and in the Spaces of Probability Measures*. Springer Science & Business Media.

Aue, A., Gabrys, R., Horváth, L., and Kokoszka, P. (2009). Estimation of a change-point in the mean function of functional data. *Journal of Multivariate Analysis*, 100(10):2254–2269.

Becq, J.-P. G., Habet, T., Collomb, N., Faucher, M., Delon-Martin, C., Coizet, V., Achard, S., and Barbier, E. L. (2020). Functional connectivity is preserved but reorganized across several anesthetic regimes. *NeuroImage*, 219:116945.

Bekierman, J. and Gribisch, B. (2019). A mixed frequency stochastic volatility model for intraday stock market returns. *Journal of Financial Econometrics*. https://doi.org/10.1093/jjfinec/nbz021.

Bel, L., Bar-Hen, A., Cheddadi, R., and Petit, R. (2008). Spatio-temporal functional regression on paleo-ecological data. *arXiv preprint arXiv:0807.2588*.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Berkes, I., Gabrys, R., Horváth, L., and Kokoszka, P. (2009). Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society (B)*, 71:927–946.

Berkes, I. and Horváth, L. (2001). Strong approximation of the empirical process of garch sequences. *Annals of Applied Probability*, pages 789–809.

Berkes, I., Horváth, L., Kokoszka, P., and Shao, Q.-M. (2006). On discriminating between long-range dependence and changes in mean. *The annals of statistics*, 34(3):1140–1165.

Bigot, J., Gouet, R., Klein, T., and López, A. (2017). Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l'Institut Henri Poincaré B: Probability and Statistics*, 53:1–26.

Billingsley, P. (2013). *Convergence of probability measures.* John Wiley & Sons.

Bosq, D. (2000). *Linear Processes in Function Spaces.* Springer.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2:107–144.

Brockwell, P. J. and Davis, R. A. (1991). *Time series: theory and methods.* Springer Science & Business Media.

Brockwell, P. J. and Lindner, A. (2010). Strictly stationary solutions of autoregressive moving average equations. *Biometrika*, 97:765–772.

Brockwell, P. J., Lindner, A., and Vollenbröker, B. (2013). Strictly stationary solutions of multivariate ARMA equations with i.i.d. noise. *Ann. Inst. Statist. Math*, 64:1089–1119.

Chaimow, D., Yacoub, E., Uğurbil, K., and Shmuel, A. (2018). Spatial specificity of the functional mri blood oxygenation response relative to neuronal activity. *Neuroimage*, 164:32–47.

Chen, Y., Härdle, W., and Pigorsch, U. (2010). Localized realized volatility modeling. *Journal of the American Statistical Association*, 105:1376–1393.

Chen, Y., Lin, Z., and Müller, H.-G. (2020). Wasserstein regression. *arXiv preprint arXiv:2006.09660.*

Christakos, G. (2000). *Modern spatiotemporal geostatistics*, volume 6. Oxford university press.

Cressie, N. (1993). *Statistics for spatial data.* John Wiley & Sons.

Cressie, N. and Lahiri, S. N. (1993). The asymptotic distribution of reml estimators. *Journal of multivariate analysis*, 45(2):217–233.

Cressie, N. and Lahiri, S. N. (1996). Asymptotics for reml estimation of spatial covariance parameters. *Journal of Statistical Planning and Inference*, 50(3):327–341.

Dabo-Niang, S. and Yao, A.-F. (2007). Kernel regression estimation for continuous spatial processes. *Mathematical Methods of Statistics*, 16(4):298–317.

Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics: The official journal of the International Environmetrics Society*, 21(3-4):224–239.

Doukhan, P. (2012). *Mixing: properties and examples*, volume 85. Springer Science & Business Media.

Egozcue, J. J., Díaz-Barrero, J. L., and Pawlowsky-Glahn, V. (2006). Hilbert space of probability density functions based on aitchison geometry. *Acta Mathematica Sinica*, 22(4):1175–1182.

Eickhoff, S. B., Yeo, B., and Genon, S. (2018). Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, 19(11):672–686.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, pages 2008–2036.

Fan, J. and Gijbels, I. (2018). *Local polynomial modelling and its applications*. Routledge.

Flaxman, S., Wilson, A., Neill, D., Nickisch, H., and Smola, A. (2015). Fast kronecker inference in gaussian processes with non-gaussian likelihoods. In *International Conference on Machine Learning*, pages 607–616. PMLR.

Gasser, T., Muller, H.-G., Kohler, W., Molinari, L., and Prader, A. (1984). Nonparametric regression analysis of growth curves. *The Annals of Statistics*, pages 210–229.

Giraldo, R., Delicado, P., and Mateu, J. (2010). Continuous time-varying kriging for spatial prediction of functional data: An environmental application. *Journal of agricultural, biological, and environmental statistics*, 15(1):66–82.

Giraldo, R., Delicado, P., and Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and ecological statistics*, 18(3):411–426.

Gromenko, O., Kokoszka, P., Zhu, L., and Sojka, J. (2012). Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. *The Annals of Applied Statistics*, pages 669–696.

Hall, P., Müller, H.-G., and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics*, 34(3):1493–1517.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). . . . and the cross-section of expected returns. *The Review of Financial Studies*, 29:5–68.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2):383–385.

Hörmann, S. and Kokoszka, P. (2011). Consistency of the mean and the principal components of spatially distributed functional data. In *Recent Advances in Functional Data Analysis and Related Topics*, pages 169–175. Springer.

Horta, E. and Ziegelmann, F. (2018). Dynamics of financial returns densities: A functional approach applied to the bovespa intraday index. *International Journal of Forecasting*, 34(1):75–88.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications.* Springer.

Horváth, L., Kokoszka, P., and Rice, G. (2014). Testing stationarity of functional time series. *Journal of Econometrics*, 179:66–82.

Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., and Filzmoser, P. (2016). Simplicial principal component analysis for density functions in bayes spaces. *Computational Statistics & Data Analysis*, 94:330–350.

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons.

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, pages 805–820.

Jiang, J. and Nguyen, T. (2007). *Linear and generalized linear mixed models and their applications*, volume 1. Springer.

Kaiser, M. (2011). A tutorial in connectome analysis: topological and spatial features of brain networks. *Neuroimage*, 57(3):892–907.

Klepsch, J., Küppelberg, C., and Wei, T. (2017). Prediction of functional ARMA processes with an application to traffic data. *Econometrics and Statistics*, 1:128–149.

Kneip, A. and Utikal, K. J. (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association*, 96(454):519–542.

Kokoszka, P., Miao, H., Petersen, A., and Shang, H. L. (2019). Forecasting of density functions with an application to cross-sectional and intraday returns. *International Journal of Forecasting*, 35(4):1304–1317.

Kokoszka, P. and Reimherr, M. (2017a). *Introduction to Functional Data Analysis*. CRC Press.

Kokoszka, P. and Reimherr, M. (2017b). *Introduction to Functional Data Analysis*. Chapman and Hall/CRC.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical statistics*, 22:79–86.

Lindstrom, M. J. and Bates, D. M. (1988). Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.

Liu, D. C. and Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528.

Lütkepohl, H. (2006). *New Introduction to Multiple Time Series Analysis*. Springer.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146.

Mazzuco, S. and Scarpa, B. (2015). Fitting age-specific fertility rates by a flexible generalized skew normal probability density function. *Journal of the Royal Statistical Society (A)*, 178:187–203.

Meskaldji, D. E., Ottet, M.-C., Cammoun, L., Hagmann, P., Meuli, R., Eliez, S., Thiran, J. P., and Morgenthaler, S. (2011). Adaptive strategy for the statistical analysis of connectomes. *PloS one*, 6(8):e23009.

Moghimi, P., Dang, A. T., Netoff, T. I., Lim, K. O., and Atluri, G. (2021). A review on mr based human brain parcellation methods. *arXiv preprint arXiv:2107.03475*.

Panaretos, V. M. and Zemel, Y. (2016). Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771–812.

Panaretos, V. M. and Zemel, Y. (2020). *An invitation to statistics in Wasserstein space*. Springer Nature.

Pawlowsky-Glahn, V., Egozcue, J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Wiley.

Petersen, A., Liu, X., and Divani, A. A. (2020). Wasserstein F-tests and confidence bands for the Fréchet regression of density response curves. *Annals of Statistics, to appear.*

Petersen, A. and Müller, H.-G. (2019). Wasserstein covariance for multiple random densities. *Biometrika*, 106:339–351.

Petersen, A., Müller, H.-G., et al. (2016a). Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 44(1):183–218.

Petersen, A., Zhang, C., and Kokoszka, P. (2022). Modeling probability density functions as data objects. *Econometrics and Statistics*, 21:159–178.

Petersen, A., Zhao, J., Carmichael, O., and Müller, H.-G. (2016b). Quantifying individual brain connectivity with functional principal component analysis for networks. *Brain Connectivity*, 6(7):540–547.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer science & business media.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis 2nd ed.*, volume Springer Series in Statistics. New York: Springer.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(1):233–243.

Richiardi, J., Achard, S., Bunke, H., and Van De Ville, D. (2013). Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience. *IEEE Signal processing magazine*, 30(3):58–70.

Salazar, P., Di Napoli, M., Jafari, M., Jafarli, A., Ziai, W., Petersen, A., Mayer, S. A., Bershad, E. M., Damani, R., and Divani, A. A. (2019). Exploration of multiparameter hematoma 3d image analysis for predicting outcome after intracerebral hemorrhage. *Neurocritical care*, pages 1–11.

Shang, H. L. and Haberman, S. (2020). Forecasting age distribution of death counts: an application to annuity pricing. *Annals of Actuarial Science*, 14:150–169.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Shumway, R. H. and Stoffer, D. S. (2018). *Time Series Analysis and Its Applications*. Springer.

Spangenberg, F. (2013). Strictly stationary solutions of ARMA equations in Banach spaces. *Journal of Multivariate Analysis*, 121:127–138.

Srivastava, A., Jermyn, I., and Joshi, S. (2007). Riemannian analysis of probability density functions with applications in vision. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Sweeting, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics*, pages 1375–1381.

Termenon, M., Jaillard, A., Delon-Martin, C., and Achard, S. (2016). Reliability of graph analysis of resting state fmri using test-retest dataset from the human connectome project. *Neuroimage*, 142:172–187.

Van Den Heuvel, M. P. and Pol, H. E. H. (2010). Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534.

Villani, C. (2003). *Topics in Optimal Transportation*. Number 58. American Mathematical Soc.

Wang, J. (2012). *A state space model approach to functional time series and time series driven by differential equations*. PhD thesis, Rutgers University-Graduate School-New Brunswick.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295.

Wellner, J. et al. (2013). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Yang, H., Baladandayuthapani, V., Rao, A. U., and Morris, J. S. (2020). Quantile function on scalar regression analysis for distributional data. *Journal of the American Statistical Association*, 115(529):90–106.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470):577–590.

Zalesky, A., Fornito, A., and Bullmore, E. (2012). On the use of correlation as a measure of network connectivity. *Neuroimage*, 60(4):2096–2106.

Zhang, X. and Shao, X. (2015). Two sample inference for the second-order property of temporally dependent functional data. *Bernoulli*, 21:909–929.

Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5):2281–2321.