Wasserstein multivariate auto-regressive models for modeling distributional time series and its application in graph learning

Yiye Jiang\textsuperscript{1,2}

\textsuperscript{1}Institut de Mathématiques de Bordeaux, Université de Bordeaux
\textsuperscript{2}Laboratoire Bordelais de Recherche en Informatique, Université de Bordeaux

Abstract

We propose a new auto-regressive model for the statistical analysis of multivariate distributional time series. The data of interest consist of a collection of multiple series of probability measures supported over a bounded interval of the real line, and that are indexed by distinct time instants. The probability measures are modelled as random objects in the Wasserstein space. We establish the auto-regressive model in the tangent space at the Lebesgue measure by first centering all the raw measures so that their Fréchet means turn to be the Lebesgue measure. Using the theory of iterated random function systems, results on the existence, uniqueness and stationarity of the solution of such a model are provided. We also propose a consistent estimator for the model coefficient. In addition to the analysis of simulated data, the proposed model is illustrated with two real data sets made of observations from age distribution in different countries and bike sharing network in Paris. Finally, due to the simplex constraints that we impose on the model coefficients, the proposed estimator that is learned under these constraints, naturally has a sparse structure. The sparsity allows furthermore the application of the proposed model in learning a graph of temporal dependency from the multivariate distributional time series.

1 Introduction

Distributional time series is a recent research field that deals with observations that can be modeled as sequences of time-dependent probability distributions. Such distributional time series are ubiquitous in many scientific fields. A pertinent example is the analysis of sequences of the indicator distributions supported over age intervals, such as mortality and fertility (Mazzuco and Scarpa, 2015; Shang and Haberman, 2020), over calendar years in demographic studies. Other examples include daily stock return distributions from financial time series (Kokoszka et al., 2019; Zhang et al., 2021), the distributions of correlations between pairs of voxels within brain regions (Petersen and Müller, 2016). In Figure 1, we illustrate the data type with the time series of age distributions for countries in the European Union. This data set will also be used in our numeric experiments.

Since distributions can be characterized by functions, such as densities, quantile functions, and cumulative distribution functions, then to analyze the distributional time series, one may turn to study one of its functional representations with the tools from functional time series analysis (Bosq, 2000). However, due to their nonlinear constraints, such as monotonicity and positivity, the representing
Figure 1: Annual records of age distributions of EU countries. On the top are 27 countries in the European union. A sequence of age distribution is recorded at each country over years. For example, at the bottom we illustrate the sequence of France, where one distribution supported over $[0, 1]$ is observed at each year. On the lower left, we visualize the resulting univariate distributional time series with a surface in the coordinate system of Age $\times$ Year $\times$ Relative frequency. The raw data in this plot consist in 40 annual distributions. We complete them with interpolated samples to draw the surface. On the lower right, we show the projection of the raw time series onto the Age $\times$ Relative frequency plane. We can see that the population is aging along time.
functions of distributions do not constitute linear spaces. Consequently, basic notions in standard vector auto-regressive (VAR) models (Helmut, 2005), such as additivity and scalar multiplication, do not adapt, in a straightforward manner. This causes models devised for random elements of a Hilbert space often to fail. One existing approach is to map the densities of distributions to unconstrained functions in the Hilbert space by the log quantile density (LQD) transformation (Petersen and Müller, 2016), and then apply the functional tools (Kokoszka et al., 2019). However LQD does not take into account the geometry of the distribution space, thus, it can lead to deformations in the distance. Recent approaches consider such geometry by adopting the Wasserstein metric (Bigot et al., 2017; Panaretos and Zemel, 2016; Petersen and Müller, 2019b). For recent reviews on the statistics in Wasserstein space, we refer to Bigot (2020); Panaretos and Zemel (2020); Petersen et al. (2022).

For univariate times series to the setting of a distributional time series supported over an interval in IR (that is when observing a random probability distribution at each time instead of a scalar), there mainly exist recent works Chen et al. (2021); Zhang et al. (2021); Zhu and Müller (2021) on the development of stochastic models that extend the standard auto-regressive model in Hilbert space to the distributional counterparts, based on the geometry of Wasserstein space. They propose to map the random distributions in the series to the Logarithmic image of Wasserstein space at their common Fréchet mean, and construct the univariate functional time series models in terms of the logarithmic maps. The common Fréchet mean is the ideal reference distribution in the sense that it cancels out the expectations of the logarithmic maps in the regression formula. However, for multivariate time series, there is no unique Fréchet mean. Moreover such ideal reference measure is very unlikely to exist. On the other hand, an extra intercept in such functional regression model will cause great difficulties, when one wants to retain the regression model especially the predictions in the logarithmic image of Wasserstein space. Thus to deal with the unequal means, we firstly center all the raw distributions in the proposed way so that their Fréchet means turn to be the Lebesgue measure. Then based on the centered data, we define the multivariate auto-regressive model in the logarithmic image at the Lebesgue measure.

Currently, the statistical approaches devised for distributional data especially the multivariate distributional data are very little developed. The proposed data centralization approach for random distribution in this work can also serve in the development of the multivariate distribution-to-distribution regressive model for the i.i.d. data. Other existing works on the distribution-to-distribution regressive model are Chen et al. (2019) where the LQD transformation and the functional tools are adopted, and Chen et al. (2021); Ghodrati and Panaretos (2022), where both works are dedicated for the regression of two distributions. Chen et al. (2021) performs the regression using the coefficient operator which maps from the tangent space of the predictor’s Fréchet means to that of the response’s. By contrast, Ghodrati and Panaretos (2022) performs the regression directly in Wasserstein space. The predictor is pushed forward by the coefficient optimal map to the conditional Fréchet mean of the response. It is not straightforward to extend the two models to more predictors since the related operation are both directional. That is, it is not evident how to map between multiple predictor tangent spaces or push among the predictor measures.

1.1 Paper organisation

In Section 2, we provide the background on the statistics in Wasserstein space that is needed for modeling distributional data. In Section 3, we introduce Wasserstein multivariate auto-regressive models that are adapted to the statistical analysis of multivariate distributional time series. We also prove the existence, uniqueness and stationarity of the solution of such models. Estimators of the regression coefficients in these models are studied in Section 4. Finally, numerical experiments with both simulated and real data are carried out in Section 5 to analyze the finite sample properties of
the estimator and to illustrate its application in graph learning from multivariate distributional time series.

1.2 Publicly available source code

For the sake of reproducible research, Python code available at https://github.com/yiyej/Wasserstein_Multivariate_Autoregressive_Model implements the proposed estimators and the experiments carried out in this paper.

2 Backgrounds on statistics in Wasserstein space

In this section, we present the background for modeling the distributional data considered in this work. Let \( \Omega \) be a (possibly unbounded) interval in \( \mathbb{R} \), and \( \mathcal{B}(\Omega) \) the associated \( \sigma \)-algebra made of Borel subsets of \( \Omega \). Let \( \mu \) be a probability measure (namely a distribution) over \( \Omega \), \( \mathcal{B}(\Omega) \) with cumulative distribution function (cdf) \( F_\mu \). Then the (generalized) quantile function is defined as the left continuous inverse of \( F_\mu \), denoted by \( F_\mu^{-1} \), that is

\[
F_\mu^{-1}(p) := \inf\{x \in \Omega : F_\mu(x) \geq p\}, \quad p \in (0, 1).
\]

The Wasserstein space \( \mathcal{W} := \mathcal{W}_2(\Omega) \) is defined as the set of probability measures over \( (\Omega, \mathcal{B}(\Omega)) \) with finite second moment, that is endowed with the \( L^2 \) Wasserstein distance

\[
d_W(\mu, \nu) = \left( \int_0^1 \left[ F_\mu^{-1}(p) - F_\nu^{-1}(p) \right]^2 dp \right)^{1/2}, \quad \mu, \nu \in \mathcal{W}_2(\Omega).
\] (2.1)

It is well known that \( \mathcal{W} \) is a complete and separable metric space (see e.g. Villani (2008) for a detailed course on optimal transport theory and Panaretos and Zemel (2020) for an introduction to the topic of statistical analysis in the Wasserstein space).

2.1 Tangent Bundle

The space \( \mathcal{W} \) has a pseudo-Riemannian structure as shown in Ambrosio et al. (2008). Letting \( \gamma \in \mathcal{W} \) be an absolutely continuous measure, the tangent space at \( \gamma \) is defined as

\[
\text{Tan}_\gamma = \{t(F_\mu^{-1} \circ F_\gamma \circ \text{id}) : \mu \in \mathcal{W}, \ t > 0\} \quad \mathcal{L}^2(\Omega),
\]

where \( \text{id} \) is the identity function, \( \mathcal{L}^2(\Omega) \) is the Hilbert space of \( \gamma \) square integrable functions on \( \Omega \), with inner product \( \langle \cdot, \cdot \rangle_\gamma \) defined by \( \langle f, g \rangle_\gamma := \int_\Omega f(x)g(x) \, d\gamma(x) \), \( f, g \in \mathcal{L}^2(\Omega) \), and the induced norm \( \| \cdot \|_\gamma \). The exponential and the logarithmic maps at \( \gamma \) are then defined as follows.

**Definition 2.1.** The exponential map \( \text{Exp}_\gamma : \text{Tan}_\gamma \rightarrow \mathcal{W} \) is defined as

\[
\text{Exp}_\gamma g = (g + \text{id})\#\gamma,
\] (2.2)

where for any measurable function \( T : \Omega \rightarrow \Omega \) and \( \mu \in \mathcal{W} \), \( T\#\mu \) is the pushforward measure on \( \Omega \) defined as \( T\#\mu(A) = \mu(\{x \in \Omega : T(x) \in A\}) \), for any set \( A \in \mathcal{B}(\Omega) \).
Definition 2.2. The logarithmic map $\text{Log}_\gamma : \mathcal{W} \to \text{Tan}_\gamma$ is defined as

$$\text{Log}_\gamma \mu = F_{\mu}^{-1} \circ F_{\gamma} - \text{id}.$$ 

Note that the exponential map (2.2) is not a local homeomorphism (Ambrosio et al., 2004). Nevertheless, when restricted to the image of the logarithmic map, it becomes an isometry (Bigot et al., 2017) as stated in the following proposition.

Proposition 2.1. Let $\gamma \in \mathcal{W}$ be any absolutely continuous measure. Then $\text{Exp}_\gamma |\text{Log}_\gamma \mathcal{W}$ to $\mathcal{W}$, with the inverse map $\text{Log}_\gamma$, satisfying

$$d_\mathcal{W}(\mu, \nu) = \| \text{Log}_\gamma \mu - \text{Log}_\gamma \nu \|_\gamma.$$ 

The Wasserstein distance in Equation (2.1) can be interpreted as $\| \text{Log}_{Leb} \mu - \text{Log}_{Leb} \nu \|_{Leb}$, which is the distance between Logarithmic maps of $\mu$ and $\nu$ in the tangent space at $\gamma$ "Leb the uniform distribution over $[0, 1]$, if $[0, 1] \in \Omega$. For the general definition of Wasserstein distance (between probability measures supported in general metric spaces) related to the optimal transport theory, we refer to Villani (2021); Panaretos and Zemel (2020).

We can also remark from the above results that the space of all quantile functions of measures in $\mathcal{W}$, namely $\text{Log}_{Leb} \mathcal{W}$, is a complete separable metric space with respect to $\| \cdot \|_{Leb}$.

We recall below some important properties of $\text{Log}_\gamma \mathcal{W}$ (Bigot et al., 2017) that are needed in the construction of the statistical models introduced in Section 3.

Proposition 2.2. $\text{Log}_\gamma \mathcal{W}$ is a closed and convex subset of $\mathcal{L}_2^2(\Omega)$.

Proposition 2.3. Let $g \in \text{Tan}_\gamma$, then $g \in \text{Log}_\gamma \mathcal{W}$ if and only if $g + \text{id}$ is nondecreasing $\gamma$-almost everywhere.

2.2 Fréchet Means in Space $\mathcal{W}$

Definition 2.3. Let $\mu_1, \ldots, \mu_T$ be measures in $\mathcal{W}$. The empirical Fréchet mean of $\mu_1, \ldots, \mu_T$, denoted by $\bar{\mu}$, is defined as the unique minimizer of

$$\min_{\nu \in \mathcal{W}} \frac{1}{T} \sum_{t=1}^{T} d_\mathcal{W}(\mu_t, \nu).$$ 

It is well known that the empirical Fréchet mean $\bar{\mu}$ admits a simple expression through its quantile function that satisfies

$$F_{\bar{\mu}}^{-1}(p) = \frac{1}{T} \sum_{t=1}^{T} F_{\mu_t}^{-1}(p), \ p \in (0, 1).$$

Definition 2.4. A random measure $\mu$ is any measurable map from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the metric space $\mathcal{W}$, endowed with its Borel $\sigma$-algebra.

In what follows, we use bold notation to distinguish random measures from constant (that is non-random) measures.

Definition 2.5. Let $\mu$ be a random measure from probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathcal{W}$. Assume that $\mu$ is square integrable, namely $\mathbb{E}d_\mathcal{W}^2(\mu, \nu) < \infty$ for some (thus for all) $\nu \in \mathcal{W}$. Then, the population Fréchet mean of $\mu$, denoted by $\mu_\mathbb{P}$, is defined as the unique minimizer of

$$\min_{\nu \in \mathcal{W}} \mathbb{E} \left[ d_\mathcal{W}^2(\mu, \nu) \right].$$
Note that $\mu_\odot$ also admits a simple expression through its quantile function as

$$F_{\mu_\odot}^{-1}(p) = \mathbb{E}[F_{\mu}^{-1}(p)] , \; p \in (0, 1).$$

In the remarks below, we point out two important facts about Fréchet mean and the Logarithmic map.

**Remark 2.1.** The random measure $\mu$ has zero expectation in the tangent space at its Fréchet mean, that is

$$\mathbb{E}\log_{\mu_\odot} \mu = 0, \; \mu_\odot$$

almost everywhere.

**Remark 2.2.** The Fréchet mean can be calculated using standard expectation in the tangent space at any absolute continuous measure $\gamma \in \mathcal{W}$, that is

$$\mu_\odot = \exp_{\gamma}(\mathbb{E}\log_{\gamma} \mu).$$

### 3 Wasserstein multivariate auto-regressive Models

In this section, we are concerned by the statistical analysis of data that are collected over a network of $N$ features or sensors (typically a set of spatial locations), recording, over time, observations of multiple measurements. Standard VAR models are adapted to the case where real measurements $x_i^t \in \mathbb{R}, \; t \in \mathbb{Z}, \; i = 1, \ldots, N$ are collected for each feature $i$ and at each time $t$. In this work, we focus on the more involved setting of multivariate distributional time series. This corresponds to the situation where one records, for each feature $i$ and time $t$, a random probability measure $\mu_i^t$ (that is assumed to be supported on an interval of the real line). Hence, we consider a collection of $N$ time-dependent probability measures $(\mu_i^t)_{t \in \mathbb{Z}}$ for $i = 1, \ldots, N$ that are indexed by distinct time instants $t$. Then, the purpose of the present work is to extend standard VAR models to analyse such data. Especially, we wish to develop a model that can identify the dependency structure in the temporal evolution of the measures.

#### 3.1 Description of the model

We now present the proposed auto-regressive model for multivariate distributional time series $(\mu_i^t)_{t \in \mathbb{Z}}, \; i = 1, \ldots, N$. We assume $\mu_i^t$ belongs to $\mathcal{W}$ for each feature $i$ and time instant $t$. We firstly recall the standard VAR model, then we generalize the corresponding notions to the Wasserstein space to develop a new notion of multivariate distributional auto-regressive model.

Given a multivariate time series $x_i^t \in \mathbb{R}, \; t \in \mathbb{Z}, \; i = 1, \ldots, N$, we primarily assume that the mean is time invariant for each component $i$ of the multivariate time series, namely, $u^t = \mathbb{E}(x_i^t), \; t \in \mathbb{Z}$. Then the VAR model of order 1 writes as Helmut (2005)

$$x_i^t - u_i = \sum_{j=1}^{N} A_{ij} (x_j^{t-1} - u_i) + \epsilon_i^t,$$

where $\epsilon_i^t$ is a white noise. The univariate series $(x_i^t - u_i)_{t \in \mathbb{Z}}$ is the centered series of $(x_i^t)_{t \in \mathbb{Z}}$ by subtracting the component mean $u_i$. The model is then built on the centered data and does not include an intercept term. For the multivariate distributional time series $\mu_i^t, \; t \in \mathbb{Z}, \; i = 1, \ldots, N$, we firstly assume that the mean is time invariant for each component $i$ as well, as stated in Assumption A1 below.
Assumption A1. For each fixed $i = 1, \ldots, N$, the random probability measures $\mu_i^t$, $t \in \mathbb{Z}$ are square integrable and they have the same Fréchet mean denoted by $\mu_i$. We then propose the following “data centering step” for $\mu_i^t$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$. The centered measures are denoted by $\tilde{\mu}_i^t$, and they are defined through their quantile functions given as

$$
\tilde{F}_{i,t}^{-1} = F_{i,t}^{-1} \ominus F_{i,\emptyset}^{-1} := F_{i,t}^{-1} \circ (F_{i,\emptyset}^{-1})^{-1},
$$

(3.2)

where $F_{i,t}^{-1}$ is the quantile function of $\mu_i^t$ extended at 0 (Bobkov and Ledoux, 2014, Section A1) by

$$
F_{i,t}^{-1}(0) := \inf\{x \in [0, 1] : F_{i,t}(x) > 0\},
$$

and $F_{i,\emptyset}^{-1}$ is the quantile function of $\mu_i$. Note that in most cases the function $(F_{i,\emptyset}^{-1})^{-1}$ is equal to the cdf $F_{i,\emptyset}$. However, when $F_{i,\emptyset}$ is only right-continuous but not continuous, $(F_{i,\emptyset}^{-1})^{-1}$ is not equal to $F_{i,\emptyset}$, since the left-continuous inverse only gives a left-continuous function. Thus, we shall keep the notation $(F_{i,\emptyset}^{-1})^{-1}$.

The centering step (3.2) at the level of the quantile functions is thus analogous to the usual centering step in VAR models for Euclidean data. From the optimal transport point of view, the centered quantile function $\tilde{F}_{i,t}^{-1}$ is interpreted as the optimal transport map from the Fréchet mean $\mu_i$ to the measure $\mu_i^t$. The notation $\ominus$ in (3.2) as a difference operator between two increasing functions is taken from the recent work in Zhu and Müller (2021) on auto-regressive model for univariate (that is $N = 1$) distributional time series. We remark that the output of this difference operator remains an increasing function.

The function $\tilde{F}_{i,t}^{-1}$ needs to be defined over $[0, 1]$ as a valid quantile function. Furthermore, as a result of data centering, we aim to turn the Fréchet mean of $\tilde{F}_{i,t}^{-1}$ to be Lebesgue measure. Thus we impose Assumption A2 below.

Assumption A2. All $\mu_i^t$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$ are supported on the same closed and bounded interval $\mathcal{D} \subset \Omega$. Without loss of generality, we assume that $\mathcal{D} = [0, 1]$.

Under Assumption A2, the support of $\mu_i^t$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$ is a closed and bounded interval, and thus, their cdf are strictly increasing function on this support [Proposition A7](Bobkov and Ledoux, 2014). Consequently, all the quantile function $F_{i,t}^{-1}$ are continuous. Thus, the Fréchet mean has a continuous quantile function $F_{i,\emptyset}^{-1}$. The continuity of $F_{i,\emptyset}$ makes Equation (3.3) holds (Bobkov and Ledoux, 2014, Lemma A.3.5).

$$
\mathbb{E} \left[ \tilde{F}_{i,t}^{-1}(p) \right] = (\mathbb{E} F_{i,t}^{-1}) \left[ (F_{i,\emptyset}^{-1})^{-1}(p) \right] = p, \quad p \in (0, 1).
$$

(3.3)

Thus, all the centered distributional time series $(\tilde{\mu}_i^t)_{t \in \mathbb{Z}}$ have the same Fréchet mean, which equals to the Lebesgue measure. We illustrate this performance with synthetic data in Figure 2. We then propose to build an auto-regressive model for multivariate distributional time series with respect to the transformed data $\tilde{\mu}_i^t$ in the tangent space of the Lebesgue measure, that takes the following expression:

$$
\tilde{\mu}_i^t = \epsilon_{i,t} \# \text{Exp}_{\text{Leb}} \left( \sum_{j=1}^{N} A_{ij} \log_{\text{Leb}} \tilde{\mu}_{i-1}^j \right), \quad t \in \mathbb{Z}, \ i = 1, \ldots, N,
$$

(3.4)

where $\{\epsilon_{i,t}\}_{i,t}$ are i.i.d. random distortion functions taking values in the space of extended quantile functions

$$
\Pi = \{F^{-1} : [0, 1] \to [0, 1], \text{ such that } F^{-1}|_{(0,1)} \in \text{Log}_{\text{Leb}} \mathcal{W} + \text{id}, \}
$$

$$
F^{-1}(0) := \inf\{x \in [0, 1] : F(x) > 0\}, \text{ and } F^{-1}(1) := \sup\{x \in [0, 1] : F(x) < 1\},
$$

7
endowed with $\| \cdot \|_{\text{Leb}}$ and the induced Borel algebra, $\epsilon_{i,t}$ is almost surely independent of $\mu_i^{t-1}$, $i = 1, \ldots, N$, for all $t \in \mathbb{Z}$, and

$$E[\epsilon_{i,t}(x)] = x, \ x \in [0,1].$$

Note that all the univariate time series of log maps are centered to 0, namely, $E[\log_{\text{Leb}} \tilde{\mu}_i^t] = 0$, $\forall t \in \mathbb{Z}$, $i = 1, ..., N$ as in Model (3.1).

The pushforward in (3.4) under $\epsilon_{i,t}$ is a valid approach to provide random distortions of probability measures as proposed in Petersen and Müller (2019a). This approach is also used in Chen et al. (2021). An example of random distortion function satisfying the conditions in Equation (3.4) as well as in Assumption A4 imposed later on, can be found, for example in Chen et al. (2021, Equation (38)). However, in these works, not many examples of valid random distortion functions which satisfy the conditions in Equation (3.4) are given. Thus, to demonstrate that the conditions imposed on the distortion function are not restrictive, we describe, in Section 5 on numerical experiments, a general mechanism to generate random distortion functions that satisfy both Equation (3.4) and Assumption A4.

Note that, the total coefficients of the regression system defines a matrix $A = (A_{ij})_{ij}$ as in VAR models. By contrast, for general functional regression models with function response, the coefficients can be concurrent $A_{ij}(\cdot)$ or the most general a surface $A_{ij}(\cdot, \cdot)$, see Wang et al. (2015, Equations (14) and (15), respectively). Nevertheless, we purpose to use matrix form for the regression coefficients in order to represent the dependency structure between $N$ time-dependent probability measures directly by its corresponding directed and weighted graph.

To fit Model (3.4), a least squares estimator of the matrix $A$ can be constructed by minimizing the expected squared Wasserstein distance (2.1) between $\tilde{\mu}_i^t$ and its prediction $\text{Exp}_{\text{Leb}} \left( \sum_{j=1}^{N} A_{ij} \log_{\text{Leb}} \tilde{\mu}_j^{t-1} \right)$. When $\sum_{j=1}^{N} A_{ij} \log_{\text{Leb}} \tilde{\mu}_j^{t-1}$ belongs to $\log_{\text{Leb}} W$, the quantile function of its Exponential map $\text{Exp}_{\text{Leb}}$ is simply given by $\sum_{j=1}^{N} A_{ij} (\tilde{F}_{j,t-1} - id) + id$. By contrast, when $\sum_{j=1}^{N} A_{ij} \log_{\text{Leb}} \tilde{\mu}_j^{t-1}$ falls out of $\log_{\text{Leb}} W$, the dependency between the quantile function and the coefficients $A_{ij}$ is non-tractable, see Cazelles et al. (2017, Proposition 3.1). On the other hand, retaining the model in $\log_{\text{Leb}} W$ will avoid the non-identifiability problem of parametric models, thanks to Proposition 2.1. Thus, it will be needed to assume that $\sum_{j=1}^{N} A_{ij} \log_{\text{Leb}} \tilde{\mu}_j^{t-1} \in \log_{\text{Leb}} W$. Since $\tilde{\mu}_i^{-1}$ can take any value in $W$, imposing such assumption amounts to the following $N$-simplex constraint on the rows of $A$, given the convexity of the logarithmic image. Similar assumptions are imposed in related works, see e.g.
Chen et al. (2021, Assumption (A1)) and Petersen and Müller (2019b, Assumption (A3)), to keep the regression model in the logarithmic image.

**Assumption A3.** $\sum_{j=1}^{N} A_{ij} \leq 1$ and $0 \leq A_{ij} \leq 1$.

An additional important advantage of Assumption (A3) is that it leads to least squares estimation of the matrix $A$ under an $\ell_1$ ball constraint on its coefficients. In this manner, the estimators of the coefficients $A_{ij}$ will naturally be sparse, which is a favorable property in graph learning.

Given Assumption A3, we can also build the auto-regressive model directly with respect to the quantile function $r_{\cdot, \cdot}$ as

$$r_{F^{-1}_{i,t}} = \epsilon_{i,t} \circ \left[ \sum_{j=1}^{N} A_{ij} \left( \hat{F}_{j,t-1}^{-1} - id \right) + id \right], \quad t \in \mathbb{Z}, i = 1, \ldots, N. \quad (3.5)$$

When reducing to the univariate case ($N = 1$), Model (3.5) is similar to the auto-regressive model proposed in Zhu and Müller (2021, Model (4)), when a regression coefficient belongs to $\mathbb{P}_0$, $\mathbb{P}_1$.

**Remark 3.1.** To deal with the unequal expectations of random measures, we may alternatively find a reference measure at whose tangent space, a proposed intercept-free regression holds in expectation. Unfortunately, such a valid reference measure is very unlikely to exist. For example, consider the following regression model of three random measures, $\mu$, $\nu_1$, and $\nu_2$

$$E \left[ \log \gamma (\mu) \mid \nu_1, \nu_2 \right] = \lambda_1 \log \gamma (\nu_1) + \lambda_2 \log \gamma (\nu_2),$$

at the tangent space of an unknown reference measure $\gamma$ to be identified. A valid measure $\gamma$ that makes the model hold in expectation, needs to satisfy

$$F_{\gamma}^{-1} = \frac{F_{\mu}^{-1} - \lambda_1 F_{\nu_1}^{-1} - \lambda_2 F_{\nu_2}^{-1}}{1 - \lambda_1 - \lambda_2}. \quad (3.6)$$

However, the right hand side of the above equation is not necessarily an increasing function, which implies that a valid reference measure fails to exist in all negative cases. Furthermore, when plugging relationship (3.6) in the regression model, we obtain its reparameterization in terms of the model coefficient and the expectations of random measures

$$E \left[ F_{\mu}^{-1} - F_{\nu_1}^{-1} \mid \nu_1, \nu_2 \right] = \lambda_1 (F_{\nu_1}^{-1} - F_{\nu_2}^{-1}) + \lambda_2 (F_{\nu_2}^{-1} - F_{\nu_1}^{-1}).$$

However, the centering in this reparameterized model is linear given through the subtraction in Hilbert space, which does not bring well-defined new measures. Indeed, such centering can be equivalently obtained by adding an intercept term to the regression model established in the tangent space of Lebesgue measure. Therefore, by contrast to these two equivalent operations, we propose to consider nonlinear centering of quantile functions, when developing multivariate regression models in Wasserstein space.

### 3.2 Existence, uniqueness and stationarity

To study the legitimacy of the iterated random functions (IRF) system defined by Model (3.5), we shall consider the product metric space

$$(\mathcal{X}, d) := (\mathcal{T}, \| \cdot \|_{\text{Leb}}) \otimes \mathbb{N},$$
where $\mathcal{T} := \text{Log}_{\text{Leb}} \mathcal{W} + id$ is the space of all quantile functions of $\mathcal{W}$, equipped with the norm $\| \cdot \|_{\text{Leb}}$ in the tangent space at the Lebesgue measure. Thus, we have

$$
d(X, Y) := \sqrt{\sum_{i=1}^{N} \| X_i - Y_i \|_{\text{Leb}}^2}, \quad X = (X_i)_{i=1}^{N} \in \mathcal{X}, \quad Y = (Y_i)_{i=1}^{N} \in \mathcal{X}.
$$

(3.7)

The auto-regressive model (3.5) can be interpreted as an iterated random functions (IRF) system operating on the state space $(\mathcal{X}, d)$, written as

$$
X_t = \Phi_{\epsilon_t}(X_{t-1}),
$$

(3.8)

where $X_t = (X_{i,t})_{i=1}^{N}$, $\epsilon_t = (\epsilon_{i,t})_{i=1}^{N}$, and $\Phi_{\epsilon_t}(X_{t-1}) = (\Phi_{\epsilon_t}(X_{i,t-1}))_{i=1}^{N}$ with

$$
\Phi_{\epsilon_t}(X_{i,t-1}) := \epsilon_{i,t} \circ \left[ \sum_{j=1}^{N} A_{ij} (X_{j,t-1} - id) + id \right].
$$

We first study the existence and the uniqueness of the solution to the IRF system in the metric space $(\mathcal{X}, d)$.

For time series models in a Hilbert space, two standard assumptions that ensure the existence and the uniqueness of the solutions are the boundedness of the $L_p$ norm of random additive noise and the contraction of the regression operator. For Model (3.8), the random noise $\epsilon_{i,t}$ is bounded between $0$ and $1$, and thus $\mathbb{E}[d^p(X, \epsilon)]$ is bounded for all $X \in \mathcal{X}$, which is the $L_p$ norm equivalent in the metric space setting. Then, to have a contractive map $\Phi_{\epsilon_t}$, we shall rely on an interplay between properties of the matrix $A$ of coefficients and the random noise distortion since it is applied in a nonlinear way. More specifically, we impose Assumptions A4 and A5 below on Model (3.8).

**Assumption A4.** $\mathbb{E}[(\epsilon_{i,t}(x) - \epsilon_{i,t}(y))^2] \leq L^2(x - y)^2$, $\forall x, y \in [0, 1]$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$.

**Assumption A5.** $\|A\|_2 < \frac{1}{2}$.

Note that, Assumption A4 implies that $\epsilon_{i,t}$ is $L$-Lipschitz in expectation. For increasing functions from $[0, 1]$ to $[0, 1]$, the smallest $L$ is $1$ that is attained by the identity function. Therefore, Assumption A5 implies that $\|A\|_2 < 1$, which is the usual contraction assumption for standard VAR models in an Euclidean space. We now state the existence and uniqueness results.

**Theorem 3.1.** Under Assumptions A3, A4 and A5, the IRF system (3.8) almost surely admits a solution $X_t$, $t \in \mathbb{Z}$, with the same marginal distribution $\pi$, namely, $X_t \overset{d}{=} \pi$, $\forall t \in \mathbb{Z}$, where the notation $\overset{d}{=} \text{ means equality in distribution.}$ Moreover, if there exists another solution $S_t$, $t \in \mathbb{Z}$, then for all $t \in \mathbb{Z}$

$$
X_t \overset{d}{=} S_t, \quad \text{almost surely.}
$$

Theorem 3.1 states that under Assumptions A4 and A5, a well defined IRF system (3.8) (namely when Assumption A3 is satisfied) permits a unique solution in $(\mathcal{X}, d)$ almost surely. Next, we show that this solution is furthermore stationary as a functional time series in a Hilbert space. To this end, we need to assume that there is an underlying Hilbert space associated to $(\mathcal{X}, d)$, with its inner product inducing $d$ as the norm. Such an Hilbert space exists, with corresponding inner product given by

$$
\langle X, Y \rangle = \sum_{i=1}^{N} \langle X_i, Y_i \rangle_{\text{Leb}}.
$$

We recall the conventional definition of stationarity for process in a separable Hilbert space, see for example Zhang et al. (2021, Definition 2.2).
Definition 3.1. A random process \( \{V_t\}_t \) in a separable Hilbert space \( (H, \langle \cdot, \cdot \rangle) \) is said to be stationary if the following properties are satisfied.

1. \( \mathbb{E} |V_t|^2 < \infty \).
2. The Hilbert mean \( U := \mathbb{E} [V_t] \) does not depend on \( t \).
3. The auto-covariance operators defined as
   \[ G_{t,t-h}(V) := \mathbb{E} \langle V_t - U, V \rangle (V_{t-h} - U), \quad V \in H, \]
   do not depend on \( t \), that is \( G_{t,t-h}(V) = G_{0,-h}(V) \) for all \( t \).

Then, Theorem 3.2 below gives the stationarity result.

Theorem 3.2. The unique solution given in Theorem 3.1 is stationary as a random process in \( (X, \langle \cdot, \cdot \rangle) \) in the sense of Definition 3.1.

Besides, Proposition 3.3 below states that the stationary solution of the IRF system (3.8) satisfies the property (3.3) of the transformed series \( \tilde{F}_{i,t}^{-1}, t \in \mathbb{Z} \). Thus, it is consistent to propose the IRF system (3.8) as the process that generated the data \( \tilde{F}_{i,t}^{-1} \), which completes the building of Model (3.5) as valid approach to analyze multivariate distributional time series.

Proposition 3.3. The stationary solution \( X_t \) of the IRF system (3.8) satisfies:

1. \( X_{i,t}(p) \in [0, 1], \quad \forall p \in (0, 1) \),
2. \( \mathbb{E} [X_{i,t}(p)] = p, \quad \forall p \in (0, 1) \).

Finally, we point out in Proposition 3.4 additional properties of the IRF system (3.8) that will serve in the following section of the estimation of the matrix of coefficients in Model (3.5).

Proposition 3.4. Given the stationary solution \( X_t \) of the IRF system (3.8), we define matrices \( \Gamma(0), \Gamma(1) \in \mathbb{R}^{N \times N} \) as

\[
\begin{align*}
\Gamma(0) \ &= \ E \langle X_{j,t-1} - id, X_{i,t-1} - id \rangle_{Leb} \\
\Gamma(1) \ &= \ E \langle X_{j,t} - id, X_{i,t-1} - id \rangle_{Leb},
\end{align*}
\]

for \( 1 \leq j, l \leq N \). We have

1. \( \Gamma(0) \) is nonsingular,
2. the coefficient matrix \( A \) of the IRF system (3.8) admits the representation

\[
A = \Gamma(1) [\Gamma(0)]^{-1}.
\]

Note that the expression (3.9) for the matrix \( A \) is analogous to the one of VAR models with matrices \( \Gamma(0), \Gamma(1) \) carrying out the information on the correlation. However, compared to the auto-covariance operators in Definition 3.1, the matrices \( \Gamma(0) \) and \( \Gamma(1) \) rather reflect the average auto-covariance taking into account additionally the correlated level along the function domain.

4 Estimation of the regression coefficients

In this section, we develop the estimators of coefficient \( A \), given \( T + 1 \) samples \( \mu_i^t, t = 0, 1, \ldots, T \) for each feature \( i = 1, \ldots, N \). We also show the consistency result of the proposed estimator. Note that we assume that the measures are fully observed, instead of indirectly observed through their samples.
4.1 A constrained least-square estimation method

As briefly explained before the statement of Assumption (A3), we could consider the estimator based on an unconstrained least squares method, which is defined as the minimizer of the sum of squared residuals measured by the Wasserstein distance:

\[
\hat{A}_i = \arg\min_{A_i} \frac{1}{T} \sum_{t=1}^{T} d_W^2 \left( \hat{\mu}_t^i, \text{Exp}_{\text{Leb}} \left( \sum_{j=1}^{N} A_{ij} \log_{\text{Leb}} \hat{\mu}_t^j \right) \right), \quad i = 1, \ldots, N, \tag{4.1}
\]

Analogous to Proposition 3.4, the estimator \( \hat{A} \) defined in Equation (4.1) admits the expression

\[
\hat{A} = \hat{\Gamma}^{-1} \left[ \hat{\Gamma}(0) \right],
\]

where

\[
\left[ \hat{\Gamma}(0) \right]_{j,l} = \frac{1}{T} \sum_{t=1}^{T} \langle \hat{F}_{j,t}^{-1} - \text{id}, \hat{F}_{l,t}^{-1} - \text{id} \rangle_{\text{Leb}}
\]

and

\[
\left[ \hat{\Gamma}(1) \right]_{j,l} = \frac{1}{T} \sum_{t=1}^{T} \langle \hat{F}_{j,t}^{-1} - \text{id}, \hat{F}_{l,t}^{-1} - \text{id} \rangle_{\text{Leb}}.
\]

For the estimator \( \hat{A} \) to hold, strictly speaking, we need to assume that \( \hat{\Gamma}(0) \) is nonsingular as in the case of classical least squares estimators. Note that \( \hat{A} \) is the exact least squares estimator constructed from the stationary solution of Model (3.5) without any constraint. However, in practice, we do not know the population Fréchet mean \( F_{i,0} \), thus we cannot calculate the exact centered data \( \hat{\mu}_t^i \) as in method (3.2). Therefore, we propose to first estimate \( F_{i,0} \) by the empirical Fréchet mean

\[
F_{\hat{\mu}_t}^{-1} = \frac{1}{T} \sum_{t=1}^{T} F_{\hat{\mu}_t,t}^{-1}, \tag{4.2}
\]

and center \( \mu_{i,t} \) by \( F_{\hat{\mu}_t}^{-1} \) as in Equation (4.3), to obtain the transformed data \( \hat{\mu}_{i,t} \).

\[
\hat{F}_{i,t}^{-1} := F_{i,t}^{-1} \ominus F_{\hat{\mu}_t}^{-1} = F_{i,t}^{-1} \circ F_{\hat{\mu}_t}. \tag{4.3}
\]

Using the data \( \hat{\mu}_{i,t} \) in the least squares formula (4.1) we obtain an approximate least squares estimator \( \hat{A}_o \) whose rows satisfy

\[
\left[ \hat{A}_o \right]_{i,j} = \arg\min_{A_i} \frac{1}{T} \sum_{t=1}^{T} \left\| \hat{F}_{i,t}^{-1} - \sum_{j=1}^{N} A_{ij} \left( \hat{F}_{j,t}^{-1} - \text{id} \right) \right\|^2_{\text{Leb}}, \quad i = 1, \ldots, N.
\]

Note that, since the empirical Fréchet means of all \( \hat{\mu}_{i,t} \) are also uniform, similarly to the population setting, we do not need to consider the additional terms in the estimation problem to cancel out the unequal empirical Fréchet means. Analogously,

\[
\hat{A}_o = \hat{\Gamma}^{-1} \left[ \hat{\Gamma}(0) \right], \tag{4.4}
\]
where
\[ \hat{\Gamma}(0)_{j,l} = \frac{1}{T} \sum_{t=1}^{T} \langle \hat{F}_{j,t-1}^{-1} - id, \hat{F}_{l,t-1}^{-1} - id \rangle_{L_{Leb}} \]
and
\[ \hat{\Gamma}(1)_{j,l} = \frac{1}{T} \sum_{t=1}^{T} \langle \hat{F}_{j,t}^{-1} - id, \hat{F}_{l,t}^{-1} - id \rangle_{L_{Leb}}. \]

As before, we assume that \( \hat{\Gamma}(0) \) is invertible. Finally, we add the coefficient constraints to the problem, corresponding to the simplex constraint (A3). Therefore, the estimator \( \hat{A} \) that we finally propose is defined as
\[
\hat{A}_i = \arg \min_{A_i \in B^1_+} \frac{1}{T} \sum_{t=1}^{T} \left\| \hat{F}_{i,t}^{-1} - \sum_{j=1}^{N} A_{ij} \left( \hat{F}_{j,t-1}^{-1} - id \right) - id \right\|_{L_{Leb}}^2, \quad i = 1, \ldots, N, \quad (4.5)
\]
where \( B^1_+ \) is \( N \)-dimensional simplex, that is the nonnegative orthant of the \( \ell_1 \) unit ball \( B^1 \) in \( \mathbb{R}^N \).

Thus, an important advantage of this constraint is to promote sparsity in \( \hat{A}_i \), which will be illustrated in Section 5. The optimisation problem (4.5) can be solved by the accelerated projected gradient descent (Parikh and Boyd, 2014, Chapter 4.3). The projection onto \( B^1_+ \) is given in Thai et al. (2015).

4.2 Consistency of the estimators

Now, we study the consistency of the proposed estimator \( \hat{A} \). The main result of this section is Theorem 4.5. The details of its proof is given in the supplementary materials. Instead in this section, we resume the proof by the key intermediate results in its development.

The proof proceeds by firstly showing the consistency of the unconstrained least squares estimator \( \hat{A} \) (see Lemma 4.1) that uses the knowledge of the population Fréchet mean \( F^{-1}_{i,\Theta} \). Secondly, we show \( F^{-1}_{\mu_i \in P_{i,\Theta}} \) element-wise (see Lemma 4.2), and aim to rely on this result to prove that \( \hat{A}_0 = \hat{A} \| P \) (see Theorem 4.3). Thirdly, we show a general result (see Theorem 4.4 below) on the consistency of constrained estimators that is not restricted to the proposed estimator in this work. Finally, we apply this result to \( \hat{A}_0 \) to derive the consistency of the proposed estimator \( \hat{A} \).

Lemma 4.1. Assume that \( \mu_i, \ i = 1, \ldots, N \) satisfy Assumption A1 for \( t = 0, 1, \ldots, T \), and the transformed sequence \( \hat{F}_{i,t}^{-1}, \ t = 0, 1, \ldots, T \) satisfies Model (3.8) with Assumption A3 true. Suppose additionally that \( \hat{F}_{0}^{-1} = \pi \) with \( \pi \) the stationary distribution defined in Theorem 3.1. Given Assumptions A4 and A5 hold true, and the population matrix \( \Gamma(0) \) is nonsingular, where we recall that \( \Gamma(0) \in \mathbb{R}^{N \times N} \) is defined as
\[
[\Gamma(0)]_{j,l} = \mathbb{E} \langle \hat{F}_{j,t-1}^{-1} - id, \hat{F}_{l,t-1}^{-1} - id \rangle_{L_{Leb}},
\]
we obtain
\[
\hat{A} - A = \mathcal{O}_p \left( \frac{1}{\sqrt{T}} \right).
\]

Lemma 4.2. Under the conditions of Lemma 4.1, we have
\[
\frac{1}{T} \sum_{t=1}^{T} \hat{F}_{i,t-1}(p) - p = \mathcal{O}_p \left( \frac{1}{\sqrt{T}} \right), \quad \forall p \in (0, 1), \ i = 1, \ldots, N. \quad (4.6)
\]
Since $F_{t,t} = F_{i,t}^{-1} \otimes F_{i,t}^{-1}$, we have equivalently,
\[ F_{\mu_1}^{-1}(p) - F_{\mu_1}^{-1}(p) = O_p\left(\frac{1}{\sqrt{T}}\right), \quad \forall p \in (0, 1), \ i = 1, \ldots, N. \] (4.7)

**Theorem 4.3.** Under the conditions of Lemma 4.1
\[ \hat{A}_o - A \xrightarrow{p} 0, \]
which implies
\[ \hat{A}_o - A \xrightarrow{p} 0. \]

**Theorem 4.4.** Let $\hat{\beta}^* \in \mathbb{R}^n$ be some constant of interest. Assume $\hat{\beta}_o$ is an estimator, defined as:
\[ \hat{\beta}_o = \arg \min_{\beta \in \mathbb{R}^n} f_T(\beta) \]
which converges to $\beta^*$ in probability at $O_p(r_T)$. We define then the constrained estimator $\hat{\beta}$ as:
\[ \hat{\beta} = \arg \min_{\beta \in \mathbb{R}^n} f_T(\beta) \]
subject to:
\[ f_i(\hat{\beta}) \leq 0, \ i = 1, \ldots, m \]
\[ h_j(\hat{\beta}) = 0, \ j = 1, \ldots, p. \] (4.8)

If $f_T$ is strongly convex with a strong convexity constant $\mu_T$ satisfying $\frac{1}{\mu_T} = O_p(1)$, the strong duality holds for Problem (4.8) and $\beta^*$ satisfies the constraints, then $\hat{\beta}$ is also a consistent estimator of $\beta^*$, with $\hat{\beta} - \beta_o = O_p(r_T)$.

The result is evident from the other direction, since $\hat{\beta}_o \xrightarrow{p} \beta^*$ and $\beta^*$ satisfies the constraints, $\hat{\beta}_o$ should gradually approach the constraint set as the simple size $T$ increases. Once $\hat{\beta}_o$ itself satisfies the constraint, $\hat{\beta}$ takes $\hat{\beta}_o$, which can be only asymptotic though.

Many results establish conditions under which strong duality holds. For example when the objective function is convex with an open domain and the constraints $f_i$, $i = 1, \ldots, m$, $h_j$, $j = 1, \ldots, p$ are all affine, then strong duality holds as soon as the problem is feasible (Boyd et al., 2004, Section 5.2.3), which is the case of Problem (4.8). Note that, the problem can be represented in vector form as
\[ \hat{A}_i = \arg \min_{A_i \in B^+_i} \frac{1}{2} A_i \left[ \hat{\Gamma}(0) \right] A_i^\top - \left[ \hat{\Gamma}(1) \right]_{i,i}, \ i = 1, \ldots, N. \] (4.9)

Since Theorem 4.3 shows the constraint free least square estimator on the approximate series converges to $A$ in probability, applying Theorem 4.4 on Theorem 4.3, we finally obtain the desired consistency result, given in Theorem 4.5.

**Theorem 4.5.** Under the conditions of Lemma 4.1, given the true coefficient $A$ satisfies Assumption $A3$, namely, $A_i \in B^+_i$, $i = 1, \ldots, N$, we have
\[ \hat{A} - A \xrightarrow{p} 0. \]

## 5 Numerical experiments

In Section 5.1, we firstly demonstrate the consistency result of the proposed estimator using synthetic data. Then, we fit the model on two real data sets in Section 5.2 and Section 5.3. For each data set, the estimated coefficient matrix $\hat{A}$ allows us to understand the dependency structure of the multivariate distributional time series. In particular, we visualize the learned structure between features using a directed weighted graph with adjacency matrix $\hat{A}$. 

14
5.1 Simulations

5.1.1 Generation of the synthetic data

We firstly propose a mechanism to generate the valid random distortion functions. To this end, we consider the random functions defined by

\[ \epsilon_g = \frac{1 + \xi}{2} g \circ h^{-1} + \frac{1 - \xi}{2} h^{-1}, \]  

(5.1)

where \( g \) is a non-decreasing right-continuous constant function from \([0, 1]\) to \([0, 1]\), \( h^{-1} \) is the left continuous inverse of \( h = \frac{1}{2}(g + id) \), and \( \xi \sim U(-1, 1) \) is a random variable. For any given function \( g \), we can sample a family of distortion functions \( \epsilon_{i,t} \), when sampling \( \xi_{i,t} \sim U(-1, 1) \). This construction of random distortion functions is inspired by the one proposed in Zhu and Müller (2021, Equation (13)), however, we have modified their construction of \( h \) and of the random coefficients. It is easy to verify that

\[ \mathbb{E} [\epsilon_g] = \frac{1}{2} (g \circ h^{-1} + h^{-1}) + \frac{1}{2} (g + id) \circ h^{-1} = id. \]

To make \( \epsilon_{i,t} \) satisfy additionally Model (3.4) and Assumption A4, we require \( g \) to be furthermore continuous and differentiable. Then on the one hand, since \( g \) is continuous and non-decreasing, any generated \( \epsilon_g \) is non-decreasing and left-continuous. On the other hand, note that

\[ [h^{-1}]' = \frac{1}{h' \circ h^{-1}} = \frac{1}{\frac{1}{2}(g' + 1) \circ h^{-1}} = \frac{2}{g' \circ h^{-1} + 1}. \]

Thus, we have

\[ \epsilon_g' = \frac{1 + \xi}{2} (g' \circ h^{-1}) \frac{2}{g' \circ h^{-1} + 1} + \frac{1 - \xi}{2} \frac{2}{g' \circ h^{-1} + 1} \]

\[ = \left( \frac{1 + \xi}{2} g' \circ h^{-1} + \frac{1 - \xi}{2} \right) \frac{2}{g' \circ h^{-1} + 1} \]

\[ = 1 + \xi - \xi \frac{2}{g' \circ h^{-1} + 1} = 1 + \xi \left( 1 - \frac{2}{g' \circ h^{-1} + 1} \right). \]

This implies

\[ |\epsilon_g'| \leq 1 + |\xi| \left| 1 - \frac{2}{g' \circ h^{-1} + 1} \right| \leq 2. \]

The bound comes from \( \xi \sim U(-1, 1) \) and \( g' \geq 0 \), which is hence tight. Thus any \( \epsilon_g \) generated by Formula (5.1) is Lipschitz continuous, with the constant uniformly bounded by 2 over \( \xi \). Note that Assumption A4 requires the Lipschitz continuity only in expectation. Thus, the i.i.d. samples \( \epsilon_{i,t} \) of any \( \epsilon_g \) satisfy obviously Assumption A4 with the largest \( L = 2 \). Figure 3 shows the function \( g \) used in the simulation and one realization of 30 i.i.d. samples of the resulting \( \epsilon_g \).

Secondly, we present the procedure to generate the true coefficient matrix \( A \). We first generate a sparse matrix with the weights all positive in a random way, denoted by \( A^0 \). We then normalize each row of \( A^0 \) by the row sum to fulfill Assumptions A3. We denote this last matrix still by \( A^0 \). Based on the previous mechanism for the random distortion function, we take \( L = 2 \) in Assumption A5. Lastly, we scale down \( A^0 \) by \( (2 + \alpha)\|A^0\|_2 \) to obtain a valid \( A \). We test two values \( \alpha = 0.1 \) and \( \alpha = 0.5 \) in our experiments.

Given a valid matrix of coefficients \( A \) and samples of \( \epsilon_{i,t} \), we can then generate the “centered” quantile functions \( \tilde{F}_{i,t}^{-1} \) from Model (3.5). Note that, \( \tilde{F}_{i,t}^{-1} \) are only the simulations of the transformed

---
Figure 3: The function $g$ on the left is given by the natural cubic spline passing through the points $(0,0), (0.2,0.1), (0.6,0.2), (1,1)$. On the right is one realization of 30 i.i.d. samples of the resulting $\epsilon_g$.

data. Thus, we have to generate furthermore the population Fréchet mean $F^{-1}_{i; \Theta}$ of each univariate series in order to finally obtain the synthesized “raw” data, as the inverse of transformation (3.2):

$$F^{-1}_{i,t} = \tilde{F}^{-1}_{i,t} \oplus F^{-1}_{i; \Theta} := \tilde{F}^{-1}_{i,t} \circ F^{-1}_{i; \Theta}.$$  

We set $F^{-1}_{i; \Theta}$ as the natural cubic spline of the points: $(0,0), (0.2,0.1), (0.6,0.2 + 0.2i/N), (1,1), i = 1, ..., N$. The empirical Fréchet mean $F^{-1}_{\hat{P}_i}$ and the proposed estimator $\hat{A}$ are calculated on the synthesized “raw” data $F^{-1}_{i,t}$. In Section 5.1.2, we aim to demonstrate the consistency result given in Theorem 4.5 with the synthetic data.

5.1.2 Experiment settings and results

In this experiment, we demonstrate the consistency of the proposed estimator $\hat{A}$ for two different values $N = 10$ and $N = 100$. For each $N$, we generate two true matrices $A$ for $\alpha = 0.1$ and 0.5 respectively, according to the procedure presented in Section 5.1.1. With each $A$, we calculate the root mean square deviation (RMSD) successively

$$\frac{\|\hat{A} - A\|_F}{\|A\|_F}$$

with the synthetic data that it generates along time. To furthermore study the mean and the variance of the RMSD (5.2), we run 100 independent simulations for the same $A$.

Note that the value of $\hat{A}$ we use in Equation (5.2) is the approximation obtained by the projected gradient descent applied to Problem (4.5). Thus the corresponding approximation error also accounts for the deviation which is on the order of the threshold we set in the stopping criteria. For all values of $N$, we use the same error threshold. We stop the algorithm as soon as the difference between the previous and the current updates in $\ell_2$ norm reaches 0.0001, for the resolution of each row $\hat{A}_{i,t}$.

We firstly show the evolution of RMSD for $N = 10, 100$ in Figures 4 and 5, respectively. We can see that, all means and variances of the RMSD decrease towards zero as the sample size $T$ increases, for each $N$ and $\alpha$ value. This demonstrates empirically that, when the model assumptions $A3, A4$ and $A5$ hold true for the data, the proposed estimator $\hat{A}$ converges to $A$ in probability, which is implied actually by the convergence of $\hat{A}$ to $A$ in $L^2$. 

16
Additionally, we can notice that, the RMSD for $\alpha = 0.1$ which corresponds to larger $\ell_2$ norm of $A$ has a smaller mean in both cases, and also a smaller variance for most of the sample sizes $T$ investigated.

Also we would like to remark that, during the first few $T$ values, the samples are insufficient for a meaningful estimation. Thus the projected gradient descent will terminate rapidly, as shown in the very beginning of two subfigures in Figure 6. The output $\hat{A}$ will be a zero matrix, since we initialize $\hat{A}$ as zero. This results in the low RMSD values during the early phase, since the true $A$ is generated as a sparse matrix with small weights, as shown in Figures 4 and 5. As $T$ increases, more entries of $\hat{A}$ become non-zero, which brings to the growth of RMSD. Upon the arrival of more new samples, $\hat{A}$ starts to converges, meanwhile the RMSD starts to decrease accordingly.

Lastly, we show in Figure 6 the complete execution time of the model fitting on the raw data with respect to the sample size $T$. We can see that the execution time increases linearly with respect to $T$, and $A$ with the smaller $\ell_2$ norm requires sightly less time ($\alpha = 0.5$) than the other. The linear increase comes mainly from the loop over time $t = 1,\ldots,T$ in calculating the empirical Fréchet mean (4.2) and in calculating the matrices $\hat{\Gamma}(0),\hat{\Gamma}(1)$ by their formulas in Equation (4.4). The running time of these calculations is determined by the granularity in the numerical methods to approximate the function composition, function inverse, and the inner product. The granularity applied during this
Figure 6: Calculation time (in seconds) of $\hat{A}$ with respect to the sample size $T$ for $N = 10$ (left) and $N = 100$ (right). The calculation time counting starts from the computation of the empirical Fréchet means for Data transformation (4.3), and ends when the accelerated projected gradient descent of Problem (4.5) finishes for the last row $i = N$.

simulation is 0.01, that is we input/output only the quantile function values at grid 0, 0.01, ..., 0.99 to/from each numerical approximation.

5.2 Age distribution of countries

We firstly test the proposed model with the data set illustrated in Figure 1 in the introduction. These data are from the US Census Bureau’s International Data Base\(^1\), which provides the population estimates and projections for countries and areas by single year of age, over years. We would like to apply the proposed model on this international age distribution data to learn about the links among the changes in the age structures of different countries. Specially, we consider the countries and the micro-states in the European Union and/or Schengen Area. Because the corresponding data used during the model fitting starts in the 1990s, we also include the former European Union member United Kingdom. Note that, Vatican City is not included since it is not available in the data base. Therefore, the list of 34 countries in this study is: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Monaco, Netherlands, Norway, Poland, Portugal, Romania, San Marino, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom.

Time-wise, we consider the 40 years between 1996 to 2035. 1996 is the earliest year for which the data for all the considered countries is available.

To apply Model (3.4), we firstly represent the distribution of age population, of country $i$, at year $T$, by $\mu_i^T$, with $T = 1, ..., 40$ and $i = 1, ..., 34$. Note that the age considered by the data base goes through 0 to 100-plus. Thus we take the 100-plus as 100, and moreover scale down the age by 100 to make the age distribution supported over $[0, 1]$. Then we retrieve the quantile function $F_{i,t}^{-1}$ of $\mu_i^T$ from the population counts by ages of country $i$ recorded at year $T$, with the numeric methods. In particular, we retrieve the quantile functions using continuous functions which take 0 on $p = 0$ and 1 on $p = 1$ so as to be consistent with Assumption A2 (for details see function \texttt{generate_qt_fun} defined in script \texttt{age_pop.py} in the code related to this paper).

\(^1\)The data base in open access through \url{https://www.census.gov/data/developers/data-sets/international-database.html}. 
We fit Model (3.5) on the retrieved functions $F_{i,t}^{-1}$, $T = 1, ..., 40, i = 1, ..., 34$. We use the same stopping criteria as in the simulation, while we apply the granularity of 0.002. The complete execution time of model fitting takes around 78 seconds. Figure 7 shows the inferred coefficient $A$ visualized as a directed weighted graph, on the geographical map.

![Inferred age structure graph](image)

Figure 7: Inferred age structure graph. The non-zero coefficients $A_{ij}$ are represented by the weighted directed edges from node $j$ to node $i$. Thicker arrow corresponds to larger weights. The blue circles around nodes represent the weights of self-loop.

Firstly, we can notice that for all countries $i \in \{1, ..., 34\}$, the weight of self-loop $A_{ii}$ dominates the weights of incoming edges $A_{ij}, j = 1, ..., 34$, which are bounded by $0 \leq \sum_{j=1}^{34} A_{ij} \leq 1$. This is because the age structure of a country does not change much from one year to another. On the other hand, this also implies the age structure differs largely across countries. Nevertheless, there are still significant links between countries’ age distribution. The first two largest weights excluding all the self-loops are respectively on the edges: Estonia $\rightarrow$ Latvia, and Sweden $\rightarrow$ Norway. To justify the inferred edges, we plot the evolution of age structure of these four countries in Figures 8 and 9.

We can see that within these four countries, the age structures between the linked countries are similar along time; by contrast, the structures between the unlinked countries are very different. Indeed, these two linked pairs consist both of the countries which share long distances along their borders. Thus, generally, the inferred edges in Figure 7 indicate the similarity of the age structures between countries from 1996 to 2036. Moreover, the directions of the edges imply, at the model level, that, when age structures in the outward countries (for example, Estonia, Sweden) change, it will induce relative changes in the inward countries (respectively, Latvia, Norway). These numeric findings can be furthermore explained in demography or not. On the other, we are interested in the
Figure 8: Evolution of age structure from 1996 to 2036 (projected) of Estonia (left) versus Latvia (right). Each curve connects the 101 relative frequencies from 0, 1/100, 2/100, ..., 1, which represents the age structure of a considered year. Lighter curves correspond to more recent years.

Figure 9: Evolution of age structure from 1996 to 2036 (projected) of Sweden (left) versus Norway (right).
neighbouring countries which are not linked. We verify for example the age structures of France, Italy, in Figure 10. We can see that the age structures are as expected very different.

Figure 10: Evolution of age structure from 1996 to 2036 (projected) of France (left) versus Italy (right).

All these observations strongly support the usefulness of our model. Lastly, in Table 1, we provide the first 5 edges of the largest weights.

| From     | To     |
|----------|--------|
| 1 Estonia| Latvia |
| 2 Sweden | Norway |
| 3 Belgium| Germany|
| 4 Finland| Netherlands |
| 5 France | Greece |

Table 1: Top 5 edges with the largest weights excluding all the self-loops

5.3 Bike-sharing network in Paris

Secondly, we test our model on the bike-sharing data set of Paris from Jiang et al. (2020). The data set records the ratio of available bikes of 274 stations, observed over 4417 consecutive hours. We are interested in the temporal evolution of this bike availability of stations, and would like to identify how the stations relate mutually in their evolution using the proposed method. To this end, we firstly represent the data by taking into account the distribution aspect. We introduce the temporal variable $T$ which represents hours in a day. Accordingly, the distribution $\mu_i$ considered by Model (3.4) represents the distribution of the bike availability of the station $i$, at hour $T$ in a day, with $T = 1, ..., 24$ and $i = 1, ..., 274$. Thus the quantile function $F_{i,t}^{-1}$ of distribution $\mu_i$ can be retrieved from all the data points of station $i$ recorded at hour $T$, by the numeric methods. We use the similar method as for the age distribution data to retrieve the valid quantile functions, however the implementation is different due to the availability of different quantity types (for details see function generate_qt_fun defined in script bike_net.py in the code related to this paper).
Figure 11: Subgraph 1 (50 nodes chosen randomly). For graphical meanings see the caption of Figure 7.

Note that since these observations are distanced from each other in time, they can be considered approximately as being independent samples. For more comments on this point, we refer to Remark 5.1. We then fit Model (3.5) on data $F^{-1}_{i,t}, T = 1, ..., 24, i = 1, ..., 274$. We use the same stopping criteria as previously, and we apply the granularity of 0.002. The complete execution time of model fitting takes around 20 minutes. We now demonstrate the visualization of the inferred matrix of coefficients $A$ on the map of Paris, represented by the directed weighted graph. Since the edges of the complete graph will be densely located in the plot when fitting the graph to the paper size, for better visual effects, we show two subgraphs each of 50 nodes, in Figure 11 and Figure 12 respectively. The complete graph and the subgraphs are available in interactive form at https://github.com/yiyej/Wasserstein_Multivariate_Autoregressive_Model.

These figures show some interesting links between the bike utilisation in different areas of Paris. For example, we can notice in Figure 11 that the station 41604-lagny-saint-mande is much more useful in predicting the stations along the flow from Saint-Mandé in southeast to Neuilly-sur-Seine in northwest. This flow actually goes along the main itinerary of Metro 1 and RER A in Paris, which starts around Saint-Mandé. Additionally, we can see in Figure 12 that the stations near Charles de Gaulle - étoile (upper left) need to be predicted by numerous stations together, among which is the station 02012-quatre-septembre. By contrast, station 02012-quatre-septembre can only be predicted by itself with no edges pointed in. These observations all support the effectiveness of the proposed model and its estimation method.

Remark 5.1. Note that another way to represent the data set by distributions is to consider every $K$ consecutive observations as the i.i.d. samples of a distribution. Thus $T = 1, 2, ..., 185/K$ represents the ongoing time in hours. However, the consecutive observations are highly correlated. Moreover, the distribution retrieved in this manner does not clearly represent a random variable. Thus we do not go further with this data representation method.
6 Conclusion

In this paper, we extend the standard VAR models to distributional multivariate AR models, which provides an approach to model a collection of multiple time-dependent probability measures, and to represent their dependency structure by a directed weighted graph at the same time. Especially, the proposed data centering method for random measures allows the development of auto-regressive and regressive models with multiple predictors. Moreover, the empirical studies on the real data sets demonstrate that, the proposed models equipped with the distributional data representation are the efficient tools for analyzing and understanding the spatial-temporal data. For future research directions, this paper provides a class of multivariate AR models for distributional time series which favors the graph learning. More classes which suit the different data analysis purposes are expected to explore.

Acknowledgements

I would like to express my gratitude to my supervisors Jérémie Bigot and Sofian Maabout for their support in this work. I would like to thank Jérémie for introducing me this interesting problem and for the helpful discussions. I would also like to thank Sofian for reading this article and suggesting improvements in the writing. Additionally, I would like to thank Alex Petersen for the inspiring question on my PhD defense related to this work which helped me realize the result in Proposition 3.4 Point 1. Lastly, I would like to thank my colleague Paul Freulon for the discussion over the interpretation of the graphical results of the Paris network.
References

L. Ambrosio, N. Gigli, and G. Savaré. Gradient flows with metric and differentiable structures, and applications to the Wasserstein space. *Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei. Matematica e Applicazioni*, 15(3-4):327–343, 2004.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

J. Bigot. Statistical data analysis in the Wasserstein space. *ESAIM: ProcS*, 68:1–19, 2020.

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

S. Bobkov and M. Ledoux. One-dimensional empirical measures, order statistics and Kantorovich transport distances. *preprint*, 7127347, 2014.

D. Bosq. *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media, 2000.

S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

E. Cazelles, V. Seguy, J. Bigot, M. Cuturi, and N. Papadakis. Log-PCA versus geodesic PCA of histograms in the Wasserstein space. *arXiv preprint arXiv:1708.08143*, 2017.

Y. Chen, Z. Lin, and H.-G. Müller. Wasserstein regression. *Journal of the American Statistical Association*, pages 1–14, 2021.

Z. Chen, Y. Bao, H. Li, and B. F. Spencer Jr. LQD-RKHS-based distribution-to-distribution regression methodology for restoring the probability distributions of missing shm data. *Mechanical Systems and Signal Processing*, 121:655–674, 2019.

L. Ghodrati and V. M. Panaretos. Distribution-on-distribution regression via optimal transport maps. *Biometrika*, 01 2022. asac005.

L. Helmut. *New introduction to multiple time series analysis*. Springer Berlin Heidelberg, 2005.

Y. Jiang, J. Bigot, and S. Maabout. Sensor selection on graphs via data-driven node sub-sampling in network time series. *arXiv preprint arXiv:2004.11815*, 2020.

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

S. Mazzuco and B. Scarpa. Fitting age-specific fertility rates by a flexible generalized skew normal probability density function. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):187–203, 2015.

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

V. M. Panaretos and Y. Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.
N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

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

A. Petersen and H.-G. Müller. Fréchet regression for random objects with euclidean predictors. *The Annals of Statistics*, 47(2):691–719, 2019a.

A. Petersen and H.-G. Müller. Wasserstein covariance for multiple random densities. *Biometrika*, 106(2):339–351, 2019b.

A. Petersen, C. Zhang, and P. Kokoszka. Modeling Probability Density Functions as Data Objects. *Econometrics and Statistics*, 21(C):159–178, 2022.

H. L. Shang and S. Haberman. Forecasting age distribution of death counts: An application to annuity pricing. *Annals of Actuarial Science*, 14(1):150–169, 2020.

J. Thai, C. Wu, A. Pozdnukhov, and A. Bayen. Projected sub-gradient with $\ell_1$ or simplex constraints via isotonic regression. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 2031–2036. IEEE, 2015.

C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annu. Rev. Statist*, 1:41, 2015.

W. B. Wu and X. Shao. Limit theorems for iterated random functions. *Journal of Applied Probability*, 41(2):425–436, 2004.

C. Zhang, P. Kokoszka, and A. Petersen. Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis*, 2021.

X. Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.

C. Zhu and H.-G. Müller. Autoregressive optimal transport models. *arXiv preprint arXiv:2105.05439*, 2021.

### 7 Supplementary Material

#### 7.1 Time series in metric space

Let $(\mathcal{X}, d)$ be a complete separate metric space with Borel set $X$, an iterated random function (IRF) system in the state space $(\mathcal{X}, d)$ is defined as

$$X_t = \Phi_{\epsilon_t}(X_{t-1}), \quad t \in \mathbb{Z}, \quad (7.1)$$

where $\epsilon_t, t \in \mathbb{Z}$ are i.i.d. random objects taking values in a measurable space $\Theta$, $\Phi_{\epsilon}(\cdot) := \Phi(\cdot, \epsilon)$ is the $\epsilon$-section of a jointly measurable function $\Phi : \mathcal{X} \times \Theta \to \mathcal{X}$. Note that $X_t, t \in \mathbb{Z}$ can also be seen as a $\mathcal{X}$-valued nonlinear auto-regressive process.

We now define $\hat{\Phi}_{t,m} := \Phi_{\epsilon_t} \circ \Phi_{\epsilon_{t-1}} \circ \ldots \circ \Phi_{\epsilon_{t-(m-1)}}$ the same way as Theorem 1 in Zhu and Müller (2021), and we recall Conditions 1 and 2 from Zhu and Müller (2021) as follows.

25
Condition 1. (Condition 1 in Wu and Shao (2004)) There exists $Y^0 \in \mathcal{X}$ and $\alpha > 0$, such that
$$I(\alpha, Y^0) := \mathbb{E}d^n(Y^0, \Phi_t(Y^0)) < \infty.$$  

Condition 2. (Theorem 1 in Zhu and Müller (2021); Condition 2 in Wu and Shao (2004)) There exists $X^0 \in \mathcal{X}$, $\alpha > 0$, $r = r(\alpha) \in (0, 1)$, and $C = C(\alpha) > 0$, such that for all $t \in \mathbb{Z}$, we have
$$\mathbb{E}d^n(\Phi_{t,m}(X^0), \Phi_{t,m}(X)) \leq C r^m d^n(X^0, X), \quad \forall X, m \in \mathbb{N}. \quad (7.2)$$

Note that $\epsilon_t, t \in \mathbb{Z}$ are i.i.d, thus for any fixed $X \in \mathcal{X}$, we have $\Phi_{t,m}(X) \overset{d}{=} \Phi_{t,m}(X), \forall t, t' \in \mathbb{Z}$. Thus Condition 2 is not a uniform requirement imposed for $t \in \mathbb{Z}$. We recall firstly in Lemma 7.1 and Lemma 7.2 below, the stability results given in Wu and Shao (2004, Theorem 2). Note that even though Theorem 1 in Zhu and Müller (2021) gives the stability result only meant to their particular IRF system, defined in the space of optimal transports $\mathcal{T}$, their corresponding proof does not rely on the particular IRF system, which thus implies the stability of the general system (7.1) as well. More importantly, their proof is presented from the viewpoint of auto-regressive process by contrast to the one in Wu and Shao (2004). Since we rely on their construction of $\Phi_{t,m}$, here we refer to their work as well.

We thus consolidate the proofs of Wu and Shao (2004) and Zhu and Müller (2021) to more detailed ones given in Section 7.3.1 and Section 7.3.2.

**Lemma 7.1.** (Wu and Shao (2004, Theorem 2); Zhu and Müller (2021, Theorem 1)) Assuming that Conditions 1 and 2 hold, it follows that, for any fixed $t \in \mathbb{Z}$, $\lim_{m \to \infty} \Phi_{t,m}(X^0)$ exists almost surely and is denoted by $\tilde{X}_t$. Moreover, for any fixed $t, t' \in \mathbb{Z}$, $\tilde{X}_{t} \overset{d}{=} \tilde{X}_{t'}$. We denote this time-invariant marginal distribution as $\pi$.

**Lemma 7.2.** (Wu and Shao (2004, Theorem 2); Zhu and Müller (2021, Theorem 1)) Assuming that Conditions 1 and 2 hold, then the limits in Lemma 7.1 do not depend on the departure point $X^0$, that is, for any fixed $X \in \mathcal{X}$ and any fixed $t \in \mathbb{Z}$, $\Phi_{t,m}(X) \overset{m \to \infty}{\rightarrow} \tilde{X}_t$, in $d$, almost surely.

The random process $\tilde{X}_t, t \in \mathbb{Z}$ is then a solution to the IRF system (7.1). This result is also indicated in Zhu and Müller (2021, Theorem 1), however, they do not elaborate the proof. We thus provide a proof in Section 7.3.3.

**Lemma 7.3.** (Zhu and Müller (2021, Theorem 1)) Suppose that Conditions 1 and 2 hold, then $\tilde{X}_t = \lim_{m \to \infty} \Phi_{t,m}(X^0), t \in \mathbb{Z}$ is a solution of IRF system (7.1), almost surely.

We gather all the above results in Theorem 7.4.

**Theorem 7.4.** (Existence; Wu and Shao (2004, Theorem 2); Zhu and Müller (2021, Theorem 1)) Suppose that Conditions 1 and 2 hold, then the IRF system (7.1) almost surely admits a solution $\tilde{X}_t, t \in \mathbb{Z}$, with the same marginal distribution $\pi$.

The uniqueness result provided in Zhu and Müller (2021, Theorem 1) is no longer valid for the general system (7.1), which is also mentioned in Wu and Shao (2004, Theorem 2). Thus, we provide the uniqueness result for the general system in Theorem 7.5. For the proof we refer to Section 7.3.4.

**Theorem 7.5.** (Uniqueness) Suppose that Conditions 1 and 2 hold, then, if there is another solution $S_t, t \in \mathbb{Z}$, such that
$$\mathbb{E}[d^\beta(S_t, Z^0)] < M, \quad t \in \mathbb{Z}, \quad (7.3)$$
for some $M, \beta > 0$, and some $Z^0 \in \mathcal{X}$ (thus for all), then for all $t \in \mathbb{Z}$
$$\tilde{X}_t = S_t, \quad \text{in } d, \quad \text{almost surely.}$$
Since \( \bar{X}_t, t \in \mathbb{Z} \) is the unique solution of IRF system (7.1), we hereafter denote this solution directly by \( X_t, t \in \mathbb{Z} \). Lastly, we recall the geometric moment contraction result given in Wu and Shao (2004, Theorem 2), which will be used later on in the proof of consistency of the estimators. We reproduce their proof using our notation, in Section 7.3.5.

**Proposition 7.6.** (Wu and Shao (2004, Theorem 2)) the IRF system (7.1) is geometric moment contracting in the following sense: let \( X \sim \pi \) be independent of \( X^1 \sim \pi \), where \( \pi \) is the shared marginal distribution in Theorem 7.4, \( X, X^1 \) are independent of \( \epsilon_m, m \geq 1 \). Let \( X_m(X), X_m(X^1), m \geq 1 \), denote the sequences generated by the model (7.1) starting respectively from \( X, X^1 \). Then, for all \( m \geq 1 \), there exist constants \( D > 0, s \in (0, 1) \), such that

\[
\mathbb{E} d^s(X_m(X), X_m(X^1)) \leq D s^m. \tag{7.4}
\]

### 7.2 Time series in Hilbert space

Now, we study the stationarity of the functional time series \( X_t, t \in \mathbb{Z} \). Theorem 7.4 has already implied that \( X_t \) is stationary in the sense that \( X_t \overset{d}{=} \pi \), \( t \in \mathbb{Z} \). However, the stationarity in weak sense for time series requires additionally the auto-covariance to be time-invariant. To this end, we need furthermore a Hilbert structure in \( \mathcal{X} \) to be able to define the notion of covariance between two random objects. Thus, we suppose \( (\mathcal{X}, d) \) is furthermore a Hilbert space, whose inner product and the induced norm are denoted respectively by \( \langle \cdot, \cdot \rangle \) and \( \| \cdot \| \). To make the previous results presented in the metric space \( (\mathcal{X}, d) \) valid as the results for the Hilbert space \( (\mathcal{X}, \langle \cdot, \cdot \rangle) \), we assume \( (\mathcal{X}, d) \) is indeed the induced metric space of \( (\mathcal{X}, \langle \cdot, \cdot \rangle) \), namely, \( d(X, Y) = \| X - Y \|, X, Y \in \mathcal{X} \).

We can now give the stationarity result in Theorem 7.7. The proof can be found in Section 7.3.6.

**Theorem 7.7.** Suppose that Conditions 1 and 2 hold with \( \alpha \geq 2 \), then the unique solution \( X_t, t \in \mathbb{Z} \) given in Theorems 7.4 and 7.5 is stationary in \( (\mathcal{X}, \langle \cdot, \cdot \rangle) \) in the sense of Definition 3.1.

The condition \( \alpha \geq 2 \) is to ensure the existence of a second moment of \( X_t \), from which we can define the auto-covariance operators of the time series. Note that for any \( t \in \mathbb{Z} \)

\[
\mathbb{E} \| X_t \|^2 \leq \mathbb{E} \| X_t - \Phi_{\epsilon_t}(X^0) \|^2 + \mathbb{E} \| \Phi_{\epsilon_t}(X^0) - \Phi_{\epsilon_t}(Y^0) \|^2 + \mathbb{E} \| \Phi_{\epsilon_t}(Y^0) - Y^0 \|^2 + \mathbb{E} \| Y^0 \|^2
\]

\[
= \mathbb{E} d^2(X_t, \Phi_{\epsilon_t}(X^0)) + \mathbb{E} d^2(\Phi_{\epsilon_t}(X^0), \Phi_{\epsilon_t}(Y^0)) + \mathbb{E} d^2(\Phi_{\epsilon_t}(Y^0), Y^0) + \| Y^0 \|^2
\]

\[
\overset{(a)}{\leq} (\mathbb{E} \| \Phi_{\epsilon_t}(X^0) \|^2)^{\frac{\alpha}{2}} + (\mathbb{E} \| \Phi_{\epsilon_t}(X^0) - \Phi_{\epsilon_t}(Y^0) \|^2)^{\frac{\alpha}{2}} + (\mathbb{E} \| \Phi_{\epsilon_t}(Y^0) - Y^0 \|^2)^{\frac{\alpha}{2}} + \| Y^0 \|^2
\]

\[
\overset{(b)}{\leq} (\mathbb{E} \| \Phi_{\epsilon_t}(X^0) \|^2)^{\frac{\alpha}{2}} + (\mathbb{E} \| \Phi_{\epsilon_t}(X^0) - \Phi_{\epsilon_t}(Y^0) \|^2)^{\frac{\alpha}{2}} + (\mathbb{E} \| \Phi_{\epsilon_t}(Y^0) - Y^0 \|^2)^{\frac{\alpha}{2}} + \| Y^0 \|^2.
\]

Inequality (a) comes from the condition \( \alpha \geq 2 \) and Jensen’s inequality. Inequality (b) comes from Conditions 1 and 2. Thus, by the upperbound (7.5), we have \( \mathbb{E} \| X_t \|^2 < \infty \) is uniformly bounded over \( t \in \mathbb{Z} \).

### 7.3 Proofs

#### 7.3.1 Proof of Lemma 7.1

We are first to show that

\[
I(\alpha, X^0) := \mathbb{E} d^\alpha(X^0, \Phi_{\epsilon_t}(X^0)) < \infty.
\]
We have
\[ I(\alpha, X^0) \leq \mathbb{E}d^\alpha(X^0, Y^0) + \mathbb{E}d^\alpha(Y^0, \Phi_{\epsilon_t}(X^0)) \]
\[ \leq \mathbb{E}d^\alpha(X^0, Y^0) + \mathbb{E}d^\alpha(Y^0, \Phi_{\epsilon_t}(Y^0)) + \mathbb{E}d^\alpha(\Phi_{\epsilon_t}(Y^0), \Phi_{\epsilon_t}(X^0)) \]
\[ \overset{(a)}{\leq} d^\alpha(X^0, Y^0) + I(\alpha, Y^0) + C r d^\alpha(Y^0, X^0) < \infty. \]

Inequality (a) comes from Condition 2 applied to \( \mathbb{E}d^\alpha(\Phi_{\epsilon_t}(Y^0), \Phi_{\epsilon_t}(X^0)) \). Then, by Inequality (7.2), we have for all \( m \in \mathbb{N} \)
\[ \mathbb{E}\left[d^\alpha(\tilde{\Phi}_{t,m}(X^0), \tilde{\Phi}_{t,m+1}(X^0)) \right] = \mathbb{E}\left[d^\alpha(\tilde{\Phi}_{t,m}(X^0), \tilde{\Phi}_{t,m}(\Phi_{\epsilon_{t-m}}(X^0))) \right] \]
\[ \overset{\text{Condition 2}}{\leq} C r^m \mathbb{E}\left[d^\alpha(X^0, \Phi_{\epsilon_{t-m}}(X^0)) \right] = I(\alpha, X^0) C r^m. \]

Then by the Markov inequality, we have
\[ \mathbb{P}\left[d^\alpha(\tilde{\Phi}_{t,m}(X^0), \tilde{\Phi}_{t,m+1}(X^0)) \geq r^{-1/2} \right] \leq r^{-1/2} \mathbb{E}\left[d^\alpha(\tilde{\Phi}_{t,m}(X^0), \tilde{\Phi}_{t,m+1}(X^0)) \right] \leq r^{-1/2}. \]

Thus
\[ \sum_{m=1}^{\infty} \mathbb{P}\left[d(\tilde{\Phi}_{t,m}(X^0), \tilde{\Phi}_{t,m+1}(X^0)) \geq r^{-1/2} \right] \leq \sum_{m=1}^{\infty} r^{-1/2} = \frac{r^2}{1 - r} < \infty. \]

Applying the Borel-Cantelli lemma, we have
\[ \mathbb{P}\left[d(\tilde{\Phi}_{t,m}(X^0), \tilde{\Phi}_{t,m+1}(X^0)) \geq r^{-1/2} \text{ infinitely often} \right] = 0. \]

Thus, we have that
\[ \mathbb{P}\left[d(\tilde{\Phi}_{t,m}(X^0), \tilde{\Phi}_{t,m+1}(X^0)) \geq r^{-1/2} \text{ happen finitely times} \right] = 1. \]

This implies
\[ \mathbb{P}\left[d(\tilde{\Phi}_{t,m}(X^0), \tilde{\Phi}_{t,m+1}(X^0)) \xrightarrow{m-\infty} 0 \right] = 1. \]

Thus \( \tilde{\Phi}_{t,m}(X^0) \) is a Cauchy sequence in \((\mathcal{X}, d)\), almost surely. By the completeness of \((\mathcal{X}, d)\), there exists a \( \tilde{X}_t \in \mathcal{X} \), such that \( \tilde{\Phi}_{t,m}(X^0) \xrightarrow{m-\infty} \tilde{X}_t \) almost surely. Moreover, since for any fixed \( t, t' \in \mathbb{Z} \), \( \tilde{\Phi}_{t,m}(X^0) \xrightarrow{d} \tilde{\Phi}_{t',m}(X^0), \forall m \in \mathbb{N} \) Thus \( \lim_{m-\infty} \tilde{\Phi}_{t,m}(X^0) \xrightarrow{d} \tilde{\Phi}_{t',m}(X^0) \), namely \( \tilde{X}_t \xrightarrow{d} \tilde{X}_t' \), almost surely.

### 7.3.2 Proof of Lemma 7.2

Under the conditions of Lemma 7.1, we have \( d(\tilde{\Phi}_{t,m}(X^0), \tilde{X}_t) \xrightarrow{m-\infty} 0 \) almost surely, with \( \tilde{X}_t \in (\mathcal{X}, d), \forall m \in \mathbb{N}, t \in \mathbb{Z} \). Thus \( \mathbb{E}d^\alpha(\tilde{\Phi}_{t,m}(X^0), \tilde{X}_t) \xrightarrow{m-\infty} 0 \). On the other hand, Since \( \mathbb{E}d^\alpha(\tilde{\Phi}_{t,m+j}(X^0), \tilde{\Phi}_{t,m+j+1}(X^0)) \leq r^{m+j} \) for any fixed \( m > 0, j \geq 0 \), then
\[ \mathbb{E} \sum_{j=0}^{m} d^\alpha(\tilde{\Phi}_{t,m+j}(X^0), \tilde{\Phi}_{t,m+j+1}(X^0)) \leq \sum_{j=0}^{m} r^{m+j} \leq \sum_{j=0}^{m} r^{m+j} = \frac{r^m}{1 - r}. \]
Thus, for any \( m, n \in \mathbb{N} \),
\[
\mathbb{E}d^\alpha(\tilde{\Phi}_{t,m}(X^0), \tilde{X}_t) \leq \mathbb{E} \sum_{j=0}^n d^\alpha(\tilde{\Phi}_{t,m+j}(X^0), \tilde{\Phi}_{t,m+j+1}(X^0))
\]
\[
+ \mathbb{E}d^\alpha(\tilde{\Phi}_{t,m+n+1}(X^0), \tilde{X}_t)
\]
\[
\lesssim r^m.
\] (7.5) Then, for any fixed \( m \in \mathbb{N} \), \( X \in \mathcal{X} \), we have
\[
d^\alpha(\tilde{\Phi}_{t,m}(X), \tilde{X}_t) \leq d^\alpha(\tilde{\Phi}_{t,m}(X), \tilde{\Phi}_{t,m}(X^0)) + d^\alpha(\tilde{\Phi}_{t,m}(X^0), \tilde{X}_t).
\]
Applying the expectation on both sides of the above inequality, we have
\[
\mathbb{E}d^\alpha(\tilde{\Phi}_{t,m}(X), \tilde{X}_t) \leq \mathbb{E}d^\alpha(\tilde{\Phi}_{t,m}(X), \tilde{\Phi}_{t,m}(X^0)) + \mathbb{E}d^\alpha(\tilde{\Phi}_{t,m}(X^0), \tilde{X}_t)
\]
\[
\lesssim Cr^m d^\alpha(X, X^0) + \mathbb{E}d^\alpha(\tilde{\Phi}_{t,m}(X^0), \tilde{X}_t) \lesssim r^m.
\]
Following the same steps as in the proof of Lemma 7.1, we then have almost surely,
\[
d(\tilde{\Phi}_{t,m}(X), \tilde{X}_t) \xrightarrow{m \to \infty} 0, \quad \forall t \in \mathbb{Z}.
\]

7.3.3 Proof of Lemma 7.3
We would like to show that, for any fixed \( t \in \mathbb{Z} \), \( d(\tilde{X}_t, \Phi_{\epsilon_t}(\tilde{X}_{t-1})) = 0 \), almost surely. Firstly, we have for any \( m \in \mathbb{N} \)
\[
\mathbb{E}d^\alpha(\tilde{X}_t, \Phi_{\epsilon_t}(\tilde{X}_{t-1})) \leq \mathbb{E}d^\alpha(\tilde{X}_t, \tilde{\Phi}_{t,m}(X^0)) + \mathbb{E}d^\alpha(\tilde{\Phi}_{t,m}(X^0), \Phi_{\epsilon_t}(\tilde{X}_{t-1}))
\]
\[
= \mathbb{E}d^\alpha(\tilde{X}_t, \tilde{\Phi}_{t,m}(X^0)) + \mathbb{E}d^\alpha(\Phi_{\epsilon_t} \circ \tilde{\Phi}_{t-1,m-1}(X^0), \Phi_{\epsilon_t}(\tilde{X}_{t-1}))
\]
\[
\leq \mathbb{E}d^\alpha(\tilde{X}_t, \tilde{\Phi}_{t,m}(X^0)) + Cr^m d^\alpha(X^0, \tilde{X}_{t-1}).
\]
Since \( d(\tilde{X}_t, \tilde{\Phi}_{t,m}(X^0)) \xrightarrow{m \to \infty} 0 \) almost surely, for any \( t \in \mathbb{Z} \). Thus, the last bound above tends to 0 as \( m \to \infty \). Thus \( \mathbb{E}d^\alpha(\tilde{X}_t, \Phi_{\epsilon_t}(\tilde{X}_{t-1})) = 0 \), which implies \( d(\tilde{X}_t, \Phi_{\epsilon_t}(\tilde{X}_{t-1})) = 0 \) almost surely.

7.3.4 Proof of Theorem 7.5
We first show that Equation (7.2) holds for any \( \alpha' \in (0, \alpha) \). For any \( X \in \mathcal{X} \), we have
\[
\mathbb{E}d^{\alpha'}(\tilde{\Phi}_{t,m}(X^0), \tilde{\Phi}_{t,m}(X)) \leq \mathbb{E}d^\alpha(\tilde{\Phi}_{t,m}(X^0), \tilde{\Phi}_{t,m}(X)) \leq C \frac{\alpha'}{\alpha} d^\alpha(X^0, X).
\] (7.6)
Let \( \gamma = \min\{\alpha, \beta\} \), then for any \( t, m \in \mathbb{N} \), we have
\[
\mathbb{E}d^\gamma(\tilde{X}_t, S_t) = \mathbb{E}d^\gamma(\tilde{X}_t, \tilde{\Phi}_{t,m}(S_{t-m})) \leq \mathbb{E}d^\gamma(\tilde{X}_t, \tilde{\Phi}_{t,m}(X^0)) + \mathbb{E}d^\gamma(\tilde{\Phi}_{t,m}(X^0), \tilde{\Phi}_{t,m}(S_{t-m}))
\]
\[
\overset{(7.6)}{\leq} \mathbb{E}d^\gamma(\tilde{X}_t, \tilde{\Phi}_{t,m}(X^0)) + C \frac{\alpha'}{\alpha} \mathbb{E}d^\gamma(X^0, S_{t-m}).
\]
Since \( d(\tilde{X}_t, \tilde{\Phi}_{t,m}(X^0)) \xrightarrow{m \to \infty} 0 \) almost surely, for any \( t \in \mathbb{Z} \), we have \( \mathbb{E}d^\gamma(\tilde{X}_t, \tilde{\Phi}_{t,m}(X^0)) \xrightarrow{m \to \infty} 0 \) for any \( t \in \mathbb{Z} \). On the other hand, since \( \gamma < \beta \), for all \( t \in \mathbb{Z} \), we have \( \mathbb{E}d^\gamma(S_t, X^0) \leq \mathbb{E}d^\gamma(S_t, Z^0) + d^\gamma(X^0, Z^0) \leq (\mathbb{E}d^\gamma(S_t, Z^0))^2 + d^\gamma(X^0, Z^0) < M^2 + d^\gamma(X^0, Z^0) \). Therefore, \( \mathbb{E}d^\gamma(\tilde{X}_t, S_t) = 0 \), which implies \( d(\tilde{X}_t, S_t) = 0 \), almost surely, for all \( t \in \mathbb{Z} \).
7.3.5 Proof of Proposition 7.6

We first show that given $X_t \overset{d}{=} X^1 \overset{d}{=} \pi$, we have for any $t \in \mathbb{Z}$, $X_m(X) \overset{d}{=} X_m(X^1) \overset{d}{=} X_t$, almost surely for any $m \in \mathbb{N}$. From Theorem 7.4, we have $X_t \overset{d}{=} X_{t-m} \overset{d}{=} \pi$, almost surely for any $m \in \mathbb{N}, t \in \mathbb{Z}$, and $X_t = \tilde{\Phi}_{t,m}(X_{t-m})$, almost surely for any $m \in \mathbb{N}, t \in \mathbb{Z}$. Thus, we obtain that $\pi = \tilde{\Phi}_{t,m}(\pi)$, almost surely, which implies $X_m(X) \overset{d}{=} X_m(X^1) \overset{d}{=} \pi = X_t$.

Therefore,

$$
\mathbb{E}d^\alpha(X_m(X), X_m(X^1)) = d^\alpha(X_m(X), X_m(X^0)) + d^\alpha(X_m(X^0), X_m(X^1))
$$

$$
= 2\mathbb{E}d^\alpha(X_t, X_m(X^0)) \overset{a.s.}{=} 2\mathbb{E}d^\alpha(X_t, \tilde{\Phi}_{t,m}(X^0))
$$

Inequality (7.5) is checked with $s$ taken as $r$, which completes the proof.

7.3.6 Proof of Theorem 7.7

We first show that the Hilbert mean $\mathbb{E}[X_t]$ for time series $X_t \in (\mathcal{X}, \langle \cdot, \cdot \rangle)$, $t \in \mathbb{Z}$ does not depend on time $t$. We are thus led to show that, for all $t, t' \in \mathbb{Z}$, $\mathbb{E}[X_t] = \mathbb{E}[X_{t'}]$. By the definition of Hilbert mean, this is equivalent to show that

$$
\mathbb{E}(X_t, X) = \mathbb{E}(X_{t'}, X), \quad \forall X \in \mathcal{X}.
$$

Firstly, we show that $\forall t \in \mathbb{Z}, X \in \mathcal{X}$, $\mathbb{E}(X_t, X) = \lim_{m \to \infty} \mathbb{E}(\tilde{\Phi}_{t,m}(X^0), X)$. We have

$$
\mathbb{E}(\tilde{\Phi}_{t,m}(X^0) - X_t, X) \leq \mathbb{E}\|\tilde{\Phi}_{t,m}(X^0) - X_t\| X
$$

$$
\overset{(a)}{\leq} \|X\| (\mathbb{E}\|\tilde{\Phi}_{t,m}(X^0) - X_t\|) \overset{\alpha}{=} \|X\| \left(\mathbb{E}d^\alpha(\tilde{\Phi}_{t,m}(X^0), X_t)\right)^\frac{1}{\alpha} m \to \infty 0.
$$

Inequality (a) comes from the condition $\alpha \geq 2$ and Jensen inequality. Thus, for any $t \in \mathbb{Z},$

$$
\lim_{m \to \infty} \mathbb{E}(\tilde{\Phi}_{t,m}(X^0) - X_t, X) = 0.
$$

On the other hand, since $\epsilon_t$ are i.i.d., there is $\tilde{\Phi}_{t,m}(X^0) \overset{d}{=} \tilde{\Phi}_{t',m}(X^0)$, $\forall t, t' \in \mathbb{Z}, m \in \mathbb{N}$. Thus, we have for any $t, t' \in \mathbb{Z}, X \in \mathcal{X}^\prime$

$$
\mathbb{E}(\tilde{\Phi}_{t,m}(X^0), X) = \mathbb{E}(\tilde{\Phi}_{t',m}(X^0), X), \quad \forall m \in \mathbb{N}.
$$

Then $\forall t \in \mathbb{Z}, X \in \mathcal{X}^\prime$

$$
\lim_{m \to \infty} \mathbb{E}(\tilde{\Phi}_{t,m}(X^0), X) = \lim_{m \to \infty} \mathbb{E}(\tilde{\Phi}_{t',m}(X^0), X), \quad \forall m \in \mathbb{N},
$$

which implies $\forall t \in \mathbb{Z}, X \in \mathcal{X}$

$$
\mathbb{E}(X_t, X) = \mathbb{E}(X_{t'}, X).
$$

We denote $\mathbb{E}[X_t]$ by $U$. Next, since $\mathbb{E}[X_t]^2 < \infty$ for all $t \in \mathbb{Z}$, the auto-covariance operator $\mathcal{G}_{t,t+h}$ is well-defined. We are now to show $\mathcal{G}_{t,t+h}$ is time-invariant, which is equivalent to show

$$
\langle \mathcal{G}_{t,t+h}(X), Y \rangle = \langle \mathcal{G}_{t',t'+h}(X), Y \rangle, \quad \forall X, Y \in \mathcal{X}^\prime,
$$

30
by Definition 3.1, that is
\[ \mathbb{E}(X_t - U, X \mid X_{t+h} - U, Y) = \mathbb{E}(X_{t+m} - U, X \mid X_{t+h+m} - U, Y), \quad \forall X, Y \in \mathcal{X}. \tag{7.7} \]

Analogously, we show firstly that \( \forall t, h \in \mathbb{Z}, X, Y \in \mathcal{X} \)
\[ \mathbb{E}(X_t - U, X \mid X_{t+h} - U, Y) = \lim_{m \to \infty} \mathbb{E}(\tilde{\Phi}_{t+m}(X^0) - U, X \mid \tilde{\Phi}_{t+h+m}(X^0) - U, Y). \tag{7.8} \]

We have
\[ \mathbb{E}(X_t - U, X \mid X_{t+h} - U, Y) - \mathbb{E}(\tilde{\Phi}_{t+m}(X^0) - U, X \mid \tilde{\Phi}_{t+h+m}(X^0) - U, Y) \]
\[ \leq \mathbb{E}(X_t - U, X \mid X_{t+h} - U, Y) - \mathbb{E}(\tilde{\Phi}_{t+m}(X^0) - U, X \mid X_{t+h} - \tilde{\Phi}_{t+h+m}(X^0), Y) \]
\[ + \mathbb{E}(\tilde{\Phi}_{t+m}(X^0) - U, X \mid X_{t+h} - U, Y) - \mathbb{E}(\tilde{\Phi}_{t+m}(X^0) - U, X \mid X_{t+h} - \tilde{\Phi}_{t+h+m}(X^0), Y) \]
\[ \leq \mathbb{E}(X_t - \tilde{\Phi}_{t+m}(X^0) - U, X \mid X_{t+h} - U, Y) + \mathbb{E}(\tilde{\Phi}_{t+m}(X^0) - U, X \mid X_{t+h} - \tilde{\Phi}_{t+h+m}(X^0), Y) \]
\[ \leq ||X||Y||E(X_t - X_{t+h} - U) + ||X||Y||E(\tilde{\Phi}_{t+m}(X^0) - U) ||X_{t+h} - \tilde{\Phi}_{t+h+m}(X^0)|| \]
\[ \leq ||X||Y||\text{Ed}(X_t, \tilde{\Phi}_{t,m}(X^0))d(X_{t+h}, U) + ||X||Y||\text{Ed}^{\tilde{\Phi}_{t,m}(X^0)}, U)d(X_{t+h}, \tilde{\Phi}_{t+h,m}(X^0)). \]

Since \( d(X_t, \tilde{\Phi}_{t,m}(X^0)) \rightarrow X \) almost surely, for any \( t \in \mathbb{Z} \), then \( d(\tilde{\Phi}_{t,m}(X^0)), U \) \( m \rightarrow X \) \( d(X_t, U) \), for any \( t \in \mathbb{Z} \). Thus, the last bound tends to 0, as \( m \to \infty \), which implies Equation (7.8).

On the other hand, since \( \epsilon_i \) are i.i.d., thus \( \forall t, h \in \mathbb{Z}, X, Y \in \mathcal{X} \) and \( \forall m \in \mathbb{N} \)
\[ \langle \tilde{\Phi}_{t,m}(X^0) - U, X \mid \tilde{\Phi}_{t+h,m}(X^0) - U, Y \rangle = d(\tilde{\Phi}_{t,m}(X^0) - U, X \mid \tilde{\Phi}_{t+h,m}(X^0) - U, Y), \]
which follows
\[ \mathbb{E}(\tilde{\Phi}_{t,m}(X^0) - U, X \mid \tilde{\Phi}_{t+h,m}(X^0) - U, Y) = \mathbb{E}(\tilde{\Phi}_{t,m}(X^0) - U, X \mid \tilde{\Phi}_{t+h,m}(X^0) - U, Y). \tag{7.9} \]

Take limit on \( m \) on both sides of Equation (7.9), we obtain Equation (7.7). Thus the auto-covariance is time-invariant.

### 7.4 Proof of Theorem 3.1 and Theorem 3.2

With \( (\mathcal{X}, d) \) given in Definition (3.7), \( \Phi_{\epsilon_i} \) defined in Equation (3.8), and \( \Theta \) defined as the product space of \( \Pi_{\mathcal{X}} ||_{\text{L}} \) with \( \Pi_{\mathcal{X}} ||_{\text{L}} \) defined in Equation (3.4), we show the resulting system (3.8) satisfies Conditions 1, 2 in Section 7.1.

The verification of Condition 1 does not require the additional assumptions. We consider any \( X \in (\mathcal{X}, d) \), \( \alpha = 2 \), then
\[ \mathbb{E}(d^2(X, \Phi_{\epsilon_i}(X))) \leq \mathbb{E} \left( \sum_{i=1}^{N} ||X_i - \Phi_{\epsilon_i}(X)||^2_{\text{Leb}} \right). \]

Since \( (\mathcal{X}, d) \) is a space of bounded functions from \( (0, 1) \) to \([0, 1] \). Thus \( \mathbb{E}(d^2(X, \Phi_{\epsilon_i}(X))) \) is bounded.
We are now to show that under Assumptions $A_4$ and $A_5$, IRF System (3.8) satisfies Condition 2. We first examine the case of $m = 1$, we consider any $X, X^1 \in \mathcal{X}$, then

$$
\mathbb{E}(d^\alpha(\tilde{\Phi}_{t,1}(X), \tilde{\Phi}_{t,1}(X^1))) \\
= \mathbb{E} \left[ \sum_{i=1}^N \left| \epsilon_{i,t} \circ \left[ \sum_{j=1}^N A_{ij}(X_j - \mu) + \mu \right] - \epsilon_{i,t} \circ \left[ \sum_{j=1}^N A_{ij}(X_j^1 - \mu) + \mu \right] \right|^2 \right] \\
= \sum_{i=1}^N \mathbb{E} \left( \epsilon_{i,t} \circ \left[ \sum_{j=1}^N A_{ij}(X_j - \mu) + \mu \right] (p) - \epsilon_{i,t} \circ \left[ \sum_{j=1}^N A_{ij}(X_j^1 - \mu) + \mu \right] (p) \right)^2 dp
$$

Assumption $A_4$

$$
\leq \sum_{i=1}^N L^2 \int_0^1 \left( \sum_{j=1}^N A_{ij}(X_j(p) - X_j^1(p)) \right)^2 dp = L^2 \int_0^1 \sum_{i=1}^N \left( \sum_{j=1}^N A_{ij}(X_j(p) - X_j^1(p)) \right)^2 dp
$$

Exchange (b) of the integral and the expectation comes from that the integrand is bounded. Then for any fixed $p \in (0, 1)$, $\int \sum_{j=1}^N A_{ij}(X_j(p) - X_j^1(p)) \leq \|A\|_{l_2} \|x_{p} - x_{p}^1\|_{l_2}$, where $\bar{x}_p = (X_1(p) - X_1^1(p), \ldots, X_N(p) - X_N^1(p))^T$. Since $\|A\|_{l_2} \leq \|A\|_{l_2} \|x_{p} - x_{p}^1\|_{l_2}$, we have

$$
L^2 \int_0^1 \sum_{i=1}^N \left( \sum_{j=1}^N A_{ij}(X_j(p) - X_j^1(p)) \right)^2 dp \\
\leq L^2 \|A\|_{l_2}^2 \int_0^1 \sum_{j=1}^N ((X_j(p) - X_j^1(p))^2) dp = L^2 \|A\|_{l_2}^2 d^2(\mathcal{X}, X^1).
$$

By Assumption $A_5$, $L^2 \|A\|_{l_2}^2 < 1$, thus by taking $r = L^2 \|A\|_{l_2}, C = 1$, Equation (7.2) is checked for $m = 1$. Now, suppose for any fixed $m$, we have

$$
\mathbb{E}(d^\alpha(\tilde{\Phi}_{t,m}(X), \tilde{\Phi}_{t,m}(X^1))) \leq r^m d^2(\mathcal{X}, X^1). \tag{7.10}
$$

32
Then for $m + 1$, we have
\[
\begin{align*}
\mathbb{E}(d^m(\tilde{\Phi}_{t,m+1}(X), \tilde{\Phi}_{t,m+1}(X^1))) &= \mathbb{E}\left[\sum_{i=1}^{N} \epsilon_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\tilde{\Phi}_{t-1,m}(X) - id\right) + id\right] - \epsilon_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\tilde{\Phi}_{t-1,m}(X^1) - id\right) + id\right]\right] \\
&= \sum_{i=1}^{N} \int_{0}^{1} \mathbb{E}(\epsilon_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\tilde{\Phi}_{t-1,m}(X) - id\right) + id\right](p) \\
&\quad - \epsilon_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\tilde{\Phi}_{t-1,m}(X^1) - id\right) + id\right](p)) dp \\
&= \sum_{i=1}^{N} \int_{0}^{1} \mathbb{E}(\epsilon_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\tilde{\Phi}_{t-1,m}(X) - id\right) + id\right](p) \\
&\quad - \epsilon_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\tilde{\Phi}_{t-1,m}(X^1) - id\right) + id\right](p)) dp \\
&\leq \sum_{i=1}^{N} L^2 \int_{0}^{1} \mathbb{E}\left[\sum_{j=1}^{N} A_{ij} \left(\tilde{\Phi}_{t-1,m}(X) - id\right) + id\right](p) \\
&\quad - \left[\sum_{j=1}^{N} A_{ij} \left(\tilde{\Phi}_{t-1,m}(X^1) - id\right) + id\right](p) dp \\
&= L^2 \int_{0}^{1} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} A_{ij} (\tilde{\Phi}_{t-1,m}(X)(p) - \tilde{\Phi}_{t-1,m}(X^1)(p))\right) dp \\
&\leq L^2 |A|^2 \mathbb{E}(d^2(\tilde{\Phi}_{t-1,m}(X), \tilde{\Phi}_{t-1,m}(X^1))) \overset{(c)}{=} r^{m+1}d^2(X, X^1).
\end{align*}
\]

Inequality (c) comes from that $\tilde{\Phi}_{t,m}(X) \overset{d}{=} \tilde{\Phi}_{t-1,m}(X), \forall t \in \mathbb{Z}$ and the hypothesis (7.10). Thus, by induction, Equation (7.2) is verified for all $m \geq 1$, thus Condition 2 is checked by IRF System (3.8).

By Theorems 7.4, 7.5, we have proved Theorem 3.1. Note that since $X$ is the set of quantile functions from $0, 1$ to $0, 1$, all the elements in $X$ are bounded uniformly in the sense of Inequality (7.3) with $d$ defined as (3.7). Thus Theorem 7.5 applies to all solution $S_i$ in $(X, d)$. Moreover, since $\alpha$ is taken as 2, by Theorem 7.7, we obtain furthermore the stationarity of the unique solution $X_i$ in the underlying Hilbert space $(X, \langle \cdot, \cdot \rangle)$ with $\langle X, Y \rangle = \sum_{i} \langle X_i, Y_i \rangle_{\text{Leb}}$.

Lastly, applying Proposition 7.6, we have additionally that IRF system (3.8) is geometric moment contracting.

\section*{7.5 Proof of Proposition 3.3}

Since $X_i$ is a solution of IRF system (3.8) defined in $T^N, d$, where $T$ is a set of quantile functions over $[0, 1]$. Thus point 1 is checked.

For any $p \in (0, 1)$, from the definition of system (3.8), we have
\[
X_{i,t}(p) = \epsilon_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} (X_{j,t-1} - id) + id\right](p), \quad t \in \mathbb{Z}, i = 1, \ldots, N.
\]
Take expectation on both sides, gives
\[
\mathbb{E} \mathbf{X}_{i,t}(p) = \mathbb{E} \left[ \epsilon_{i,t} \circ \left[ \sum_{j=1}^{N} A_{ij} (\mathbf{X}_{j,t-1} - id) + id \right] (p) \right]
\]
\[
\overset{(a)}{=} \mathbb{E} \left[ \sum_{j=1}^{N} A_{ij} (\mathbf{X}_{j,t-1} - id) + id \right] (p)
\]
\[
= \sum_{j=1}^{N} A_{ij} (\mathbb{E} \mathbf{X}_{j,t-1}(p) - p) + p.
\]

Equation (a) comes from that \( \epsilon_{i,t} \) is independent of \( \mathbf{X}_{t-1} \). Thus we have
\[
\mathbb{E} \mathbf{X}_{i,t}(p) - p = \sum_{j=1}^{N} A_{ij} (\mathbb{E} \mathbf{X}_{j,t-1}(p) - p), i = 1, \ldots, N, p \in (0,1).
\]

When \( \mathbf{X}_{t}, t \in \mathbb{Z} \) is a stationary solution, the equation becomes
\[
\mathbb{E} \mathbf{X}_{i,t}(p) - p = \sum_{j=1}^{N} A_{ij} (\mathbb{E} \mathbf{X}_{j,t}(p) - p), i = 1, \ldots, N, p \in (0,1).
\]

Let \( \bar{x}_p = (\mathbb{E} \mathbf{X}_{1,t}(p) - p, \ldots, \mathbb{E} \mathbf{X}_{N,t}(p) - p)^\top \), we then have
\[
\bar{x}_p = A \bar{x}_p \iff (I - A) \bar{x}_p = 0.
\]

Since \( \|A\| < 1 \), \( I - A \) is invertible, which implies \( \bar{x}_p = 0 \) for any \( p \in (0,1) \). Thus \( \mathbb{E} \mathbf{X}_{i,t}(p) = p, i = 1, \ldots, N, p \in (0,1) \).

\[ \blacksquare \]

7.6 Proof of Proposition 3.4

\textbf{Point 1:} Firstly, we show that \( \Gamma(0) \) is positive semi-definite. Let \( w = (w_1, \ldots, w_N)^\top \in \mathbb{R}^N \), we have
\[
w^\top \Gamma(0)w = \sum_{i,j=1}^{N} w_i w_j \mathbb{E} \langle \mathbf{X}_{i,t-1} - id, \mathbf{X}_{j,t-1} - id \rangle_{\text{Leb}}
\]
\[
= \mathbb{E} \left\| \sum_{i=1}^{N} w_i (\mathbf{X}_{i,t-1} - id) \right\|^2_{\text{Leb}} \geq 0.
\]

We now show for any \( w, w^\top \Gamma(0)w \neq 0 \). Given
\[
\mathbf{X}_{i,t-1} = \epsilon_{i,t-1} \circ \left[ \sum_{j=1}^{N} A_{ij} (\mathbf{X}_{j,t-2} - id) + id \right],
\]

\[
\mathbb{E} \mathbf{X}_{i,t}(p) - p = \sum_{j=1}^{N} A_{ij} (\mathbb{E} \mathbf{X}_{j,t}(p) - p), i = 1, \ldots, N, p \in (0,1).
\]

Let \( \bar{x}_p = (\mathbb{E} \mathbf{X}_{1,t}(p) - p, \ldots, \mathbb{E} \mathbf{X}_{N,t}(p) - p)^\top \), we then have
\[
\bar{x}_p = A \bar{x}_p \iff (I - A) \bar{x}_p = 0.
\]

Since \( \|A\| < 1 \), \( I - A \) is invertible, which implies \( \bar{x}_p = 0 \) for any \( p \in (0,1) \). Thus \( \mathbb{E} \mathbf{X}_{i,t}(p) = p, i = 1, \ldots, N, p \in (0,1) \).

\[ \blacksquare \]
we denote $\sum_{j=1}^{N} A_{ij} (X_{j,t-2} - id) + id$ by $H_{t-2}$, we have

$$
\mathbb{E} \left[ \sum_{i=1}^{N} w_i (X_{i,t-1} - id) \right]_{Leb}^{2} = \mathbb{E} \left[ \sum_{i=1}^{N} w_i (\epsilon_{i,t-1} \circ H_{t-2} - id) \right]_{Leb}^{2} = \mathbb{E} \left( \mathbb{E} \left[ \left\| \sum_{i=1}^{N} w_i (\epsilon_{i,t-1} \circ H_{t-2} - id) \right\|_{Leb}^{2} \mid H_{t-2} \right] \right) = (a) \sum_{i=1}^{N} w_i^2 \mathbb{E} \left( \mathbb{E} \left[ \| \epsilon_{i,t-1} \circ H_{t-2} - id \|_{Leb}^{2} \mid H_{t-2} \right] \right).
$$

Equation (a) is because that $\epsilon_{i,t-1}, i = 1, \ldots, N$ are independent of $H_{t-2}$, and $\epsilon_{i,t-1}, i = 1, \ldots, N$ are mutually independent. Since $\epsilon_{i,t-1}, i = 1, \ldots, N$ are independent of $H_{t-2}$, we have furthermore

$$
\mathbb{E} \left[ \| \epsilon_{i,t-1} \circ H_{t-2} - id \|_{Leb}^{2} \mid H_{t-2} \right] \neq 0.
$$

Otherwise, we would have

$$
P \left[ \| \epsilon_{i,t-1} \circ H_{t-2} - id \|_{Leb}^{2} \mid H_{t-2} \right] = 1 \iff P \left\{ \epsilon_{i,t-1} \circ H_{t-2} = id \text{ almost everywhere} \mid H_{t-2} \right\} = 1
$$

$$
\iff P \left\{ \epsilon_{i,t-1} = H_{t-2}^{-1} \text{ a.e.} \mid H_{t-2} \right\} = 1 \iff P \left\{ \epsilon_{i,t-1} = H_{t-2}^{-1} \text{ a.e.} \right\} = 1,
$$

which is a contradiction to the independence between $\epsilon_{i,t-1}$ and $H_{t-2}^{-1}$. Thus, we can conclude that $\Gamma(0)$ is nonsingular.

**Point 2:** Process (3.8) writes in terms of components $X_{i,t}$ as

$$
X_{i,t} = \epsilon_{i,t} \circ \left[ \sum_{j=1}^{N} A_{ij} (X_{j,t-1} - id) + id \right].
$$

Subtract id from the both sides, we obtain

$$
X_{i,t} - id = \epsilon_{i,t} \circ \left[ \sum_{j=1}^{N} A_{ij} (X_{j,t-1} - id) + id \right] - id.
$$

Pass the both sides to the inner product with $X_{i,t-1} - id$, then take the expectation, we get

$$
\mathbb{E} \langle X_{i,t} - id, X_{i,t-1} - id \rangle = \mathbb{E} \left( \epsilon_{i,t} \circ \left[ \sum_{j=1}^{N} A_{ij} (X_{j,t-1} - id) + id \right] - id, X_{i,t-1} - id \right) = \mathbb{E} \left( \mathbb{E} \left( \epsilon_{i,t} \circ \left[ \sum_{j=1}^{N} A_{ij} (X_{j,t-1} - id) + id \right] - id, X_{i,t-1} - id \mid X_{\tau}, \tau \leq t - 1 \right) \right) = \mathbb{E} \left( \sum_{j=1}^{N} A_{ij} (X_{j,t-1} - id), X_{i,t-1} - id \right) = \sum_{j=1}^{N} A_{ij} \mathbb{E} \langle X_{j,t-1} - id, X_{i,t-1} - id \rangle.
$$

Compare the definitions of $\Gamma(0), \Gamma(1)$, we can then retrieve Representation (3.9).
7.7 Proof of Lemma 4.1

Since $\tilde{F}_t^{-1}$ follows the stationary distribution $\pi$ defined in Theorem 7.4, the sequence $\tilde{F}_t^{-1}$, $t = 1, \ldots, T$ generated successively by the stationary model (3.8) follow $\pi$, furthermore, their auto-covariance is time-invariant and equal to the stationary solution $X_t$, $t \in \mathbb{Z}$. Thus we can use the data $\tilde{F}_t^{-1}$ in Representation 3.4, which writes as

$$A = \Gamma(1) [\Gamma(0)]^{-1},$$

where $[\Gamma(0)]_{j,l} = E \langle \tilde{F}_{t-1} - id, \tilde{F}_{t-1} - id \rangle_{Leb}$ and $[\Gamma(1)]_{j,l} = E \langle \tilde{F}_{t-1} - id, \tilde{F}_{t-1} - id \rangle_{Leb}$. Since $\tilde{A} = \tilde{\Gamma}(1) [\tilde{\Gamma}(0)]^{-1}$. We first show $[\tilde{\Gamma}(0)]_{j,l} - [\Gamma(0)]_{j,l} = \mathcal{O}(\frac{1}{\sqrt{T}})$ by applying Theorem 3 in Wu and Shao (2004).

To this end, we define $g_{jl} : (\mathcal{X}, \langle \cdot, \cdot \rangle) \rightarrow \mathbb{R}$ as

$$g_{jl}(X) = \langle X_j - id, X_l - id \rangle_{Leb} - E \langle X_j - id, X_l - id \rangle_{Leb}.$$

The construction implies that for any random object $Y = (Y_i)_{i=1}^N$ in $(\mathcal{X}, \langle \cdot, \cdot \rangle)$, $Eg_{jl}(Y) = 0$. Moreover, since all $Y_i$ are bounded function from $(0, 1)$ to $[0, 1]$, $|g_{jl}(Y)|^p < \infty$ for all $p > 2$, which leads to $E|g_{jl}(Y)|^p < \infty$ for all $p > 2$. It is then left to show that $g_{jl}$ is stochastic dini-continuous (Wu and Shao, 2004, Equation (9)).

For any $Y$ and $Y^1$ identically distributed in $(\mathcal{X}, \langle \cdot, \cdot \rangle)$, we have

$$|g_{jl}(Y) - g_{jl}(Y^1)| = |\langle Y_j - id, Y_l - id \rangle_{Leb} - \langle Y_j^1 - id, Y_l^1 - id \rangle_{Leb}|$$

$$\leq |\langle Y_j - id, Y_l - id \rangle_{Leb} - \langle Y_j - id, Y_l^1 - id \rangle_{Leb}| + |\langle Y_j - id, Y_l^1 - id \rangle_{Leb} - \langle Y_j - id, Y_l^1 - id \rangle_{Leb}|$$

$$\leq |\langle Y_j - id, Y_l^1 \rangle_{Leb}| + |\langle Y_j - id, Y_l^1 - Y_j^1 \rangle_{Leb}|$$

$$\leq \|Y_j - id\| \|Y_l - Y_l^1\| + \|Y_j - Y_j^1\| \|Y_l - id\|.$$

Since $Y_j$ and $Y_j^1$ are increasing function from $(0, 1)$ to $[0, 1]$, $\|Y_j - id\| \leq \frac{1}{2}$ and $\|Y_j^1 - id\| \leq \frac{1}{2}$. Thus, we have furthermore

$$|g_{jl}(Y) - g_{jl}(Y^1)| \leq \frac{1}{2} (\|Y_l - Y_l^1\| + \|Y_j - Y_j^1\|)$$

$$\leq \frac{\sqrt{2}}{2} (\|Y_l - Y_l^1\| + \|Y_j - Y_j^1\|)^{\frac{1}{2}}.$$

Then

$$\sup_{Y, Y^1} \left\{ |g_{jl}(Y) - g_{jl}(Y^1)| 1_{\left(\sum_{i=1}^T |Y_i - Y_i^1|^2 < \delta\right)} \right\} \leq \frac{\sqrt{2}}{2} \delta.$$

Thus $g_{jl}$ is stochastic dini-continuous. Since IRF system (3.8) is geometric moment contracting indicated in the proof of Theorem (3.1), then by Theorem 3 in Wu and Shao (2004),

$$\frac{S_T(g_{jl})}{\sqrt{T}} := \sqrt{T} \left( [\tilde{\Gamma}(0)]_{j,l} - [\Gamma(0)]_{j,l} \right) \overset{d}{\rightarrow} \sigma_{g_{jl}} \mathcal{N}(0, 1),$$

which is followed

$$[\tilde{\Gamma}(0)]_{j,l} - [\Gamma(0)]_{j,l} = \mathcal{O}_p\left(\frac{1}{\sqrt{T}}\right).$$
We first show a convergence result in Lemma 7.8 which will be required by the demonstration of the Assumption A1, where
\[ g_{jl}(X, X') = \langle X_j - id, X'_l - id \rangle_{Leb} - \mathbb{E} \langle X_j - id, X'_j - id \rangle_{Leb}. \]
It can be shown by the similar proof as before that \( g_{jl}(X, X') \) is stochastic dih-continuous with respect to the product metric \( \rho((X, X'), (Z, Z')) = \sqrt{\sum_{i=1}^{N} \|X_i - Z_i\|^2 + \sum_{i=1}^{N} |X'_i - Z'_i|^2} \). Therefore by Theorem 3 in Wu and Shao (2004), we have
\[
\left[ \tilde{F}(1) \right]_{j,l} - [\Gamma(1)]_{j,l} = O_p\left( \frac{1}{\sqrt{T}} \right), \quad j, l = 1, \ldots, N.
\]
Since matrix inversion is a continuous mapping, thus by continuous mapping theorem
\[
\left[ \tilde{F}(0)^{-1} \right]_{j,l} - [\Gamma(0)^{-1}]_{j,l} = O_p\left( \frac{1}{\sqrt{T}} \right), \quad j, l = 1, \ldots, N.
\]
Representations of \( \tilde{A} \) and \( A \) then bring to \( \tilde{A}_{j,l} = A_{j,l} = O_p\left( \frac{1}{\sqrt{T}} \right), \quad j, l = 1, \ldots, N. \)

7.8 Proof of Lemma 4.2

To prove Equation (4.6), we define, for any fixed \( p \in (0, 1), \ i = 1, \ldots, N, \ g_i^p : (\mathcal{X}, \langle \cdot, \cdot \rangle) \to \mathbb{R} \) defined as
\[ g_i^p(X) = X_i(p) - \mathbb{E}X_i(p). \]
Follow the similar steps in the proof of Lemma 4.1, we have
\[ \frac{ST(g_i^p)}{T} := \frac{1}{T} \sum_{t=1}^{T} \tilde{F}_{i,t-1}(p) - \mathbb{E} \tilde{F}_{i,t-1}(p) \overset{Equation 3.3}{=} \frac{1}{T} \sum_{t=1}^{T} \tilde{F}_{i,t-1}(p) - p = O\left( \frac{1}{\sqrt{T}} \right). \]
Then plug the transformation relation in \( \tilde{F}_{i,t-1} \), we have
\[ \frac{1}{T} \sum_{t=1}^{T} \tilde{F}_{i,t-1}^{-1}\left[\tilde{F}_{i,t-1}^{-1}(p)\right] - p = O\left( \frac{1}{\sqrt{T}} \right), \]
where \( (F_{i,\odot}^{-1})^{-1} \) is the left continuous inverse of the quantile function \( F_{i,\odot}^{-1} \). By the statement under Assumption A1, \( F_{i,\odot}^{-1} \) is continuous, thus \( (F_{i,\odot}^{-1})^{-1} \) is strictly increasing. We therefore substitute \( q \) for \( (F_{i,\odot}^{-1})^{-1}(p) \), it follows \( p = F_{i,\odot}(q) \), which brings to Equation (4.7).

7.9 Proof of Theorem 4.3

We first show a convergence result in Lemma 7.8 which will be required by the demonstration of the theorem.

Lemma 7.8. Under the conditions of Lemma 4.1, we have
\[ \frac{1}{T} \sum_{t=1}^{T} \langle id, \tilde{F}_{j,t-1}^{-1} - id \rangle = O\left( \frac{1}{\sqrt{T}} \right), \quad j = 1, \ldots, N. \]
The proof of Lemma 4.2 follows the same steps in the one of Lemma 4.1, by considering
\( g_j : (X, \langle \cdot, \cdot \rangle) \to \mathbb{R} \) defined as
\[
g_j(X) = \langle id, X_j - id \rangle_{Leb} - \mathbb{E} \langle id, X_j - id \rangle_{Leb}.
\]
Then use \( \mathbb{E} \hat{F}_{t,i}(p) = p, \forall p \in (0, 1) \), we obtain the result.

We are now to prove \( \hat{A}_o - \hat{A} \overset{p}{\rightarrow} 0 \). Note that analogous to \( A \) and \( \hat{A} \),
\[
\hat{A}_o = \hat{\Gamma}(0) \left[ \hat{\Gamma}(1) \right]^{-1},
\]
\( \hat{\Gamma}(0), \hat{\Gamma}(1) \) are given in Equation (4.4). Thus, we first show \( \hat{\Gamma}(0)_{j,l} - \hat{\Gamma}(0)_{j,l} \overset{p}{\rightarrow} 0 \). By calculation, we have

\[
[\hat{\Gamma}(0)]_{j,l} = \frac{1}{T} \sum_{t=1}^{T} \langle \hat{F}_{j,t-1}^{-1} - id, \hat{F}_{l,t-1}^{-1} - id \rangle_{Leb} = \frac{1}{T} \sum_{t=1}^{T} \langle \hat{F}_{j,t}^{-1} - id, \hat{F}_{l,t}^{-1} - id \rangle_{Leb} + O(\frac{1}{T})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{l,t}^{-1} - id, F_{l,t}^{-1} \otimes F_{l,t}^{-1} - id \rangle_{Leb} + O(\frac{1}{T})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{l,t}^{-1} - id \rangle_{Leb} - \frac{1}{T} \sum_{t=1}^{T} \langle id, F_{l,t}^{-1} \otimes F_{l,t}^{-1} - id \rangle_{Leb} + O(\frac{1}{T})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{l,t}^{-1} - id \rangle_{Leb} + O(\frac{1}{T})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{l,t}^{-1} - id \rangle_{Leb} - \frac{1}{T} \sum_{t=1}^{T} \langle id, F_{l,t}^{-1} \otimes F_{l,t}^{-1} - id \rangle_{Leb} + O(\frac{1}{T})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{l,t}^{-1} - id \rangle_{Leb} - \langle id, id \rangle_{Leb} + O(\frac{1}{T}),
\]

and

\[
[\hat{\Gamma}(0)]_{j,l} = \frac{1}{T} \sum_{t=1}^{T} \langle \hat{F}_{j,t-1}^{-1} - id, \hat{F}_{l,t-1}^{-1} - id \rangle_{Leb} = \frac{1}{T} \sum_{t=1}^{T} \langle \hat{F}_{j,t}^{-1} - id, \hat{F}_{l,t}^{-1} - id \rangle_{Leb} + O(\frac{1}{T})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{j,t}^{-1} - id, F_{l,t}^{-1} \otimes F_{j,t}^{-1} - id \rangle_{Leb} + O(\frac{1}{T})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{j,t}^{-1} - id \rangle_{Leb} - \frac{1}{T} \sum_{t=1}^{T} \langle id, F_{l,t}^{-1} \otimes F_{j,t}^{-1} - id \rangle_{Leb} + O(\frac{1}{T})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{j,t}^{-1} - id \rangle_{Leb} + O(\frac{1}{T})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{j,t}^{-1} - id \rangle_{Leb} - \frac{1}{T} \sum_{t=1}^{T} \langle id, F_{l,t}^{-1} \otimes F_{j,t}^{-1} - id \rangle_{Leb} + O(\frac{1}{T})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{j,t}^{-1} - id \rangle_{Leb} + O(\frac{1}{T})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{j,t}^{-1} - id \rangle_{Leb} + O_p(\frac{1}{\sqrt{T}})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{j,t}^{-1} - id \rangle_{Leb} + O_p(\frac{1}{\sqrt{T}})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{j,t}^{-1} - id \rangle_{Leb} + O_p(\frac{1}{\sqrt{T}})
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{j,t}^{-1} - id \rangle_{Leb} + O_p(\frac{1}{\sqrt{T}}).
\]
Equations (a) and (b) come from $\widehat{F}_{i,t}^{-1}$ and $\widehat{F}_{i,t-1}^{-1}$ are bounded between 0 and 1 for all $i = 1, ..., N$ and $t \in \mathbb{N}$. Thus,

$$[\widehat{F}(0)]_{j,l} - [\widehat{F}(0)]_{j,l} = \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{p_j}^{-1}, F_{i,t}^{-1} \otimes F_{p_i}^{-1} \rangle_{\text{Leb}}$$

$$- \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{p_j}^{-1}, F_{i,t}^{-1} \otimes F_{p_i}^{-1} \rangle_{\text{Leb}} + O_p\left(\frac{1}{\sqrt{T}}\right)$$

$$= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{p_j}^{-1}, F_{i,t}^{-1} \otimes F_{p_i}^{-1} \rangle_{\text{Leb}} - \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{p_j}^{-1}, F_{i,t}^{-1} \otimes F_{p_i}^{-1} \rangle_{\text{Leb}}$$

$$+ \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{p_j}^{-1}, F_{i,t}^{-1} \otimes F_{p_i}^{-1} \rangle_{\text{Leb}} - \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{p_j}^{-1}, F_{i,t}^{-1} \otimes F_{p_i}^{-1} \rangle_{\text{Leb}} + O_p\left(\frac{1}{\sqrt{T}}\right)$$

$$= \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{p_j}^{-1}, F_{i,t}^{-1} \otimes F_{p_i}^{-1} \rangle_{\text{Leb}}$$

$$+ \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{p_j}^{-1}, F_{i,t}^{-1} \otimes F_{p_i}^{-1} \rangle_{\text{Leb}} + O_p\left(\frac{1}{\sqrt{T}}\right).$$

(7.12)

We are now to show $\frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{p_j}^{-1}, F_{i,t}^{-1} \otimes F_{p_i}^{-1} \rangle_{\text{Leb}} \rightarrow 0$. We bound the absolute value

$$\left| \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \otimes F_{p_j}^{-1}, F_{i,t}^{-1} \otimes F_{p_i}^{-1} \rangle_{\text{Leb}} \right|$$

$$\leq \frac{1}{T} \sum_{t=1}^{T} \int_{0}^{1} \left| (F_{j,t}^{-1} \otimes F_{p_j}^{-1})(p) - (F_{i,t}^{-1} \otimes F_{p_i}^{-1})(p) \right| dp$$

$$\leq \frac{a}{T} \sum_{t=1}^{T} \int_{0}^{1} \left| (F_{j,t}^{-1} \otimes F_{p_j}^{-1})(p) - (F_{i,t}^{-1} \otimes F_{p_i}^{-1})(p) \right| dp.$$ 

Inequality (a) is because $F_{i,t}^{-1} \otimes F_{p_i}^{-1}$ is bounded between 0 and 1. Note that for any fixed $p \in (0, 1)$, we have

$$\frac{1}{T} \sum_{t=1}^{T} \left| (F_{j,t}^{-1} \otimes F_{p_j}^{-1})(p) - (F_{i,t}^{-1} \otimes F_{p_i}^{-1})(p) \right| = \frac{1}{T} \sum_{t=1}^{T} \left| F_{j,t}^{-1} \circ (F_{p_j}^{-1})^{-1}(p) - F_{i,t}^{-1} \circ (F_{p_i}^{-1})^{-1}(p) \right|$$

$$= \left\{ \begin{array}{ll}
\frac{1}{T} \sum_{t=1}^{T} F_{j,t}^{-1} \circ (F_{p_j}^{-1})^{-1}(p) - F_{i,t}^{-1} \circ (F_{p_i}^{-1})^{-1}(p), & \text{if } (F_{p_j}^{-1})^{-1}(p) > (F_{p_i}^{-1})^{-1}(p), \\
\frac{1}{T} \sum_{t=1}^{T} F_{j,t}^{-1} \circ (F_{p_j}^{-1})^{-1}(p) - F_{i,t}^{-1} \circ (F_{p_i}^{-1})^{-1}(p), & \text{otherwise},
\end{array} \right.$$ 

$$= \left\{ \begin{array}{ll}
p - (F_{p_j}^{-1}) \circ (F_{p_i}^{-1})^{-1}(p), & \text{if } (F_{p_j}^{-1})^{-1}(p) > (F_{p_i}^{-1})^{-1}(p), \\
p - (F_{p_j}^{-1}) \circ (F_{p_i}^{-1})^{-1}(p), & \text{otherwise},
\end{array} \right.$$ 

$$= |p - (F_{p_j}^{-1}) \circ (F_{p_i}^{-1})^{-1}(p)|.$$ 

39
Thus

\[
\frac{1}{T} \sum_{t=1}^{T} \int_0^1 \left| (F_{j,t}^{-1} \ominus F_{\mu_j}^{-1})(p) - (F_{j,t}^{-1} \ominus F_{j,\Theta})(p) \right| \, dp
\]

is a continuous mapping, thus by continuous mapping theorem exists for all \( T \) converges to \( \Gamma \). Exchange(b) comes from that for all \( p \in (0,1) \), \( |p_0 - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p_0)| = O(\frac{1}{\sqrt{T}}) \). Since \( |p_0 - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p_0)| \) is a bounded sequence uniformly over \( T \), \( \mathbb{E} |p_0 - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p_0)| \) converges to 0. Thus, \( \mathbb{E} |p - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p)| \) converges pointwise to 0. On the other hand, note that for any fixed \( \epsilon > 0 \), we have

\[
\lim_{T \to \infty} \mathbb{P} \left( \int_0^1 |p - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p)| \, dp > \epsilon \right) = 0.
\]

By Lemma 4.2, we have for any fixed \( p_0 \in (0,1) \), \( |p_0 - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p_0)| = O(\frac{1}{\sqrt{T}}) \). Since \( |p_0 - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p_0)| \) is a bounded sequence uniformly over \( T \), \( \mathbb{E} |p_0 - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p_0)| \) converges to 0. Thus, \( \mathbb{E} |p - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p)| \) converges pointwise to 0. On the other hand, note that for any fixed \( \epsilon > 0 \), we have

\[
\lim_{T \to \infty} \mathbb{P} \left( \int_0^1 |p - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p)| \, dp > \epsilon \right) = 0.
\]

Exchange(b) comes from that \( |p - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p)| \) is bounded for all \( p \in (0,1) \), thus \( \mathbb{E} |p - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p)| \) exists for all \( p \in (0,1) \). Moreover, \( \mathbb{E} |p - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p)| \) is bounded uniformly over \( p \in (0,1) \), thus we can furthermore exchange the integral and the limit by bounded convergence theorem, which brings to Equation (c). Therefore,

\[
\int_0^1 |p - (F_{\mu_j}^{-1}) (F_{j,\Theta})^{-1}(p)| \, dp \to 0.
\]

This implies \( \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \ominus F_{\mu_j}^{-1} - F_{j,t}^{-1} \ominus F_{j,\Theta}, F_{t,l}^{-1} \ominus F_{\mu_l}^{-1} \rangle_{Leb} \to 0 \). Since the proof above is valid for all \( j, l = 1, ..., N \), we have immediately \( \frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \ominus F_{j,\Theta}, F_{t,l}^{-1} \ominus F_{l,\Theta} \rangle_{Leb} \to 0 \). Combined the two convergence results in Equation (7.12) gives \( \hat{\Gamma}(0)_{j,l} - \tilde{\Gamma}(0)_{j,l} \to 0 \), for all \( j, l = 1, ..., N \).

Analogously, we can prove \( \hat{\Gamma}(1)_{j,l} - \tilde{\Gamma}(1)_{j,l} \to 0 \), for all \( j, l = 1, ..., N \). Since matrix inversion is a continuous mapping, thus by continuous mapping theorem

\[
\hat{\Gamma}(0)^{-1} - \tilde{\Gamma}(0)^{-1} \to 0, \quad j, l = 1, ..., N.
\]

Representations of \( \hat{\Lambda}_o, \hat{\Lambda} \) bring to \( \hat{\Lambda}_o - \hat{\Lambda} \to 0 \).

The last term in Equation (7.11) implies that it is sufficient to obtain

\[
\frac{1}{T} \sum_{t=1}^{T} \langle F_{j,t}^{-1} \ominus F_{j,\Theta}, F_{t,l}^{-1} \ominus F_{t,\Theta} \rangle_{Leb} - \mathbb{E} \langle F_{j,t}^{-1} \ominus F_{j,\Theta}, F_{t,l}^{-1} \ominus F_{t,\Theta} \rangle_{Leb} \to 0,
\]

to prove \( \hat{\Lambda}_o - A \to 0 \). Thus we do not have to achieve the primary result until \( \hat{\Lambda} - A \to 0 \). However, to show clearly the logic of the proof of Theorem 4.3, and the difference of the random objects involved, especially \( \hat{F}_t \) and \( \tilde{F}_t \), we complete the result for \( \hat{A} \) and emphasize it in Theorem 4.3.
7.10 Proof of Theorem 4.4

The Lagrangian of Problem (4.8) is given by

\[ L(\beta, \Lambda) = f_T(\beta) + \sum_{i=1}^{m} \lambda_i f_i(\beta) + \sum_{j=1}^{p} \nu_j h_j(\beta). \]

Let \(\Lambda^*\) be a dual solution, then the strong duality implies the primal solution \(\hat{\beta}\) minimizes \(L(\beta, \Lambda^*)\), with \(\sum_{i=1}^{m} \lambda_i^* f_i(\beta) + \sum_{j=1}^{p} \nu_j^* h_j(\beta) = 0\), see Boyd et al. (2004, Section 5.5.2). Therefore we have

\[ L(\hat{\beta}, \Lambda^*) \leq L(\beta^*, \Lambda^*). \]

Furthermore,

\[ f_T(\hat{\beta}) \leq f_T(\beta^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\beta^*) + \sum_{j=1}^{p} \nu_j^* h_j(\beta^*). \]

We subtract \(f_T(\hat{\beta}_0)\) from both sides of the inequality above, which gives

\[ 0 \leq f_T(\hat{\beta}) - f_T(\hat{\beta}_0) \leq f_T(\beta^*) - f_T(\hat{\beta}_0) + \sum_{i=1}^{m} \lambda_i^* f_i(\beta^*) + \sum_{j=1}^{p} \nu_j^* h_j(\beta^*). \]

(7.13)

Note that the non-negativity comes from that \(\hat{\beta}_0\) is the minimizer of \(f_T\).

On the other hand, because \(f_T\) is strongly convex with the constant \(\mu_T\), thus for any \(s \in \partial f_T(\hat{\beta}_0)\), there is (Zhou, 2018, Lemma 3 (iii)),

\[ f_T(\beta^*) - f_T(\hat{\beta}_0) \leq \langle s, \beta^* - \hat{\beta}_0 \rangle + \frac{1}{2\mu_T} \|\beta^* - \hat{\beta}_0\|^2_{\ell_2}. \]

(7.14)

Because \(\hat{\beta}_0\) is the minimizer of \(f_T\), thus \(0 \leq f_T(\beta^*) - f_T(\hat{\beta}_0)\). Moreover, \(0 \in f_T(\hat{\beta}_0)\), thus, we pick \(s = 0\), Equation (7.14) becomes

\[ 0 \leq f_T(\beta^*) - f_T(\hat{\beta}_0) \leq \frac{1}{2\mu_T} \|\beta^* - \hat{\beta}_0\|^2_{\ell_2}. \]

Since \(\|\beta^* - \hat{\beta}_0\|^2_{\ell_2} = O_p(r_T)\), and \(\frac{1}{2\mu_T}\) is stochastically bounded, thus \(f_T(\beta^*) - f_T(\hat{\beta}_0) = O_p(r_T)\).

Furthermore, \(\beta^*\) satisfies the constraints, thus \(\sum_{i=1}^{m} \lambda_i^* f_i(\beta^*) + \sum_{j=1}^{p} \nu_j^* h_j(\beta^*) \leq 0\), combined with the previous convergence result in Inequality (7.13), we have

\[ f_T(\hat{\beta}) - f_T(\hat{\beta}_0) = O_p(r_T). \]

Use once again the strong convexity of \(f_T\), we obtain (Zhou, 2018, Lemma 2 (iii))

\[ f_T(\hat{\beta}) - f_T(\hat{\beta}_0) \geq \langle s, \beta^* - \hat{\beta}_0 \rangle + \frac{\mu_T}{2} \|\beta^* - \hat{\beta}_0\|^2_{\ell_2}, \quad s \in \partial f_T(\hat{\beta}_0). \]

(7.15)

Plug \(s = 0\) in Equation (7.15), we get

\[ \frac{2}{\mu_T} (f_T(\hat{\beta}) - f_T(\hat{\beta}_0)) \geq \|\beta^* - \hat{\beta}_0\|^2_{\ell_2}. \]

The stochastic boundedness and the convergence finally leads to \(\hat{\beta} - \hat{\beta}_0 = O_p(r_T)\). 

\[ \square \]
7.11 Proof of Theorem 4.5

Note that $\mathbf{\hat{\Gamma}}(0)$ is nonsingular at any given time $T$, thus the objective function in Problem (4.9) is strongly convex with a strong convexity constant equal to $2\lambda_{min}^T$, with $\lambda_{min}^T$ the smallest eigenvalue of $\mathbf{\hat{\Gamma}}(0)$. On the other hand, because $\mathbf{\hat{\Gamma}}(0) \cong \mathbf{\Gamma}(0)$ element-wise, meanwhile eigenvalues are continuous mappings of a matrix’s entries, by applying continuous mapping theorem, we have $\lambda_{min}^T \equiv \lambda_{min}$, where $\lambda_{min}$ is the smallest eigenvalue of $\mathbf{\Gamma}(0)$. Since $\lambda_{min}$ is nonzero, this implies $1/\lambda_{min}^T \overset{p}{\rightarrow} 1/\lambda_{min}$, then $1/\lambda_{min}^T = 1/\lambda_{min} + o_p(1)$ implies $1/\lambda_{min}^T = O_p(1)$. 