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

Yiye Jiang [1]

Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France
yiye.jiang@inria.fr

Résumé. Nous proposons un nouveau modèle auto-régressif pour l'analyse statistique des séries chronologiques distribuées multivariées. Les données d'intérêt consistent en une collection de plusieurs séries de mesures de probabilité supportées sur un intervalle borné de la ligne réelle, et qui sont indexées par des instants de temps. Les mesures de probabilité sont modélisées comme des objets aléatoires dans un espace de Wasserstein. Nous établissons le modèle auto-régressif dans l'espace tangent à la mesure de Lebesgue en centrant d'abord toutes les mesures brutes de manière à ce que leur moyenne de Fréchet corresponde à la mesure de Lebesgue. En utilisant la théorie des systèmes de fonctions aléatoires itérés, des résultats sur l'existence, l'unicité et la stationnarité de la solution d'un tel modèle sont fournis. Nous proposons également un estimateur consistant pour le coefficient du modèle. En plus de l'analyse de données simulées, le modèle proposé est illustré avec deux ensembles de données réelles, constitués d'observations des distributions d'âge de différents pays / états pour l'un, et du réseau de vélos en libre-service à Paris pour l'autre. Enfin, grâce aux contraintes du simplexe que nous imposons sur les coefficients du modèle, l'estimateur proposé, qui est appris sous ces contraintes, a naturellement une structure peu dense. Cette structure permet en outre d'appliquer le modèle proposé à l'apprentissage d'un graphe de dépendance temporelle à partir de séries chronologiques distribuées multivariées.

Mots-clés. Modèles autorégressifs, espaces de Wasserstein, graph learning, analyse de données distribuées.

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 a 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 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 / states 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.

Keywords. Autoregressive models, Wasserstein spaces, graph learning, distributional data analysis

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. In Figure 1, we illustrate the data type with a real data set.

[1] This work was conducted while the author was preparing her PhD at Université de Bordeaux.

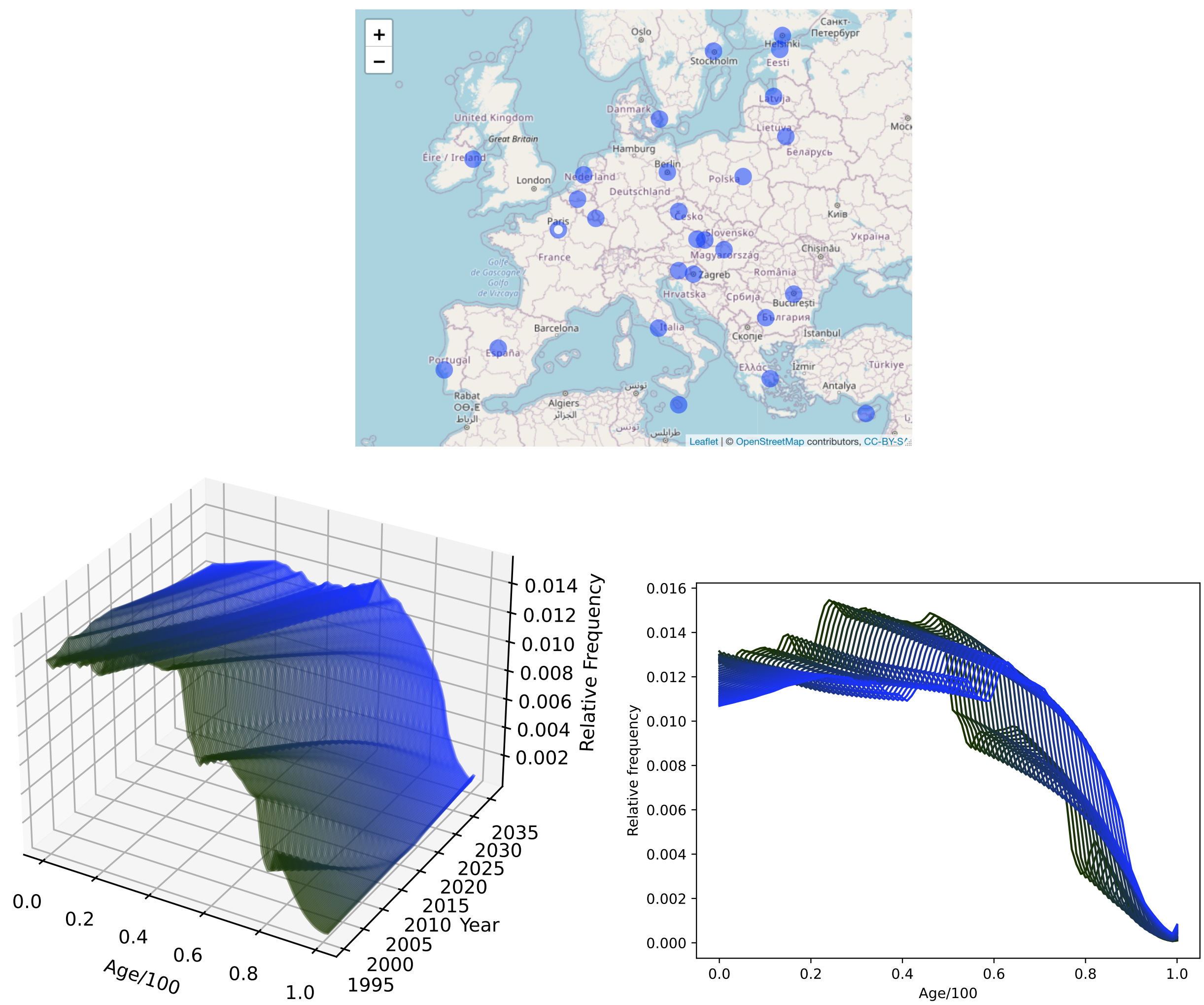

Figure 1: *Annual records of age distributions of countries/states.* On the top are 27 countries in the European union. A sequence of age distribution is recorded at each country over years. For example, at the bottom we illustrate the sequence of France, where one distribution supported over $[0, 1]$ 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.

Since distributions can be characterized by functions, such as densities, quantile functions, and cumulative distribution functions, 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 the nonlinear constraints, such as monotonicity and positivity, the representing functions of distributions do not constitute linear spaces. Consequently, basic notions needed by the classical tools, 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 distributions to unconstrained functions 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, 2019) of distributions. For further reading on the topic of statistics in Wasserstein space, we refer to Bigot (2020); Panaretos and Zemel (2020); Petersen et al. (2022).

Relying on the theories of Wasserstein spaces, Chen et al. (2021); Zhang et al. (2021); Zhu and Müller (2021) have successfully extended one of the most important models in classical time series analysis, autoregressive (AR) model, to univariate distributional (with closed bounded supports) time series, namely, $(\boldsymbol{\mu}_t)_{t\in\mathbb{Z}} \in \mathcal{P}([0,1])$. In this work, we will furthermore extend the AR model to the multivariate distributional (with closed bounded supports) case, that is, $(\boldsymbol{\mu}_t^i)_{t\in\mathbb{Z}}$ for $i = 1,\ldots,N \in \mathcal{P}([0,1])$, based on the tools of Wasserstein spaces as well. In the following, we firstly recall the classical AR models. Next we introduce the necessary notions of the Wasserstein spaces to present our model. Then we present our model and its theoretical properties. Fourthly, we propose a consistent estimator for the model coefficient. Lastly we fit our model over a real data sets, and show the numerical results.

2 Backgrounds

2.1 Vector autoregressive models of order 1

In this section, we recall the classical AR models in the multivariate setting. Let $\boldsymbol{x}_{it} \in \mathbb{R}$, $t \in \mathbb{Z}$, $i = 1,\ldots,N$, a multivariate time series, and assume $\mathbb{E}\boldsymbol{x}_{it} = u_i$ exists and time invariant, then the vector autoregressive model of order 1 (VAR(1)) writes as

$$\boldsymbol{x}_{it} - u_i = \sum_{j=1}^{N} A_{ij}(\boldsymbol{x}_{j,t-1} - u_j) + \boldsymbol{\epsilon}_{it},$$

where $\boldsymbol{\epsilon}_{it} \sim WN(0,\sigma^2)$ with σ non-zero, and $\epsilon_{it}, \epsilon_{jt}, i \neq j$ are not correlated. Then the goal of this work is to extend this model by "replacing" each scalar $\boldsymbol{x}_{it}$ by a univariate distribution $\mu_i^t \in \mathcal{P}(\mathbb{R})$. The main ingredient is using the geometrical notions of Wasserstein space. In the next section, we introduce the ones which are needed by our models.

2.2 Notions in Wasserstein spaces

We firstly define the Wasserstein spaces. Since we will need the notion of Tangent space, we use the spaces of 2-Wasserstein distances. More specifically, we focus on the space

$$\mathcal{W}_2(\mathbb{R}) = \left\{\mu \in \mathcal{P}(\mathbb{R}) \middle| \int_{\mathbb{R}} x^2 d\mu(x) < \infty\right\}, \tag{2.1}$$

endowed with the 2-Wasserstein distance[2]

$$d_W(\mu,\nu) = \int_0^1 \left(F_\mu^{-1}(u) - F_\nu^{-1}(u)\right)^2 du, \tag{2.2}$$

[2]In general, the Wasserstein distances are defined by the Kantorovich problems. However, in the special case we consider here, that is, the cost is quadratic and the domain is $\mathbb{R}$, the solution of Kantorovich problem is explicit and given by the definition in Equation (2.2).

where $F_\mu^{-1}(u), F_\nu^{-1}(u)$ are the quantile functions of μ and ν.

The space defined in Equation (2.1) is not linear. Nevertheless, Wasserstein spaces of order 2 permit the notion of Tangent spaces (Ambrosio et al., 2008; Bigot et al., 2017; Zemel and Panaretos, 2019), where the linear operations are enabled again. To present the Tangent spaces, we firstly recall the notion of pushforward for a pair of measures.

Given two measurable spaces $(\mathcal{X}_i, \sigma_i)$, $i = 1, 2$, a measurable function $f : \mathcal{X}_1 \to \mathcal{X}_2$, and a measure $\mu_1 : \sigma_1 \to [0, +\infty]$, the pushforward measure of μ_1 by f denoted by $f\#\mu_1$ is defined as

$$f\#\mu_1(A) = \mu_1(\{x : f(x) \in A\}), \forall A \in \sigma_2.$$

Let $\gamma \in \mathcal{W}_2$ be an atomless measure, that is, it possesses a continuous cumulative distribution function F_γ, then the tangent space at γ is defined as follows.

Definition 2.1.

$$\mathrm{Tan}_\gamma = \overline{\{t(T_\gamma^\mu - id) : \mu \in \mathcal{W}_2,\ t > 0\}}^{\mathcal{L}_\gamma^2},$$

where $T_\gamma^\mu = F_\mu^{-1} \circ F_\gamma$.

Note that Tan_γ is endowed with the inner product $\langle \cdot, \cdot \rangle_\gamma$ defined by

$$\langle f, g \rangle_\gamma := \int_{\mathbb{R}} f(x)g(x)\, d\gamma(x),\ f, g \in \mathcal{L}_\gamma^2(\mathbb{R}),$$

and the induced norm $\| \cdot \|_\gamma$. By the definition of pushforward, it is easy to verify that T_γ^μ pushforwards μ to γ. In fact, it is the optimal pushforward from μ to γ in the sense that it results in the least transport cost. The technical explanation see for example (Panaretos and Zemel, 2020, Chapter 1).

We now define the exponential and logarithmic maps, which are the ways to communicate between a Tangent space and its Wasserstein space.

Definition 2.2. *The logarithmic map* $\mathrm{Log}_\gamma : \mathcal{W}_2 \to \mathrm{Tan}_\gamma$ *is defined as*

$$\mathrm{Log}_\gamma \mu = T_\gamma^\mu - id.$$

The exponential map $\mathrm{Exp}_\gamma : \mathrm{Tan}_\gamma \to \mathcal{W}_2$ *is defined as*

$$\mathrm{Exp}_\gamma g = (g + id)\#\gamma.$$

We can see that, given a reference measure, all the measures in the nonlinear Wasserstein space can be mapped into the linear Tangent space of the reference. This implies that we can construct the extended AR model in a Tangent space. Before we move onto the proposed model in the next section, lastly, we introduce the notion of Fréchet mean, which will replace the classical expectation.

Definition 2.3. *Let* $\mu_1, \ldots, \mu_T$ *be measures in* $\mathcal{W}_2$*. The empirical Fréchet mean of* $\mu_1, \ldots, \mu_T$*, denoted by* $\bar{\mu}$*, is defined as the unique minimizer of*

$$\min_{\nu \in \mathcal{W}_2} \frac{1}{T} \sum_{t=1}^{T} d_W^2(\mu_t, \nu).$$

Definition 2.4. *A random measure* $\boldsymbol{\mu}$ *is any measurable map from a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *to the metric space* $\mathcal{W}_2$*, endowed with its Borel* σ*-algebra.*

Definition 2.5. *Let* $\boldsymbol{\mu}$ *be a random measure from probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *to* $\mathcal{W}_2$*. Assume that* $\boldsymbol{\mu}$ *is square integrable, namely* $\mathbb{E}d_W^2(\boldsymbol{\mu}, \nu) < \infty$ *for some (thus for all)* $\nu \in \mathcal{W}_2$*. Then, the population Fréchet mean of* $\boldsymbol{\mu}$*, denoted by* $\mu_\oplus$*, is defined as the unique minimizer of*

$$\min_{\nu \in \mathcal{W}_2} \mathbb{E}\left[d_W^2(\boldsymbol{\mu}, \nu)\right].$$

For our Wasserstein distance, $\bar{\mu}$ and $\mu_\oplus$ admit simple expressions through their quantile functions:

$$F_{\bar{\mu}}^{-1}(p) = \frac{1}{T} \sum_{t=1}^{T} F_{\mu_t}^{-1}(p),\ F_{\mu_\oplus}^{-1}(p) = \mathbb{E}\left[F_{\boldsymbol{\mu}}^{-1}(p)\right],\ p \in (0, 1). \tag{2.3}$$

3 Wasserstein multivariate auto-regressive Models

Given the data of a multivariate distributional time series, $\boldsymbol{\mu}_t^i$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$, we primarily assume that

Assumption A1. $\boldsymbol{\mu}_t^i \in \mathcal{W}_2(\mathbb{R})$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$, *and* $\mathbb{E}_\oplus \boldsymbol{\mu}_t^i = \mu_{i,\oplus}, t \in \mathbb{Z}$.

We now extend the VAR(1) models for such data in two steps. In the first step, we will fix a reference measure, on which we will set up a Tangent space. A common choice in literature is Fréchet mean, such as the population Fréchet mean for a set of iid samples. Similarly, the previous works on univariate distributional time series took the population Fréchet mean of the series $\mathbb{E}_\oplus \boldsymbol{\mu}_t$ as the reference. However, as we deal with N series, we have N population Fréchet means, which makes the natural choice not obvious. Thus in this step, we are inspired by the centering operation $\boldsymbol{x}_{j,t-1} - u_j$ in the VAR(1) models and propose a way to center our samples $\boldsymbol{\mu}_t^i$, so that the centered samples will share the common Fréchet mean. The proposed centering applies to general random distributions. Given a random distribution $\boldsymbol{\mu}$ with its quantile function $\boldsymbol{F}^{-1}$, the centered distribution $\tilde{\boldsymbol{\mu}}$ is defined by its quantile function $\tilde{\boldsymbol{F}}^{-1}$ as

$$\widetilde{\boldsymbol{F}}^{-1} = \boldsymbol{F}^{-1} \circ (F_\oplus^{-1})^{-1}, \tag{3.1}$$

where $F_\oplus^{-1}(p) = \mathbb{E}\boldsymbol{F}^{-1}(p)$, according to the fact given in Equation (2.3), $F_\oplus^{-1}$ is actually the quantile function of Fréchet mean $\mathbb{E}_\oplus \boldsymbol{\mu}$. Our goal is to make the centered distribution have $U(0,1)$ as its Fréchet mean, that is $\mathbb{E}_\oplus \widetilde{\boldsymbol{\mu}} = U(0,1)$. To this end, we need to assume[3] that $\boldsymbol{\mu}$ is supported on $[0,1]$, essentially a closed bounded interval of $\mathbb{R}$. We need to apply this centering transformation on each distribution in our data, thus we add the following data assumption.

Assumption A2. *All* $\boldsymbol{\mu}_t^i$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$ *are supported on* $[0,1]$.

Therefore, the centering operation in our extended VAR(1) models is given by

$$\widetilde{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{F}_{i,t}^{-1} \circ (F_{i,\oplus}^{-1})^{-1}, \tag{3.2}$$

where $\boldsymbol{F}_{i,t}^{-1}$ and $\boldsymbol{\mu}_t^i$ are respectively the quantile functions of $\boldsymbol{\mu}_t^i$ and $\mu_{i,\oplus}$, and $\widetilde{\boldsymbol{F}}_{i,t}^{-1}$ defines the centered distribution $\widetilde{\boldsymbol{\mu}}_t^i$. Then in the second step, we will set up the regressive formula for $\widetilde{\boldsymbol{\mu}}_t^i$. The common Fréchet mean, namely $U(0,1)$, therefore will be taken as the reference measure. We propose an extension of VAR(1) model as

$$\widetilde{\boldsymbol{\mu}}_t^i = \boldsymbol{\epsilon}_{i,t} \# \operatorname{Exp}_{Leb}\left(\sum_{j=1}^{N} A_{ij} \operatorname{Log}_{Leb} \widetilde{\boldsymbol{\mu}}_{t-1}^j\right), \quad t \in \mathbb{Z},\, i = 1, \ldots, N,$$

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

$$\begin{aligned} &\Pi = \{F^{-1} : [0,1] \to [0,1], \text{ such that } F^{-1}\big|_{(0,1)} \in \operatorname{Log}_{Leb} \mathcal{W} + id, \\ &F^{-1}(0) := \inf\{x \in [0,1] : F(x) > 0\}, \text{and} F^{-1}(1) := \sup\{x \in [0,1] : F(x) < 1\}\}, \end{aligned}$$

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

$$\mathbb{E}\left[\boldsymbol{\epsilon}_{i,t}(x)\right] = x,\; x \in [0,1].$$

Note that, we first map the predictor measures $\widetilde{\boldsymbol{\mu}}_{t-1}^j$ into the Tangent space of $U(0,1)$ namely the Lebesgue measure over $[0,1]$ as in our notation. Then over the corresponding Tangent vectors $\operatorname{Log}_{Leb} \widetilde{\boldsymbol{\mu}}_{t-1}^j$, we apply the same regressive operation as in classical VAR(1) model. The proposed model is not yet identifiable since the exponential map is not injective. Therefore, we need the following assumption.

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

[3] For the technical explanation, we refer to Jiang (2022).

Given the identifiability, the model allows another representation in terms of quantile functions as follows.

$$\widetilde{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{\epsilon}_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\widetilde{\boldsymbol{F}}_{j,t-1}^{-1} - id \right) + id \right], \quad t \in \mathbb{Z},\, i = 1, \ldots, N, \tag{3.3}$$

where $\widetilde{\boldsymbol{F}}_{i,t}^{-1}$ is the quantile function of $\widetilde{\boldsymbol{\mu}}_{i,t}^{-1}$.

4 Existence, uniqueness and stationarity

Now, we present the theoretical properties of the proposed model. The main tool in the derivation is a general result from Wu and Shao (2004), where conditions for an iterated random function (IRF) system in a metric space to be stable are given. Applying the result allows us to show the existence of the solution of our system. Then we build from that, and we are able to show the uniqueness and stationarity of the solution. We rely on the quantile representation (3.3) in the theoretical development by considering it as an IRF system in the following metric space.

$$(\mathcal{X}\,,\, d) := (\mathcal{T}\,,\, \|\cdot\|_{Leb})^{\otimes N}\,,$$

where $\mathcal{T} := \mathrm{Log}_{Leb}\,\mathcal{W}_2(\mathbb{R}) + id$ is the space of all quantile functions of $\mathcal{W}_2(\mathbb{R})$, equipped with the norm $\|\cdot\|_{Leb}$. Thus, we have

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

We propose the following assumptions, which allow our system to satisfy the stability conditions required by Wu and Shao (2004).

Assumption A4. $\mathbb{E}\left[\boldsymbol{\epsilon}_{i,t}(x) - \boldsymbol{\epsilon}_{i,t}(y)\right]^2 \leqslant L^2(x-y)^2,\ \forall x, y \in [0,1], \quad t \in \mathbb{Z},\, i = 1, \ldots, N,$

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

Note that, Assumption A4 implies that $\boldsymbol{\epsilon}_{i,t}$ 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 stability condition for VAR(1) models. Theorem 4.1 states the existence and uniqueness results.

Theorem 4.1. *Under Assumptions A3, A4 and A5, the IRF system (3.3) almost surely admits a solution $\mathbf{X}_t$, $t \in \mathbb{Z}$, with the same marginal distribution π, namely, $\mathbf{X}_t \stackrel{d}{=} \pi$, $\forall\, t \in \mathbb{Z}$, where the notation $\stackrel{d}{=}$ means equality in distribution. Moreover, if there exists another solution $\boldsymbol{S}_t$, $t \in \mathbb{Z}$, then for all $t \in \mathbb{Z}$*

$$\mathbf{X}_t \stackrel{d}{=} \boldsymbol{S}_t, \quad \textit{almost surely.}$$

We are able to show that this unique (in the sense of distributions) solution is furthermore stationary[4] as a functional process 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 Hilbert space exists, with corresponding inner product given by

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^{N} \langle \mathbf{X}_i, \mathbf{Y}_i \rangle_{Leb}\,.$$

We recall the conventional definition of stationarity for process in a separable Hilbert space, see for example Zhang et al. (2021, Definition 2.2).

[4]Different from stability, stationarity requires the second order invariance with respect to some inner product, thus it is a notion defined in a Hilbert space.

Definition 4.1. *A random process $\{\mathbf{V}_t\}_t$ in a separable Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ is said to be stationary if the following properties are satisfied.*

1. $\mathbb{E}\|\mathbf{V}_t\|^2 < \infty$.
2. *The Hilbert mean $U := \mathbb{E}[\mathbf{V}_t]$ does not depend on t.*
3. *The auto-covariance operators defined as*

$$\mathcal{G}_{t,t-h}(V) := \mathbb{E}\langle \mathbf{V}_t - U, V\rangle (\mathbf{V}_{t-h} - U), \quad V \in \mathcal{H},$$

do not depend on t, that is $\mathcal{G}_{t,t-h}(V) = \mathcal{G}_{0,-h}(V)$ for all t.

Then, Theorem 4.2 below gives the stationarity result.

Theorem 4.2. *The unique solution given in Theorem 4.1 is stationary as a random process in $(\mathcal{X}, \langle\cdot,\cdot\rangle)$ in the sense of Definition 4.1.*

5 Estimation of the regression coefficients

In this section, we develop a consistent estimator of coefficient A, given $T+1$ samples $\boldsymbol{\mu}_t^i$, $t = 0, 1, \ldots, T, i = 1, \ldots, N$. We calculate the estimator in two steps as well. Firstly, we estimate the exact centered data $\widetilde{\boldsymbol{\mu}}_t^i$ defined in Equation (3.2) using empirical Fréchet means. We denote the estimates by $\widehat{\boldsymbol{\mu}}_t^i$, which are defined by their quantile functions as:

$$\widehat{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{F}_{i,t}^{-1} \circ (\boldsymbol{F}_{\bar{\mu}_i}^{-1})^{-1}, \quad \text{where } \boldsymbol{F}_{\bar{\mu}_i}^{-1} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{F}_{\mu_i,t}^{-1},$$

Then the proposed estimator is defined through the following least squares formula.

$$\widehat{\boldsymbol{A}}_{i:} = \underset{A_{i:} \in B_+^1}{\arg\min} \frac{1}{T}\sum_{t=1}^{T} \left\| \widehat{\boldsymbol{F}}_{i,t}^{-1} - \sum_{j=1}^{N} A_{ij}\left(\widehat{\boldsymbol{F}}_{j,t-1}^{-1} - id\right) - id \right\|_{Leb}^2, \quad i = 1, \ldots, N, \tag{5.1}$$

where B_+^1 is N-dimensional simplex, that is the nonnegative orthant of the ℓ_1 unit ball B^1 in $\mathbb{R}^N$, corresponding to Assumption A3. Thus, an important advantage of this constraint is to promote sparsity in $\widehat{\boldsymbol{A}}_{i:}$. When using the proposed model in graph learning, the retrieved (directed) graph from $\widehat{\boldsymbol{A}}$ will be naturally sparse. The optimisation problem (5.1) 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). Theorem 5.1 shows that the proposed estimator is consistent.

Theorem 5.1. *Assume that $\boldsymbol{\mu}_t^i$, $i = 1, \ldots, N$ satisfy Assumption A1 for $t = 0, 1, \ldots, T$, and the transformed sequence $\widetilde{\boldsymbol{F}}_{i,t}^{-1}$ satisfies Model (3.3) with Assumption A3 true. Suppose additionally that $(\widetilde{\boldsymbol{F}}_{i,0}^{-1})_{i=1}^N \overset{d}{=} \pi$ with π the stationary distribution defined in Theorem 4.1. Given Assumptions A4 and A5 hold true, and the following $N \times N$ matrix $\Gamma(0)$ is nonsingular*

$$[\Gamma(0)]_{j,l} = \mathbb{E}\langle \widetilde{\boldsymbol{F}}_{j,t-1}^{-1} - id\,,\, \widetilde{\boldsymbol{F}}_{l,t-1}^{-1} - id\rangle_{Leb},$$

we have

$$\widehat{\boldsymbol{A}} - A \xrightarrow{p} 0.$$

6 Numerical experiment

In this section, we fit our model with a real data set which records annual age distributions of countries/states. The data set has been illustrated in Figure 1 in the introduction. We represent the distribution of age population, of country i, at year t, by $\boldsymbol{\mu}_t^i$, with $T = 1, ..., 40$ and $i = 1, ..., 34$. To justify the proposed model and to illustrate its application in graph learning from distributional time series, we visualize the estimation of regression coefficients A on the real geographical map in Figure 2, so as to inspect the learned patterns. For details of the data and experimental settings as well as more experiment results, we refer to Jiang (2022).

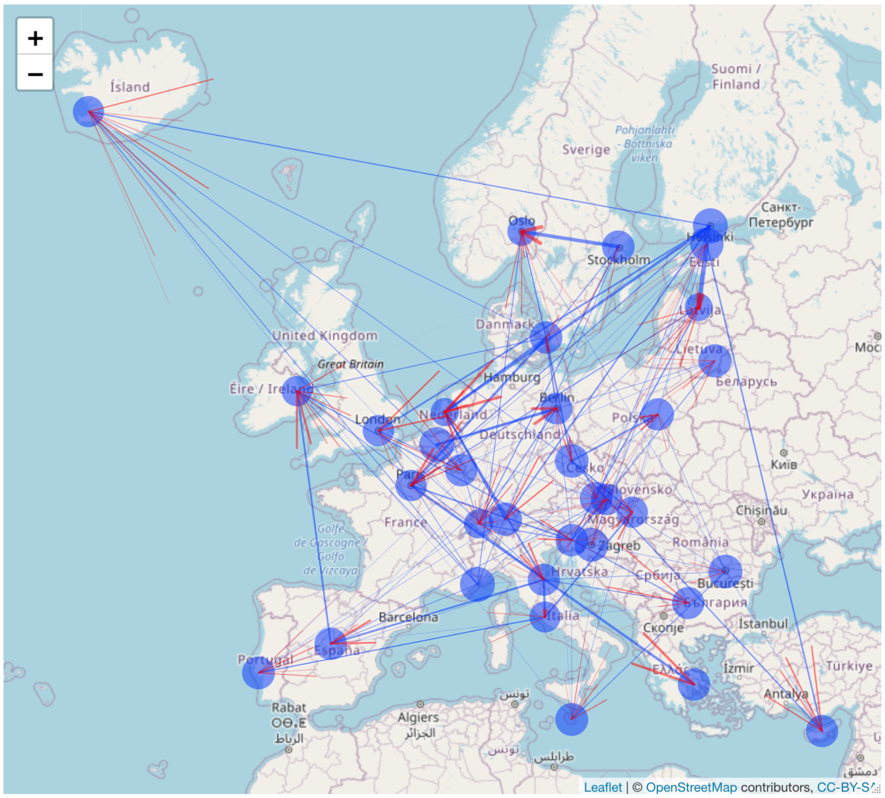

Figure 2: *Inferred age structure graph.* The non-zero coefficients A_{ij} are represented by the weighted directed edges from node j to node i. Thicker arrow corresponds to larger weights. The blue circles around nodes represent the weights of self-loop.

We can notice that for all countries $i \in \{1, ..., 34\}$, the coefficients of self-loop A_{ii} dominate others A_{ij}, $i \neq j$. 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 different countries' age distributions. The first two largest international coefficients are: Estonia → Latvia, and Sweden → Norway. To justify these inferred patterns, we plot the time series of the four countries in Figure 3. We can see that the countries on the same row (their regression coefficients have large values) have similar time series along time, by contrast, those on the same column (their regression coefficients have small or zero values) differ a lot. All these observations strongly support the usefulness of our model.

References

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

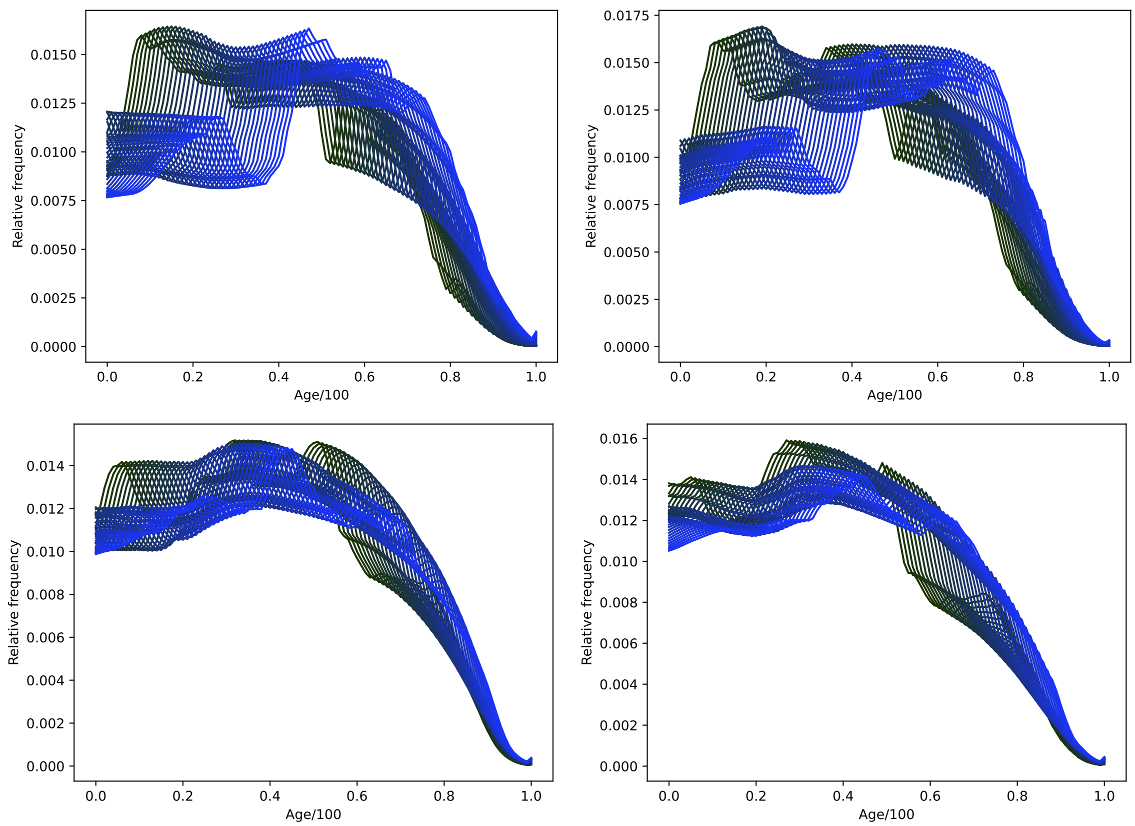

Figure 3: *Evolution of age structure from* 1996 *to* 2036 *(projected) of Estonia (left upper), Latvia (right upper), Sweden (left bottom) versus Norway (right bottom).* Each curve connects the 101 relative frequencies from $0, 1/100, 2/100, ..., 1$, which represents the age structure of a considered year. Lighter curves correspond to more recent years.

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.

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

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

Y. Jiang. Wasserstein multivariate auto-regressive models for modeling distributional time series and its application in graph learning. *arXiv preprint arXiv:2207.05442*, 2022.

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. Wasserstein covariance for multiple random densities. *Biometrika*, 106(2): 339–351, 2019.

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 ℓ_1 or simplex constraints via isotonic regression. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 2031–2036. IEEE, 2015.

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

Y. Zemel and V. M. Panaretos. Fréchet means and procrustes analysis in wasserstein space. 2019.

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

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