Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **00**, 1–22
https://doi.org/10.1093/jrsssb/qkad051

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology

Original Article

Autoregressive optimal transport models

Changbo Zhu[1] and Hans-Georg Müller[2]

[1]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, USA
[2]Department of Statistics, University of California, Davis, Davis, CA 95616, USA

Address for correspondence: Changbo Zhu, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, USA. Email: czhu4@nd.edu

Abstract

Series of univariate distributions indexed by equally spaced time points are ubiquitous in applications and their analysis constitutes one of the challenges of the emerging field of distributional data analysis. To quantify such distributional time series, we propose a class of intrinsic autoregressive models that operate in the space of optimal transport maps. The autoregressive transport models that we introduce here are based on regressing optimal transport maps on each other, where predictors can be transport maps from an overall barycenter to a current distribution or transport maps between past consecutive distributions of the distributional time series. Autoregressive transport models and their associated distributional regression models specify the link between predictor and response transport maps by moving along geodesics in Wasserstein space. These models emerge as natural extensions of the classical autoregressive models in Euclidean space. Unique stationary solutions of autoregressive transport models are shown to exist under a geometric moment contraction condition of Wu & Shao [(2004) Limit theorems for iterated random functions. *Journal of Applied Probability 41*, 425–436)], using properties of iterated random functions. We also discuss an extension to a varying coefficient model for first-order autoregressive transport models. In addition to simulations, the proposed models are illustrated with distributional time series of house prices across U.S. counties and annual summer temperature distributions.

Keywords: distributional data analysis, distributional regression, distributional time series, iterated random function, optimal transport, Wasserstein space

1 Introduction

Distributional data analysis (DDA) deals with data that include random distributions as data elements. While such data are prevalent in many applied problems (Matabuena & Petersen, 2021; Menafoglio et al., 2018), this area is still in its early development. An important instance where one encounters distributional data arises for sequences of dependent distributions that are indexed by discrete time. Such distributional time series are ubiquitous. For instance, the distribution of the log returns of the stocks included in a stock index is expected to contain more information than the index itself, which only conveys the mean of the distribution but not any further information inherent in the distribution such as quantiles. Elucidating the nature of such financial time series is for example of interest for risk management (Bekierman & Gribisch, 2021; Kokoszka et al., 2019). We will illustrate the proposed methods with the time series of distributions of house prices that are formed from U.S. county house price data and may inform economic policy (Bogin et al., 2019; Oikarinen et al., 2018) and also with time series of annual distributions of temperatures aggregated over the summer, where a rise in night time temperatures and more frequent extremes have been related to global warming.

Received: May 12, 2021. **Revised:** April 15, 2023. **Accepted:** April 17, 2023

Other pertinent examples include the analysis of sequences of age-at-death distributions over calendar years, which is instrumental for the study of human longevity (Mazzuco & Scarpa, 2015; Ouellette & Bourbeau, 2011; Shang & Hyndman, 2017) and also the study of the distributions of correlations between pairs of voxels within brain regions that can be derived from fMRI Bold signals (Petersen & Müller, 2016), where such distributions may be observed repeatedly for the same subject in longitudinal studies.

Distributions can be equivalently represented as either density, quantile or cumulative distribution functions, assuming that all of these exist. Each of these representations comes with certain constraints (for example, density functions are nonnegative and integrate to 1). An important observation is that the spaces where these objects live are nonlinear. As a consequence, common statistical tools that are available in linear function spaces such as the Hilbert space L^2 that is utilized in functional time series analysis (Bosq, 2000) are inadequate and there is a need for the development of adequate statistical methodology. It is the goal of this paper to contribute to the development of autoregressive models for one-dimensional distributions, given that autoregressive models are popular in time series analysis and have been also considered for distributional time series in recent work based on mapping to tangent spaces in the Wasserstein manifold (Chen et al., 2022; Zhang et al., 2022).

Existing approaches for distributional regression are based on various transformation approaches that include mapping the distributions into a Hilbert space as implemented in the log quantile distribution approach (Kokoszka et al., 2019; Petersen & Müller, 2016) or through logarithmic maps in the Wasserstein manifold (Chen et al., 2022), where one uses the Wasserstein metric in the distribution space and maps the distributions to a tangent space that is a L^2 space, anchored at a suitable distribution, often chosen as a barycenter. One then can implement functional regression models in the ensuing L^2 space, and analyse these models by employing parallel transport. While the log quantile distribution transformation approach to distributional regression can lead to large metric deformations, the tangent bundle approach is extrinsic and there are some difficulties with the required inverse exponential maps that are caused by the injectivity requirement that one needs to numerically address in finite sample situations. Various projection methods have been devised to tackle this problem (Bigot et al., 2017; Chen et al., 2022; Pegoraro & Beraha, 2022), while in other recent work on extrinsic modelling it has been ignored (Zhang et al., 2022), which can lead to inferior performance.

Since the autoregressive transport models we propose here are intrinsic, they bypass the construction of a tangent space and the ensuing problems with mapping and projection. For the case where only the responses are distributional but predictors are vectors, one can apply Fréchet regression (Petersen & Müller, 2019). Concurrently with this paper, a distributional regression model with one predictor was proposed for the independent case (Ghodrati & Panaretos, 2022), where the goal is to learn a single best transport map that maps the predictor distribution to the response distribution, so the model parameter is the transport map learned from the data. This is akin to fitting a linear regression model where the slope parameter is fixed at 1. A nice feature of this simple model is that finding the best transport map has been shown to be equivalent to an isotonic regression problem, which can be solved by standard optimization techniques.

In this paper, we propose a novel class of intrinsic distributional regression models for the autoregressive modelling of distributional time series. The proposed models are based on transports of the probability measures. The most popular notion of transport of distributions is optimal transport, which commonly refers to moving distributions along geodesics in the Wasserstein space, i.e., the space of distributions equipped with the Wasserstein metric. The key innovation in the proposed regression model is that both predictors and responses are taken to be transports of distributions, rather than distributions themselves, in contrast to the currently available distributional regression models. Our focus is on univariate distributions with bounded support on the real line, which is the most relevant case in statistical data analysis. Moreover, in data applications the distributions that are part of the data sample are not known a priori and in practice need to be estimated from data they generate by nonparametric methods. Such methods include kernel density estimation and related approaches, and for practical implementations a bounded interval that defines the domain needs to be fixed beforehand. For the relatively uncommon applications that require the distributions to be supported on the entire real line it is common practice to truncate the target distribution at a large enough interval and to target the truncated distribution, with negligible error.

Typical examples for predictor or response transports are the transports defined by pushing distributional barycenters (Fréchet means) forward to individual distributions, and in the distributional time series framework also the transports pushing the distribution at time $(j-1)$ to that at time j, which may serve as predictors for the transports pushing the distribution at time j forward to that at time $j+1$. The idea of considering transports rather than distributions as predictors or responses, especially transports from barycenters, is motivated by the classical simple linear regression model for scalar predictors and responses. This model can be written in transport form as $E(Y-\mu_Y|X-\mu_X)=\beta(X-\mu_X)$, where $\mu_Y=EY$, $\mu_X=EX$ and β is the slope parameter, where both responses $Y-\mu_Y$ and predictors $X-\mu_X$ can be interpreted as transports pushing the barycenters μ_X, μ_Y forward to the individual data X, Y. As we show here, this transport interpretation of linear regression provides a natural approach to extend classical regression to distributional regression modelling by regressing transports on each other.

We focus here on autoregressive transport models (ATM) that permit an inherent geometrical interpretation by relating geodesics in transport space to each other, where a first order ATM (or ATM(1)) connects transports related to time $(j-1)$ to transports related to time j. As in the independent case, geometric transport interpretations can also be applied to the case of scalar or vector time series in Euclidean space, motivating the extension to distributional time series where transports are very natural. One of our main results is the existence and uniqueness of a stationary solution for ATM(1) processes, for which we utilize the geometric-moment contraction condition (Wu & Shao, 2004) for iterated random maps. While the proposed models generally involve scalar coefficients and are well interpretable, we also consider an extension for ATM(1) processes, where the ATM features a functional rather than scalar coefficient. We show that this functional coefficient can also be estimated consistently from samples. The definition of ATMs of order p (ATM(p)) is obtained as a straightforward extension; these models possess a multi-layer structure. We demonstrate that ATMs are useful to capture the dynamic evolution of distributions for both real and synthetic data.

The rest of the paper is organized as follows. In Section 2, we provide some preliminary discussion on basic concepts such as Wasserstein space, optimal transport maps and geodesics. We also introduce addition and scalar multiplication operations for the space of transport maps. Section 3 includes methodology and theoretical results for ATM(1) models. Extensions to ATM(p) models and versions of ATM(1) models with functional coefficients are the topics of Sections 4 and 5. Numerical considerations and applications to simulated and real data can be found in Section 6. Conclusions are in Section 7, while the Online supplementary material contains proofs and technical details.

2 The space of transport maps

Define $\mathcal{W}$ to be the set of probability distributions on $(\mathcal{S}, \mathscr{B}(\mathcal{S}))$ with finite second moments, where $\mathcal{S}=[s_1, s_2]$ is a bounded closed interval in $\mathbb{R}$ and $\mathscr{B}(\mathcal{S})$ is the Borel σ-algebra on $\mathcal{S}$. We assume there is an underlying probability space $(\Omega, \mathcal{A}, P)$ of $\mathcal{W}$-valued random variables that induces a probability measure on the space $\mathcal{W}$ with respect to which we can calculate moments for random variables taking values in $\mathcal{W}$.

For any measurable function $T:\mathcal{S}\to\mathcal{S}$ and $\mu\in\mathcal{W}$, let $T_{\#}\mu$ denote the pushforward measure of μ, i.e., for any $B\in\mathscr{B}(\mathcal{S})$, $T_{\#}\mu(B)=\mu(\{x:T(x)\in B\})$. If μ_1 is absolutely continuous with respect to the Lebesgue measure, then the 2-Wasserstein metric ($d_{\mathcal{W}}$) on $\mathcal{W}$ can be written using the Monge formulation (Villani, 2003)

$$\begin{aligned} d_{\mathcal{W}}(\mu_1,\mu_2) &= \inf_{T:T_{\#}\mu_1=\mu_2}\left\{\int_{\mathcal{S}}(T(x)-x)^2\,\mathrm{d}\mu_1(x)\right\}^{1/2} \\ &= \left\{\int_{\mathcal{S}}(T_{12}(x)-x)^2\,\mathrm{d}\mu_1(x)\right\}^{1/2}=\left\{\int_0^1(F_2^{-1}(u)-F_1^{-1}(u))^2du\right\}^{1/2}. \end{aligned} \tag{1}$$

Here $\mu_1, \mu_2 \in \mathcal{W}$, $F_1 = F(\mu_1)$ and $F_2 = F(\mu_2)$ are the cumulative distribution functions (cdf) of μ_1, μ_2, respectively, and

$$F_1^{-1}(u) := \inf\{x \in \mathcal{S} : F_1(x) \geq u\}, \quad F_2^{-1}(u) := \inf\{x \in \mathcal{S} : F_2(x) \geq u\}$$

are the corresponding quantile functions, defined as left-continuous inverses of the cdf. A map T that satisfies $T_{\#}\mu_1 = \mu_2$ is a transport map from μ_1 to μ_2 and $T_{12} = F_2^{-1} \circ F_1$ is referred to as the optimal transport map that pushes the probability measure μ_1 forward to the measure μ_2.

For a nonempty interval $I \subset \mathbb{R}$, the length of a given curve $\gamma : I \to \mathcal{W}$ is $L(\gamma) := \sup \sum_{i=1}^{k} d_{\mathcal{W}}(\gamma(t_{i-1}), \gamma(t_i))$, where the supremum is taken over all $k \in \mathbb{N}$ and $t_0 \leq t_1 \leq \cdots \leq t_k$ in I. For absolutely continuous μ_1, McCann's interpolant (McCann, 1997) is the curve $\gamma : [0, 1] \to \mathcal{W}$ given by

$$\gamma(a) = (id + a(T_{12} - id))_{\#}\mu_1,$$

where $a \in [0, 1]$ and id is the identity map. McCann's interpolant is the geodesic in $\mathcal{W}$ that corresponds to the optimal transport from μ_1 to μ_2, where we do not distinguish between this geodesic and the transport map T_{12}; we note that $L(\gamma) = d_{\mathcal{W}}(\mu_1, \mu_2)$ and γ has constant speed $d_{\mathcal{W}}(\gamma(a_1), \gamma(a_2)) = (a_2 - a_1)d_{\mathcal{W}}(\mu_1, \mu_2)$ for any $0 \leq a_1 \leq a_2 \leq 1$.

Our focus is on a time series of distributions $\{\mu_i\}_{i=1,2,\ldots,n} \subset \mathcal{W}$, which is assumed to possess some stationarity properties, including stationarity of the mean. This means that there exists a common Fréchet mean or barycenter $\mu_{\mathcal{F}}$, given by

$$\mu_{\mathcal{F}} := \underset{\nu \in \mathcal{W}}{\operatorname{argmin}}\ \mathbb{E} d_{\mathcal{W}}^2(\nu, \mu_i) \quad \text{for all } i = 1, 2, \ldots, n,$$

where existence and uniqueness are assured by the fact that the Wasserstein space for one-dimensional distributions is a Hadamard space (Kloeckner, 2010).

We now consider the space of all Lebesgue integrable functions on $\mathcal{S}$, $L^p(\mathcal{S}) = \{f : \mathcal{S} \to \mathbb{R} \mid \|f\|_{\mathcal{L}^p} < \infty\}$, where $1 \leq p < \infty$, λ is the Lebesgue measure and $\|f\|_{\mathcal{L}^p} := (\int_{\mathcal{S}} |f|^p \,\mathrm{d}\lambda)^{1/p}$ is the usual L^p-norm. Define the set $\mathcal{T}$ as

$$\mathcal{T} = \{T : \mathcal{S} \to \mathcal{S} \mid T(s_1) = s_1,\ T(s_2) = s_2,\ T \text{ is nondecreasing}\}. \tag{2}$$

Since $\mathcal{T}$ is a closed subset of $L^p(\mathcal{S})$, it is a complete metric space with respect to the L^p-norm, i.e., the limit of every Cauchy sequence of points in $\mathcal{T}$ is still in $\mathcal{T}$. In addition, $\mathcal{T} \subset L^p(\mathcal{S})$ can be equivalently identified as $\mathcal{T} = \{T : \mathcal{S} \to \mathcal{S} \mid T := F_1^{-1} \circ F_2\}$, where, as above, F_1, F_2 are the cdfs of probability measures $\mu_1, \mu_2 \in \mathcal{W}$. Here, F_1, F_2 may not be continuous and are not necessarily strictly increasing. For any $T \in \mathcal{T}$, the representation $T = F_1^{-1} \circ F_2$ is not unique and one may choose F_2 to be the cdf of a uniform distribution, in which case T is represented by F_1^{-1} only, which then is unique. This not only induces a metric on $\mathcal{T}$ but also shows that $\mathcal{W}$ and $\mathcal{T}$ are isometric with this induced metric. This isometry induces a probability measure on $\mathcal{T}$ that is inherited from the corresponding measure on $\mathcal{W}$. Furthermore, for every $T \in \mathcal{T}$, there exists a uniquely defined inverse transport map $T^{-1} \in \mathcal{T}$; for any given representation $T = F_1^{-1} \circ F_2$, $T^{-1} = F_2^{-1} \circ F_1$.

To build an autoregressive model for elements in $\mathcal{T}$, we introduce addition and scalar multiplication operations in the transport space $\mathcal{T}$ as follows.

- Addition: $T_1 \oplus T_2 := T_2 \circ T_1$, where $T_1, T_2 \in \mathcal{T}$.
- Scalar multiplication: For any $x \in \mathcal{S}$ and $T \in \mathcal{T}$, for any $\alpha \in \mathbb{R}$ with $-1 \leq \alpha \leq 1$, let

$$\alpha \odot T(x) := \begin{cases} x + \alpha(T(x) - x), & 0 < \alpha \leq 1 \\ x, & \alpha = 0 \\ x + \alpha(x - T^{-1}(x)), & -1 \leq \alpha < 0. \end{cases}$$

For any $|\alpha| > 1$, let $b = \lfloor |\alpha| \rfloor$, the integer part of α, and set $a = |\alpha| - b$. We then define a scalar multiplication in transport space by

$$\alpha \odot T(x) := \begin{cases} (a \odot T) \circ \underbrace{T \circ T \circ \ldots \circ T}_{b \text{ compositionsof } T}(x), & \alpha > 1 \\ (a \odot T^{-1}) \circ \underbrace{T^{-1} \circ T^{-1} \circ \ldots \circ T^{-1}}_{b \text{ compositionsof } T^{-1}}(x), & \alpha < -1, \end{cases}$$

These operations are motivated as follows. Addition of transports is defined as their simple concatenation, which is a straightforward extension from the case of transports in $\mathbb{R}^p$, where transports correspond to vectors V that are added to a vector argument c, so that $T_V(c) = V + c$. Consecutively applying two transport maps T_{V_1} and then T_{V_2} then means adding the sum of the two vectors $V_1 + V_2$ to the argument vector c, so that $T_{V_2} \circ T_{V_1}(c) = V_1 + V_2 + c$. For scalar multiplication, given $0 < \alpha < 1$, a transport vector αV defines the transport $T_{\alpha V}(c) = \alpha V + c$ and therefore transports an argument vector c to a point on the straight line (geodesic) between c and $c + V$. So if T_1 is the optimal transport that pushes $\mu_{\mathcal{F}}$ to $(T_1)_{\#}\mu_{\mathcal{F}}$, it is natural to define $\alpha \odot T_1$ such that it pushes $\mu_{\mathcal{F}}$ to a distribution lying on the geodesic from $\mu_{\mathcal{F}}$ to $(T_1)_{\#}\mu_{\mathcal{F}}$ where its location on the geodesic is characterized by a fraction of length α when measuring length from the starting point $\mu_{\mathcal{F}}$. When α is negative, $c + \alpha V = c + |\alpha|(- V)$, where $-V$ can be interpreted as the transport map that pushes $c + V$ to c and thus is the inverse transport T_V^{-1} of the transport T_V. The obvious extension to optimal transport maps in distribution spaces then leads to the above definition of scalar multiplication, which is further illustrated in Figure 1. A distinction from the vector space case is that the addition $\oplus$ for optimal transport maps is not commutative. For scalar multiplication with factors α that are such that $|\alpha| > 1$, if α is an integer we decompose the map T^α into an iterative sum of maps T, and if α is not an integer we apply the integer part of α first and after this apply an additional transport map that is a scalar multiplication of T with the left-over fractional part of α.

Observe that$(\mathcal{T},\oplus)$ is a (non-Abelian) group with the identity map as identity. For any $T \in \mathcal{T}$, the inverse is T^{-1}. By the definition of $\oplus$, we have

$$(T_1 \oplus T_2) \oplus T_3 = T_3 \circ (T_2 \circ T_1) = (T_3 \circ T_2) \circ T_1 = T_1 \oplus (T_2 \oplus T_3),$$

which entails the associativity of $\oplus$. Regarding the relation between $\odot$ and $\oplus$, distributive laws do not hold, i.e., there exists $\alpha, \beta \in \mathbb{R}$ and T_1, $T_2 \in \mathcal{T}$ such that

$$\alpha \odot (T_1 \oplus T_2) \neq (\alpha \odot T_1) \oplus (\alpha \odot T_2), \quad (\alpha + \beta) \odot T_1 \neq (\alpha \odot T_1) \oplus (\beta \odot T_1).$$

A simple example is as follows. Set $\mathcal{S} = [0, 1]$, $T_1(x) = x^2$, $T_2(x) = (x + x^2)/2$ and $\alpha = 0.6, \beta = 0.7$. Simple algebra shows that $(\alpha + \beta) \odot T_1 = 0.7x^2 + 0.3x^4 \neq 0.3(0.4x + 0.6x^2) + 0.7(0.4x + 0.6x^2)^2 = (\alpha \odot T_1) \oplus (\beta \odot T_1)$. In addition, the coefficient of x^4 in the 4th order polynomial (with respect to x) $\alpha \odot (T_1 \oplus T_2)$ is 0.3, while x^4 has coefficient 0.6^3 in $(\alpha \odot T_1) \oplus (\alpha \odot T_2)$, which indicates that $\alpha \odot (T_1 \oplus T_2) \neq (\alpha \odot T_1) \oplus (\alpha \odot T_2)$.

3 Autoregressive transport models of order 1

3.1 Model and stationary solution

We first consider a time series $\{X_i\}_{i=1,2,\ldots,n} \subset \mathbb{R}^p$ with constant mean $E[X_i] = \mu \in \mathbb{R}^p$. The vector autoregressive model of order 1 (VAR(1)) with scalar coefficient is

$$X_i - \mu = \beta(X_{i-1} - \mu) + \epsilon_i, \tag{3}$$

where $\beta \in \mathbb{R}$ and $\{\epsilon_i\}_{i=1,2,\ldots,n} \subset \mathbb{R}^p$ are the i.i.d innovations with mean 0. In this Euclidean time series model, the vector $X_i - \mu$ can be interpreted as the optimal transport map pushing μ to X_i, which provides the inspiration for the proposed ATM.

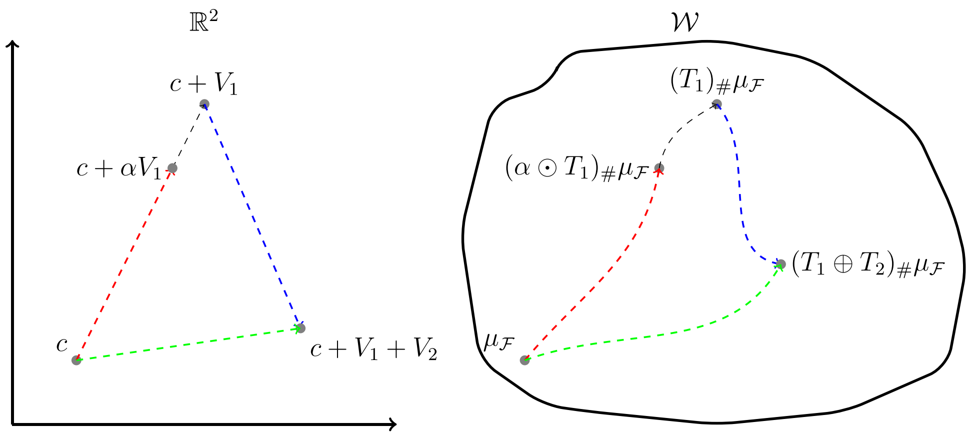

Figure 1. Motivating the definition of the addition $\oplus$ and scalar multiplication $\odot$ operations for $0 < \alpha < 1$ in the Wasserstein optimal transport space for transports T_1, T_2 (right), while in $\mathbb{R}^2$ optimal transports are defined by vectors V_1, V_2 (left).

In general metric spaces, differences cannot be formed and thus a direct extension of model (3) is not feasible. However, in transport spaces with uniquely defined optimal transports along geodesics we can reinterpret differences of elements in terms of such optimal transports. Specifically, in Wasserstein space, we define the difference between two distributions μ_2 and μ_1 to be the optimal transport map that pushes μ_1 to μ_2, i.e.,

$$\mu_2 \ominus \mu_1 = F_2^{-1} \circ F_1, \tag{4}$$

where in (4) $F_1 = F(\mu_1)$, $F_2 = F(\mu_2)$ are the cdfs of measures μ_1, μ_2, respectively. We also require appropriate generalizations for the random innovations ϵ_i that now become random transports. Extending the notion of additive noise for Euclidean data, we model noise in transport space as random transport maps in $\mathcal{T}$ constrained in such a way that their Fréchet mean (barycenter) is the identity transport. A noise contaminated version of a transport map $T \in \mathcal{T}$ is thus $T \oplus \epsilon$, where $E(\epsilon) = id$.

Motivated by model (3), the autoregressive transport model of order 1 (ATM(1)) is

$$T_i = \alpha \odot T_{i-1} \oplus \varepsilon_i, \quad \text{where } T_i = \mu_i \ominus \mu_{\mathcal{F}}, \tag{5}$$

where $\alpha \in \mathbb{R}$ is the model parameter and the ε_i are i.i.d random distortion transport maps with mean $E(\varepsilon_i) = id$. The proposed ATM approximates the optimal transport map at time $t = i$ with the scaled transport map $\alpha \odot T_{i-1}$, in analogy to the VAR(1) model $X_i - c = \beta(X_{i-1} - c) + \epsilon_i$, which can be interpreted as approximating the optimal transport map $X_i - c$ with the scaled transport map $\beta(X_{i-1} - c)$; see Figure 2 for an illustration. While (3) provides the usual formulation of the VAR(1) model, another way to view the model is by relating past differences to current differences, i.e., model (3) gives rise to the alternative model

$$X_i - X_{i-1} = \beta(X_{i-1} - X_{i-2}) + \epsilon_i. \tag{6}$$

The difference $X_i - X_{i-1}$ can be interpreted as the optimal transport map between X_{i-1} and X_i. In Wasserstein space, autoregressive transport models of order 1 (ATM(1)) can analogously be built with optimal transports between adjacent distributions,

$$T_i = \alpha \odot T_{i-1} \oplus \varepsilon_i, \quad \text{where } T_i = \mu_{i+1} \ominus \mu_i, \tag{7}$$

where the ε_i are again i.i.d random distortion maps with $E(\varepsilon_i) = id$.

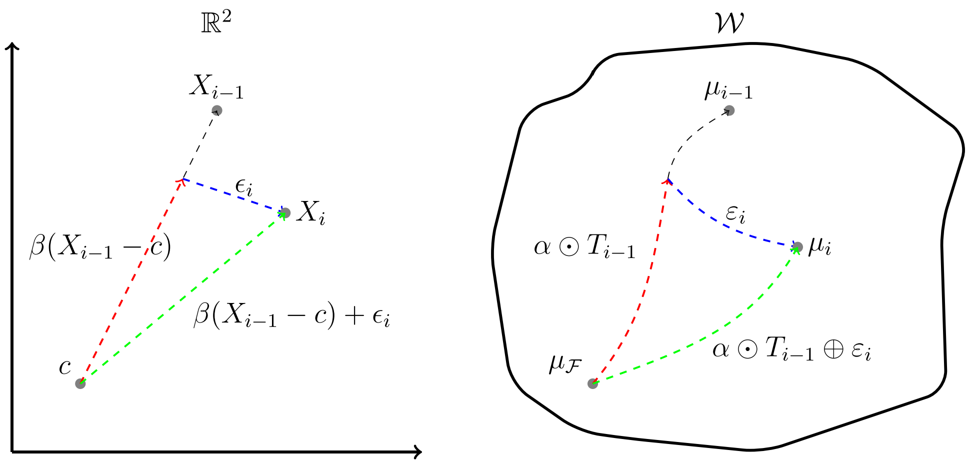

Figure 2. Illustration of the VAR(1) model $X_i - c = \beta(X_{i-1} - c) + \epsilon_i$ in $\mathbb{R}^2$ (left) and the ATM(1) model $T_i = \alpha \odot T_{i-1} \oplus \varepsilon_i$, $T_i = F_i^{-1} \circ F_{\mathcal{F}}$ in $\mathcal{W}$ (right). The coloured dashed lines are geodesics and correspond to the respective optimal transport maps.

Next we show the existence of stationary solutions for models (5) and (7). For any $S, T \in \mathcal{T}$, $1 \leq q < \infty$ and random distortion map ε, we utilize the distances $d_q(S, T) = \|S - T\|_{\mathcal{L}^q}$ on $\mathcal{T}$ and define $\phi_\varepsilon, \widetilde{\phi}_{i,m} : \mathcal{T} \to \mathcal{T}$ by

$$\phi_\varepsilon(S) = \alpha \odot S \oplus \varepsilon, \qquad \widetilde{\phi}_{i,m}(S) = \phi_{\varepsilon_i} \circ \phi_{\varepsilon_{i-1}} \circ \cdots \circ \phi_{\varepsilon_{i-m+1}}(S).$$

Then under a suitable contraction condition, stationary solutions exist.

Theorem 1 Suppose there exists $\eta > 0$, $S_0 \in \mathcal{T}$, $C > 0$ and $r \in (0, 1)$ such that

$$E\left[d_q^\eta\left(\widetilde{\phi}_{i,m}(S_0), \widetilde{\phi}_{i,m}(T)\right)\right] \leq C r^m d_q^\eta(S_0, T) \tag{8}$$

holds for a given $1 \leq q < \infty$ and all $m \in \mathbb{N}$ and all $T \in \mathcal{T}$. Then, for all $S \in \mathcal{T}$, $\widetilde{T}_i := \lim_{m\to\infty} \widetilde{\phi}_{i,m}(S) \in \mathcal{T}$ exists almost surely and does not depend on S. In addition, $\widetilde{T}_i$ is a stationary solution to the following system of stochastic transport equations

$$T_i = \alpha \odot T_{i-1} \oplus \varepsilon_i, \quad i \in \mathbb{Z} \tag{9}$$

and is unique almost surely.

The proof utilizes the theory of iterated random function systems (Diaconis & Freedman, 1999), where a crucial element is the geometric-moment contraction condition (8) of Wu and Shao (2004). Regarding sufficient conditions for (8) when $q = 1$, easy algebra shows that $d_1(\alpha \odot S, \alpha \odot T) = \alpha d_1(S, T)$ for a positive α. From the corresponding result on the L_1 distance of cdfs (see, e.g., Shorack & Wellner, 2009), one immediately finds

$$d_1(S, T) = \int_{\mathcal{S}} |S(x) - T(x)|\, dx = \int_{\mathcal{S}} |S^{-1}(x) - T^{-1}(x)|\, dx,$$

which then entails that $d_1(\alpha \odot S, \alpha \odot T) = -\alpha d_1(S, T)$ when $\alpha < 0$. Suppose for any $S, T \in \mathcal{T}$, $E[d_1(\varepsilon_i \circ S, \varepsilon_i \circ T)] \leq L d_1(S, T)$, where L is some positive constant such that $\alpha L \in (0, 1)$, then (8) is seen to hold with $\eta = 1$ and $r = \alpha L$ by iterating the argument. Moreover, $E[d_1(\varepsilon_i \circ S, \varepsilon_i \circ T)] \leq L d_1(S, T)$ holds if the $\{\varepsilon_i\}$ satisfy $E[|\varepsilon_i(x) - \varepsilon_i(y)|] \leq L|x - y|$.

3.2 Estimation

As mentioned before, ATM(1) is an extension of the classical AR(1) model in Euclidean space. A necessary and sufficient condition for AR(1) to admit a stationary solution is that the model parameter lies in $(-1, 1)$, and Theorem 1 together with the subsequent discussion indicates that a similar framework applies for ATM(1). Thus, it is natural to assume that the true model parameter α of ATM(1) lies in $(-1, 1)$. Furthermore, in distributional data analysis and distributional time series the distributions that serve as data atoms are usually not known but one rather has available i.i.d. samples of real-valued data that have been generated by these distributions and this needs to be taken into account in the analysis. In the following, we describe a consistent estimator for α based on these samples of real-valued data. We denote the available estimates of transport maps T_i by $\widehat{T}_i$, $i = 1, \cdots, n$. Depending on whether α is positive or negative, $\widehat{T}_i$ or $\widehat{T}_i^{-1}$ is used accordingly in the proposed method.

If $\{T_1, \cdots, T_n\}$ satisfies model (9), then it holds that

$$\alpha = \begin{cases} \dfrac{\int_{\mathcal{S}} E[(T_{i+1}(x) - x)(T_i(x) - x)]\,\mathrm{d}x}{\int_{\mathcal{S}} E[(T_i(x) - x)^2]\,\mathrm{d}x}, & \text{if } \alpha \geq 0, \\ \dfrac{\int_{\mathcal{S}} E[(T_{i+1}(x) - x)(x - T_i^{-1}(x))]\,\mathrm{d}x}{\int_{\mathcal{S}} E[(x - T_i^{-1}(x))^2]\,\mathrm{d}x}, & \text{if } \alpha < 0. \end{cases}$$

This motivates the following least squares type estimators of α,

$$\widehat{\alpha} = \begin{cases} \widehat{\alpha}_+ & \text{if } l_+(\widehat{\alpha}_+) \leq l_-(\widehat{\alpha}_-), \\ \widehat{\alpha}_- & \text{if } l_+(\widehat{\alpha}_+) > l_-(\widehat{\alpha}_-). \end{cases}$$

where $\widehat{\alpha}_+ = \operatorname{argmin}_\alpha l_+(\alpha)$, $\widehat{\alpha}_- = \operatorname{argmin}_\alpha l_-(\alpha)$ and

$$l_+(\alpha) = \sum_{i=2}^{n} \int_{\mathcal{S}} \left(\widehat{T}_i(x) - x - \alpha(\widehat{T}_{i-1}(x) - x)\right)^2 \mathrm{d}x,$$

$$l_-(\alpha) = \sum_{i=2}^{n} \int_{\mathcal{S}} \left(\widehat{T}_i(x) - x - \alpha(x - \widehat{T}_{i-1}^{-1}(x))\right)^2 \mathrm{d}x.$$

Theorem 2 Suppose $T_0 \overset{i.i.d}{\sim} \widetilde{T}_0$ and $\{T_i\}_{i=1}^n$ are strictly increasing, continuous and generated from equation (9) with $-1 < \alpha < 1$ and T_0 as the initial transport. Under the assumptions of Theorem 1 with $q = 1$, if $\int_{\mathcal{S}} E[(T_1(x) - x)^2]\,\mathrm{d}x > 0$,

$$|\widehat{\alpha} - \alpha| = O_p\left(\tau + \frac{1}{\sqrt{n}}\right),$$

where $\tau = \sup_i E[d_1(\widehat{T}_i, T_i)]$.

Intuitively, the condition $\int_{\mathcal{S}} E[(T_1(x) - x)^2]\,\mathrm{d}x > 0$ ensures that the sequence of transport maps deviates from a sequence of identity maps. This is required to arrive at a consistent estimator, since if $T_i = id$ almost surely, equation (9) would hold for any $\alpha \in \mathbb{R}$ and it is then not possible to estimate α consistently. More specifically, if $\int_{\mathcal{S}} E[(T_1(x) - x)^2]\,\mathrm{d}x > 0$, then the following application of the Cauchy-Schwarz inequality excludes the case of equality and therefore gives rise to the strict inequality

$$c' = \left(\int_{\mathcal{S}} E[(T_1(x) - x)^2]\,\mathrm{d}x\right)\left(\int_{\mathcal{S}} E[(x - T_1^{-1}(x))^2]\,\mathrm{d}x\right) - \left(\int_{\mathcal{S}} E[(T_1(x) - x)(x - T_1^{-1}(x))]\,\mathrm{d}x\right)^2 > 0,$$

where $1/c'$ is an implicit constant in the O_p for the rate of convergence result $O_p(\tau + 1/\sqrt{n})$.

For practical applications it needs to be taken into account that the underlying distributions are almost always unknown. Accordingly, a realistic starting point is that one has available samples of independent realizations $\{X_{i,l}\}_{l=1}^{N_i}$ that are obtained for each of the distributions μ_i. There are then two independent random mechanisms that generate the data. The first of these generates random distributions $\{\mu_i\}_{i=1}^{n}$; the second generates randomly drawn samples $\{X_{i,l}\}$ from each μ_i. Based on the $\{X_{i,l}\}_{l=1}^{N_i}$, cdfs $\{F_i\}$ or quantile functions $\{Q_i\}$ can be estimated with available methodology (Falk, 1983; Leblanc, 2012). Denoting the estimated cdfs by $\widehat{F}_i$, the corresponding quantile function estimates are $\widehat{Q}_i(a) = \inf\{x \in \mathcal{S} \mid \widehat{F}_i \geq a\}$, $a \in [0, 1]$. Alternatively, one can directly estimate quantile functions (Cheng & Parzen, 1997) or start with density estimates and convert these to cdfs using numerical integration, obtaining rates such as $\sup_{\mu \in \mathcal{W}} E[d_{\mathcal{W}}^2(\widehat{\mu}, \mu)] = O(1/\sqrt{N})$, where $N = \min\{N_i : i = 1, 2, \cdots, n\}$ (Panaretos & Zemel, 2016) under suitable assumptions or alternatively $\sup_{\mu \in \mathcal{W}_R^{ac}} E[d_{\mathcal{W}}^2(\widehat{\mu}, \mu)] = O(N^{-2/3})$ on the set of absolutely continuous distributions (Petersen & Müller, 2016). With estimates for quantile functions and cdf in hand, one then obtains optimal transport map estimates $\widehat{T}_i = \widehat{Q}_i \circ \widehat{F}_{\mathcal{F}}$ or $\widehat{T}_i = \widehat{Q}_{i+1} \circ \widehat{F}_i$, where $\widehat{F}_{\mathcal{F}} = \widehat{Q}_{\mathcal{F}}^{-1}$ and $\widehat{Q}_{\mathcal{F}} = \sum_{i=1}^{n} \widehat{Q}_i/n$, implying

$$\tau \lesssim \sup_i E[d_1(\widehat{T}_i, T_i)] \lesssim \max\left\{ \sup_i (E[d_{\mathcal{W}}^2(\widehat{\mu}_i, \mu_i)])^{1/2}, (E[d_{\mathcal{W}}^2(\widehat{\mu}_{\mathcal{F}}, \mu_{\mathcal{F}})])^{1/2} \right\}$$

for the rate τ in Theorem 2, where $a \lesssim b$ means that there exists a constant $C > 0$ such that $a \leq Cb$. Depending on assumptions and estimation procedures as mentioned above, one then obtains convergence rates ranging from $\tau \sim N^{-1/4}$ to $\tau \sim N^{-1/3}$.

4 Autoregressive transport models of order p

4.1 Stationary solution

Autoregressive transport models of order p (ATM(p)) are defined as

$$T_i = \alpha_p \odot T_{i-p} \oplus \alpha_{p-1} \odot T_{i-p+1} \oplus \cdots \oplus \alpha_1 \odot T_{i-1} \oplus \varepsilon_i, \tag{10}$$

where $\alpha_1, \cdots, \alpha_p \in \mathbb{R}$ are model parameters and ε_i are i.i.d. random distortion maps with $E(\varepsilon_i) = id$. To show the existence of stationary solutions, we construct a chain of functions and again apply the geometric-moment contraction condition (Wu & Shao, 2004). Let $\mathcal{T}^p = \mathcal{T} \times \cdots \times \mathcal{T}$ be the product space, $\mathbf{S} = (S_1, S_2, \cdots, S_p)$, $\mathbf{R} = (R_1, R_2, \cdots, R_p) \in \mathcal{T}^p$ and define the random functions $\Upsilon_\varepsilon, \widetilde{\Upsilon}_{i,m} : \mathcal{T}^p \to \mathcal{T}^p$ as

$$\Upsilon_\varepsilon(\mathbf{S}) = (S_2, \cdots, S_p, \alpha_p \odot S_1 \oplus \cdots \oplus \alpha_1 \odot S_p \oplus \varepsilon),$$
$$\widetilde{\Upsilon}_{i,m}(\mathbf{S}) = \Upsilon_{\varepsilon_i} \circ \Upsilon_{\varepsilon_{i-1}} \circ \cdots \circ \Upsilon_{\varepsilon_{i-m+1}}(\mathbf{S}),$$

where $\varepsilon, \varepsilon_i$ are random distortion transports. We employ the product L^q-metric on $\mathcal{T}^p$ given by $d_q(\mathbf{S}, \mathbf{R}) = \left\{ \sum_{i=1}^{p} d_q^2(S_i, R_i) \right\}^{1/2}$, where $q \geq 1$ is a fixed constant in the following.

Theorem 3 Suppose there exists $\eta > 0$, $\mathbf{S}_0 \in \mathcal{T}^p$, $C > 0$ and $r \in (0, 1)$ such that

$$E\left[d_q^\eta\left(\widetilde{\Upsilon}_{i,m}(\mathbf{S}_0), \widetilde{\Upsilon}_{i,m}(\mathbf{R}) \right) \right] \leq C r^m d_q^\eta(\mathbf{S}_0, \mathbf{R}) \tag{11}$$

holds for all $\mathbf{R} \in \mathcal{T}^p$ and $m \in \mathbb{N}$. Then, for all $\mathbf{S} \in \mathcal{T}^p$,

$$(\widetilde{T}_{i-p+1}, \widetilde{T}_{i-p+2}, \cdots, \widetilde{T}_i) := \lim_{m \to \infty} \widetilde{\Upsilon}_{i,m}(\mathbf{S}) \in \mathcal{T}^p$$

exists almost surely and does not depend on $\mathbf{S}$. In addition, $(\widetilde{T}_{i-p+1}, \widetilde{T}_{i-p+2}, \cdots, \widetilde{T}_i)$ is a stationary solution of the following system of stochastic equations

$$T_i = \alpha_p \odot T_{i-p} \oplus \alpha_{p-1} \odot T_{i-p+1} \oplus \cdots \oplus \alpha_1 \odot T_{i-1} \oplus \varepsilon_i, \quad i \in \mathbb{Z}$$

and is unique almost surely.

For motivation of Υ_ε and condition (11), consider the classical AR(p) model in $\mathbb{R}$, i.e., $Y_i = \sum_{j=1}^p \beta_j Y_{i-j} + \epsilon_i \in \mathbb{R}$, which can be represented as a vector autoregressive model of order 1 (VAR(1)) in the form $\mathbf{Y}_i = B\mathbf{Y}_{i-1} + \boldsymbol{\epsilon}_i$, where $\mathbf{Y}_i = (Y_i, \cdots, Y_{i-p+1})^T$, $\boldsymbol{\epsilon}_i = (\epsilon_i, 0, \cdots, 0)^T \in \mathbb{R}^p$ and

$$B = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_{p-1} & \beta_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

With (nonrandom) starting points $\mathbf{Y}_0$ and $\mathbf{Y}'_0$, running the VAR(1) model recursively m times, one obtains $\mathbf{Y}_m = B^m\mathbf{Y}_0 + \sum_{j=1}^m B^{m-j}\boldsymbol{\epsilon}_j$ and $\mathbf{Y}'_m = B^m\mathbf{Y}'_0 + \sum_{j=1}^m B^{m-j}\boldsymbol{\epsilon}_j$. With $\|\cdot\|_2$ denoting the Euclidean norm, condition (11) for this model becomes $E[\|\mathbf{Y}_m - \mathbf{Y}'_m\|_2] \lesssim r^m \|\mathbf{Y}_0 - \mathbf{Y}'_0\|_2$ for some $0 < r < 1$. With a slight abuse of notation, denoting the spectral norm of B as $\|B\|_2$,

$$E\big[\|\mathbf{Y}_m - \mathbf{Y}'_m\|_2\big] = \|B^m(\mathbf{Y}_0 - \mathbf{Y}'_0)\|_2^\eta \leq \|B^m\|_2 \|\mathbf{Y}_0 - \mathbf{Y}'_0\|_2.$$

Now if the absolute values of the eigenvalues of B are bounded above by a constant $0 < r < 1$, i.e., they are inside the unit circle, then $\|B^m\|_2 \lesssim r^m$, and this is equivalent to the fact that the roots of $\phi(z) = 1 - \sum_{j=1}^p \beta_j z^j$ all lie outside the unit circle. The latter is a standard assumption for the existence of stationary solutions of AR(p) processes in Euclidean space. In linear spaces the terms containing the innovation errors in $\mathbf{Y}_m$ and $\mathbf{Y}'_m$ cancel, which for this case simplifies the verification of Condition (11).

To select the order of the ATM, we propose an approach based on rolling-window validation and refer to Zivot and Wang (2007) for more details on rolling-window analysis for time series. To train the ATM(p) on a given sequence $\{\mu_t, \mu_{t+1}, \cdots, \mu_{t+m-1}\}$ of length m with starting time t, we assume that there exists a pre-sample of length k, i.e., $\{\mu_{t-k}, \cdots, \mu_{t-1}\}$. For each fixed p in a candidate set, the sample $\{\mu_{t-k}, \mu_{t-k+1}, \cdots, \mu_{t-k+m-1}\}$ is used as training set to predict the distribution at time $t - k + m$. Denoting this predicted distribution as $\widehat{\mu}_{t-k+m}$, the prediction accuracy can be measured by Wasserstein distance $d_{\mathcal{W}}(\mu_{t-k+m}, \widehat{\mu}_{t-k+m})$. Then roll the window one step forward and use $\{\mu_{t-k+1}, \mu_{t-k+1}, \cdots, \mu_{t-k+m}\}$ as training set to make a prediction at time $t - k + m + 1$ and compute the error $d_{\mathcal{W}}(\mu_{t-k+m+1}, \widehat{\mu}_{t-k+m+1})$. Rolling the training window forward repeatedly until the last window covering time $t-1$ to $t+m-2$ is reached and computing the error $d_{\mathcal{W}}(\mu_{t+m-1}, \widehat{\mu}_{t+m-1})$ then leads to the selection of the autoregressive order p as the minimizer of $\sum_{i=t+m-k}^{t+m-1} d_{\mathcal{W}}(\mu_i, \widehat{\mu}_i)$ over a candidate set of orders.

4.2 Estimation of model parameters

Hereafter, we denote the true model parameters as $(\alpha_1^*, \cdots, \alpha_p^*)$ to avoid confusion. Obvious estimates of the ATM(p) parameters $\alpha_1^*, \alpha_2^*, \cdots, \alpha_p^*$ are obtained as minimizers of

$$L_n(\alpha_1, \alpha_2, \cdots, \alpha_p) = \frac{1}{n-p} \sum_{i=p+1}^{n} \int_{\mathcal{S}} \left(T_i(x) - \alpha_p \odot T_{i-p} \oplus \cdots \oplus \alpha_1 \odot T_{i-1}(x)\right)^2 \mathrm{d}x.$$

When $p > 1$, the minimization of $L_n(\alpha_1, \cdots, \alpha_p)$ is challenging, as the functional L_n in general is not convex. We propose a back propagation-type algorithm to address this minimization problem.

The partial derivatives of $\alpha \odot T_i(x)$ with respect to x are

$$\frac{\partial}{\partial x}\alpha \odot T_i(x) = \begin{cases} (1 + a(g_b(x, T_i) - 1)) \times \left(\prod_{l=0}^{b-1} g_l(x, T_i)\right), & \text{if } \alpha > 0, \\ 1, & \text{if } \alpha = 0, \\ (1 + a(1 - g_b(x, T_i^{-1}))) \times \left(\prod_{l=0}^{b-1} g_l(x, T_i^{-1})\right), & \text{if } \alpha < 0, \end{cases}$$

where $b = \lfloor |\alpha| \rfloor$, $a = |\alpha| - b$, T', (T^{-1}) are the derivatives of T, T^{-1}, respectively, $\prod_{l=0}^{b-1} g_l(x, T)$ is defined to be 1 if $b - 1 < 0$ and

$$g_l(x, T) = \begin{cases} T'(x), & \text{if } l = 0, \\ T'(\underbrace{T \circ T \circ \cdots \circ T}_{l \text{ compositions of } T}(x)) & \text{if } l = 1, 2, \cdots \end{cases}$$

The partial derivative with respect to α when $a > 0$ is

$$\frac{\partial}{\partial \alpha}\alpha \odot T_i(x) = \begin{cases} T_i(h(x, T_i)) - h(x, T_i), & \text{if } \alpha > 0 \\ h(x, T_i^{-1}) - T_i^{-1}(h(x, T_i^{-1})), & \text{if } \alpha < 0, \end{cases}$$

where

$$h(x, T) = \begin{cases} x & \text{if } b = 0, \\ \underbrace{T \circ T \circ \cdots \circ T}_{b \text{ compositions of } T}(x) & \text{if } b > 0. \end{cases}$$

Since $\alpha \odot T_i(x)$ is not differentiable w.r.t α if $\alpha \in \mathbb{Z}$, we use its subdifferential (subgradient). When $\alpha = 0$, we set $\partial\alpha \odot T_i(x)/\partial\alpha$ at $\alpha = 0$ to be any value in the closed interval between $T_i(x) - x$ and $x - T_i^{-1}(x)$. In our simulations, $\partial\alpha \odot T_i(x)/\partial\alpha$ at $\alpha = 0$ is selected uniformly from $T_i(x) - x$ and $x - T_i^{-1}(x)$. When $0 \neq \alpha \in \mathbb{Z}$, $\partial\alpha \odot T_i(x)/\partial\alpha$ is set to be the partial derivative of $\alpha \odot T_i(x)$ at a point α such that α has the same sign as α and $|\alpha| < |\alpha| < (|\alpha| + 1)$. For more details on the back-propagation type algorithm for ATM of order p see the display for Algorithm 1. We employ gradient clipping, a common technique used in deep neural networks to prevent exploding gradients.

Next, we establish consistency for the minimizer of $L_n(\alpha_1, \cdots, \alpha_p)$, i.e.,

$$\tilde{\boldsymbol{\alpha}} := (\tilde{\alpha}_1, \tilde{\alpha}_2, \cdots, \tilde{\alpha}_p)^T \in \underset{-c \le \alpha_1, \cdots, \alpha_p \le c}{\operatorname{argmin}} L_n(\alpha_1, \alpha_2, \cdots, \alpha_p),$$

where c is the same constant as in Theorem 4 below, which demonstrates that $(\tilde{\alpha}_1, \tilde{\alpha}_2, \cdots, \tilde{\alpha}_p)$ converges to the true model parameters in probability with respect to the discrepancy

$$\Delta(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\alpha}^*) := \int_{\mathcal{S}} E\left[(\tilde{\alpha}_p \odot T_1 \oplus \cdots \oplus \tilde{\alpha}_1 \odot T_p(x) - \alpha_p^* \odot T_1 \oplus \cdots \oplus \alpha_1^* \odot T_p(x))^2\right] \mathrm{d}x,$$

where $\boldsymbol{\alpha}^* = (\alpha_1^*, \cdots, \alpha_p^*)^T$ are the true model parameters. The key step, where the constant c is used, is to show that $\sup_{-c \le \alpha_1, \cdots, \alpha_p \le c} |L_n(\alpha_1, \cdots, \alpha_p) - E[L_n(\alpha_1, \cdots, \alpha_p)]| = o_p(1)$ based on Corollary 3.1 of Newey (1991). In practice, we simply set c to be a large enough number.

Theorem 4 Under the assumptions of Theorem 3 with $q = 1$, if $T_0 \overset{i.i.d}{\sim} \tilde{T}_0$ and $\{T_i\}_{i=1}^n$ are strictly increasing, differentiable, bi-Lipschitz continuous with Lipschitz constant K and generated from equation (10) with T_0 as the initial transport and $(\alpha_1, \cdots, \alpha_p) = (\alpha_1^*, \cdots, \alpha_p^*)$ where $-c \le \alpha_1^*, \cdots, \alpha_p^* \le c$ for some constant $c > 0$, then

$$\Delta(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\alpha}^*) \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Algorithm 1 Back Propagation Algorithm for Fitting ATM(p), $p > 1$.

Select a grid $s_1 < x_1 < x_2 < \cdots < x_m < s_2$.
Select step size η.
Initialize $\alpha_k^0 = 0$ for $k = 2, 3, \ldots, p$ and

$$\alpha_1^0 = \underset{\alpha}{\operatorname{argmin}} \frac{1}{n-p} \sum_{i=p+1}^{n} \sum_{j=1}^{m} \big(T_i(x_j) - \alpha \odot T_{i-1}(x_j)\big)^2.$$

for $t = 1, 2, \ldots$ **do**

Forward Pass

For all $i = p+1, \ldots, n$, $j = 1, \ldots, m$, compute $R_{1,ji}^t = \alpha_p^{t-1} \odot T_{i-p}(x_j)$.
for $k = 2, 3, \ldots, p$ **do**
For all j, i, compute

$$R_{k,ji}^t = \alpha_{p+1-k}^{t-1} \odot T_{i-(p+1-k)}(R_{k-1,ji}^t).$$

For all j, i, compute $L_{ji}^t = 2\Big(T_i(x_j) - R_{p,ji}^t\Big)$.

Backward Pass

For all j, i, set $D_{0,ji}^t = 1$.
for $k = 1, 2, \ldots, p-1$ **do**
For all j, i, compute

$$D_{k,ji}^t = (D_{k-1,ji}^t) \times \left(\frac{\partial}{\partial x} \alpha_k^{t-1} \odot T_{i-k}(x) \Big|_{x = R_{p-k,ji}^t} \right) \text{ for all } j, i,$$

Update α_k as

$$\alpha_k^t = \alpha_k^{t-1} + \frac{\eta}{n-p} \sum_{i=p+1}^{n} \sum_{j=1}^{m} \left(L_{ji}^t D_{k-1,ji}^t \frac{\partial}{\partial \alpha} \alpha \odot T_{i-k}(R_{p-k,ji}^t) \Big|_{\alpha = \alpha_k^{t-1}} \right).$$

Compute $\alpha_p^t = \alpha_p^{t-1} + \frac{\eta}{n-p} \sum_{i=p+1}^{n} \sum_{j=1}^{m} \Big(L_{ji}^t (D_{p-1,ji}^t) \frac{\partial}{\partial \alpha} \alpha \odot T_{i-p}(x_j) \big|_{\alpha = \alpha_p^{t-1}} \Big)$.
if *stopping conditions hold* **then**
return $(\alpha_1^t, \alpha_2^t, \ldots, \alpha_p^t)$

5 Concurrent autoregressive transport model

A promising extension of ATMs of order 1 is to consider model coefficients that vary with $x \in \mathcal{S}$. For a function $\beta : \mathcal{S} \to [-1, 1]$, define the operation

$$\beta \circledcirc T(x) := \begin{cases} x + \beta(x)(T(x) - x), & 0 < \beta(x) \le 1 \\ x, & \beta(x) = 0 \\ x + \beta(x)(x - T^{-1}(x)), & -1 \le \beta(x) < 0. \end{cases}$$

This leads to the following concurrent autoregressive transport model (CAT),

$$T_i = \beta \circledcirc T_{i-1} \oplus \varepsilon_i, \tag{12}$$

with i.i.d. random distortion transports ε_i satisfying $E(\varepsilon_i) = id$.

To ensure monotonicity that is required for the transports to be well defined, given the true function β, we consider a subset of transports $\tilde{\mathcal{T}} \subset \mathcal{T}$ such that $\beta \circledcirc \tilde{\mathcal{T}} := \{\beta \circledcirc T : T \in \tilde{\mathcal{T}}\} \subseteq \tilde{\mathcal{T}}$ and assume that $P(\varepsilon_i \circ \tilde{\mathcal{T}} \subseteq \tilde{\mathcal{T}}) = 1$ where $\varepsilon_i \circ \tilde{\mathcal{T}} := \{\varepsilon_i \circ T : T \in \tilde{\mathcal{T}}\}$; this obviously holds if the function β does not vary, i.e., is constant, whence $\tilde{\mathcal{T}} = \mathcal{T}$ and $P(\varepsilon_i \circ \tilde{\mathcal{T}} \subseteq \tilde{\mathcal{T}}) = 1$. Whenever $\beta \circledcirc \tilde{\mathcal{T}} \subseteq \tilde{\mathcal{T}}$ and $P(\varepsilon_i \circ \tilde{\mathcal{T}} \subseteq \tilde{\mathcal{T}}) = 1$, the random functions

$$\varphi_\varepsilon(S) = \beta \circledcirc S \oplus \varepsilon, \qquad \widetilde{\varphi}_{i,m}(S) = \varphi_{\varepsilon_i} \circ \varphi_{\varepsilon_{i-1}} \circ \cdots \circ \varphi_{\varepsilon_{i-m+1}}(S), \qquad \varphi_\varepsilon, \widetilde{\varphi}_{i,m} : \tilde{\mathcal{T}} \to \tilde{\mathcal{T}}$$

are well-defined for any $S \in \tilde{\mathcal{T}}$.

An example for this concurrent autoregressive transport model (CAT) is as follows. Let $s_1 = t_1 < \cdots < t_k = s_2$ be a grid over $\mathcal{S}$ and $\beta : \mathcal{S} \to [0, 1]$ be such that β is positive and is either increasing or decreasing on each grid interval $[t_i, t_{i+1}]$. Here $\tilde{\mathcal{T}}$ is selected as a set of transports such that for any $T \in \tilde{\mathcal{T}}$, $T(t_i) = t_i$, $T(x) \geq x$ if $\beta(x)$ is increasing and otherwise $T(x) < x$. The properties required for the CAT model are satisfied as $\tilde{\mathcal{T}}$ is complete and $\{\varepsilon_i\}$ can be defined as random distortion maps taking values in $\tilde{\mathcal{T}}$. To state our next result, we equip $\tilde{\mathcal{T}}$ with the sup-metric $d_\infty(f, g) = \sup_{x \in \mathcal{S}} |f(x) - g(x)|$.

Theorem 5 Suppose that $\tilde{\mathcal{T}}$ is a complete metric space, $P(\varepsilon_i \circ \tilde{\mathcal{T}} \subseteq \tilde{\mathcal{T}}) = 1$ and there exists $\eta > 0$, $S_0 \in \tilde{\mathcal{T}}$, $C > 0$ and $r \in (0, 1)$ such that

$$E\left[d_\infty^\eta\left(\widetilde{\varphi}_{i,m}(S_0), \widetilde{\varphi}_{i,m}(T)\right)\right] \leq C r^m d_\infty^\eta(S_0, T) \tag{13}$$

holds for all $T \in \tilde{\mathcal{T}}$ and $m \in \mathbb{N}$. Then, for all $S \in \tilde{\mathcal{T}}$, $\widetilde{T}_i := \lim_{m\to\infty} \widetilde{\varphi}_{i,m}(S) \in \tilde{\mathcal{T}}$ exists almost surely and does not depend on S. In addition, $\widetilde{T}_i$ is a stationary solution of the system of stochastic equations

$$T_i = \beta \circledcirc T_{i-1} \oplus \varepsilon_i, \quad i \in \mathbb{Z} \tag{14}$$

and is unique almost surely.

The estimation of the CAT model function β proceeds similarly to the estimation of the scalar coefficient in the ATM(1). If $\{T_1, \cdots, T_n\}$ satisfy model (12), then for all $x \in \mathcal{S}$

$$\beta(x) = \begin{cases} \frac{E[(T_{i+1}(x)-x)(T_i(x)-x)]}{E[(T_i(x)-x)^2]}, & \text{if } \beta(x) \geq 0, \\ \frac{E[(T_{i+1}(x)-x)(x-T_i^{-1}(x))]}{E[(x-T_i^{-1}(x))^2]}, & \text{if } \beta(x) < 0. \end{cases}$$

This suggests estimates $\widehat{\beta}(x)$ for $\beta(x)$ given by

$$\widehat{\beta}(x) = \begin{cases} \widehat{\beta}_+(x), & \text{if } l_+(\widehat{\beta}_+(x)|x) \leq l_-(\widehat{\beta}_-(x)|x), \\ \widehat{\beta}_-(x), & \text{if } l_+(\widehat{\beta}_+(x)|x) > l_-(\widehat{\beta}_-(x)|x), \end{cases}$$

where $\widehat{\beta}_+(x) = \operatorname{argmin}_\beta l_+(\beta|x)$, $\widehat{\beta}_-(x) = \operatorname{argmin}_\beta l_-(\beta|x)$ and

$$l_+(\beta|x) = \sum_{i=2}^n \left(\widehat{T}_i(x) - x - \beta(\widehat{T}_{i-1}(x) - x)\right)^2,$$

$$l_-(\beta|x) = \sum_{i=2}^n \left(\widehat{T}_i(x) - x - \beta(x - \widehat{T}_{i-1}^{-1}(x))\right)^2.$$

Then we obtain pointwise convergence of $\widehat{\beta}(x)$ to $\beta(x)$ in probability.

Theorem 6 Suppose $T_0 \overset{i.i.d}{\sim} \widetilde{T}_0$ and $\{T_i\}_{i=1}^n$ are strictly increasing, continuous and generated from equation (14) with β such that $-1 < \beta(x) < 1$ for all $x \in \mathcal{S}$ and with T_0 as the initial transport. Under the assumptions of Theorem 5, if $E[(T_1(x) - x)^2] > 0$,

$$|\widehat{\beta}(x) - \beta(x)| = O_p\left(\tau(x) + \frac{1}{\sqrt{n}}\right),$$

where $\tau(x) = \sup_i E[|\widehat{T}_i(x) - T_i(x)|]$.

6 Numerical studies

In the following, ATM_m and CAT_m indicate models that are based on optimal transport maps $\{T_i\}$ from the Fréchet mean to individual distributions, while ATM_d and CAT_d indicate models based on optimal transport maps between adjacent distributions. Specifically, model (5) is denoted as $\mathrm{ATM}_m(p)$, model (7) as $\mathrm{ATM}_d(p)$, model (12) with $T_i = F_i^{-1} \circ F_{\mathcal{F}}$ as CAT_m and model (12) with $T_i = F_{i+1}^{-1} \circ F_i$ as CAT_d.

To examine the performance of these ATMs, we compare them in simulations with a recently proposed autoregressive model for distributional time series (Chen et al., 2022), which we refer to as WR (Wasserstein Regression). This approach is based on using manifold logarithmic maps in the Wasserstein manifold to map distributions to a tangent space anchored by the overall barycenter. Since the tangent space is a subspace of a L^2-space, functional linear regression techniques can be applied in this space, followed by a projection on the convex injectivity set and an application of the exponential map to get back to the Wasserstein manifold. Due to the local linearization this is an extrinsic approach, while the proposed ATMs are intrinsic to the Wasserstein manifold.

We also include comparisons with the log quantile (LQD) approach, which ignores the manifold structure of the distribution space, providing a direct 1:1 mapping of distributions to a Hilbert space by the invertible log quantile transformation or other transformations (Petersen & Müller, 2016). After applying the LQD transformation, standard autoregressive models for functional time series can be employed in the ensuing Hilbert space (Bosq, 2000), followed by mapping back into distribution space by the inverse LQD map. For autoregressive modelling of functional time series we used the R package 'ftsa'.

6.1 Interpretation of ATMs

We illustrate the process of transporting distributions by ATMs with a simple example. Let μ_1, μ_2, μ_3 be three normal distributions $N(-1, 2.25)$, $N(2.5, 1.44)$ and $N(-2, 0.81)$ (here all distributions are truncated to the interval $[-10, 10]$, where the miniscule mass left outside of the truncation interval is ignored). We apply $\mathrm{ATM}_m(2)$ and $\mathrm{ATM}_d(2)$ models to produce the distribution μ_4 at time $t = 4$. Figure 3 illustrates the densities of μ_2, μ_3 as well as the density of μ_4 generated by $\mathrm{ATM}_m(2)$, where the Fréchet mean is chosen as the standard normal distribution μ_0. For the optimal transport maps T_2, T_3 that map μ_0 to μ_2 and μ_3, respectively, we observe that T_2 shifts the density of μ_0 to the right and increases its variance, while T_3 shifts μ_0 to the left and decreases its variance.

Note that $\mathrm{ATM}_m(2)$ transforms μ_0 by first applying transport map $\alpha_1 \odot T_3$, followed by an application of transport map $\alpha_2 \odot T_2$. For example, from the first row of the figure, when $\alpha_1 = 0.5, \alpha_2 = 0.5$, the resulting overall transport is close to the identity, whereas for $\alpha_1 = 0.5, \alpha_2 = 0$, the resulting density of μ_4 is on the Wasserstein geodesic connecting μ_0 and μ_3. When $\alpha_1 = 0$, it can be seen from the middle row of Figure 3 that the density of μ_4 is moving to the left with decreasing variance when α_2 moves from 0.5 to -0.5. This illustrates the effect of the changing value of α_2 for the transport $\alpha_2 \odot T_2$. Similar effects can be seen in the third row of the figure.

The densities of μ_1, μ_2, μ_3 and the density of μ_4 that is obtained by applying the difference based models $\mathrm{ATM}_d(1)$ and $\mathrm{ATM}_d(2)$ are depicted in Figure 4. Denoting by T_1 the optimal transport map that maps μ_1 to μ_2 and by T_2 the transport map that maps μ_2 to μ_3, one finds that T_1 represents a shift to the right with a simultaneous decrease in variance, while T_2 represents a shift to the left, also accompanied by a decrease in variance. Applying ATM_d with T_1, T_2 as predictors, i.e., model (10) with $T_3 = \alpha_1 \odot T_2 \oplus \alpha_2 \odot T_1$, leads to the transport map T_3, which is then applied to μ_3, resulting in μ_4. To illustrate the effect of the coefficients, all panels show that decreasing α_2 enhances a shift to the left, while decreasing α_1 is associated with a shift to the right.

6.2 Reducing non-stationarity

The following example illustrates that the difference-based models ATM_d and CAT_d are advantageous compared to ATM_m, CAT_m and WR if the assumption that $\{\mu_1, \cdots, \mu_n\}$ is a stationary sequence does not hold. Stationarity of the sequence $\{\mu_1, \cdots, \mu_n\}$ is a basic assumption for models ATM_m, CAT_m and WR, whereas models ATM_d and CAT_d only require stationarity for differences, i.e., the sequence of optimal transport maps constructed by taking transports between consecutive distributions $\{\mu_1, \cdots, \mu_n\}$ as predictors.

Consider a sequence of Gaussian distributions $\{\mu_1, \cdots, \mu_6\}$ with mean 0 and decreasing standard deviations 4.8, 4, 3, 1.6, 1.15, 1, respectively. This sequence of distributions is non-stationary. We use $\{\mu_t : t = 1, 2, \cdots, 6\}$ as training data and aim to predict the distribution μ_7 with models $\text{ATM}_m(1)$, $\text{ATM}_d(1)$, CAT_m, CAT_d, WR and LQD. The densities of the training data are visualized in the left panel of Figure 5. One would expect μ_7 to follow this trend, i.e., to also have mean 0 with even smaller variance than μ_6. The right panel shows the predicted densities obtained with the different methods. We find that only ATM_d and CAT_d capture the underlying trend and provide reasonable predictions for the next element μ_7 in the sequence.

6.3 Simulations

We generate random transports according to

$$T_i = \alpha_4 \odot T_{i-4} \oplus \alpha_3 \odot T_{i-3} \oplus \alpha_2 \odot T_{i-2} \oplus \alpha_1 \odot T_{i-1} \oplus \varepsilon_i, \quad i \in \mathbb{Z}, \tag{15}$$

where $\varepsilon_i(x) = \frac{1}{2}((1+\xi_i)g(h^{-1}(x)) + (1+\xi_i)h^{-1}(x))$, $h(x) = \frac{1}{2}((1-\xi_i)g(x) + (1+\xi_i)x)$, $x \in \mathcal{S} = [0, 1]$ and $\{\xi_i\} \overset{i.i.d}{\sim} \text{Uniform}(-1, 1)$. Here $g(x)$ is the natural cubic spline passing through points (0, 0), (0.33, 0.7), (0.66, 0.8), (1, 1). We note that this construction ensures that the ε_i are transports. When representing these transports as quantile functions, for $0 < \xi_i < 1$ the function $g(x)$ is shifted along the direction perpendicular to the diagonal towards the identity map and for $-1 < \xi_i < 0$ this shift is applied to g^{-1} instead; see Figure 6 for an illustration of g and $\varepsilon_i(x)$. By construction, $E(\varepsilon_i) = id$.

To compare prediction accuracy across different models, we generated $\{T_i\}_{i=1}^{101}$ from the above model, using $\{T_i\}_{i=1}^{100}$ as training set, aiming to predict T_{101}. The Wasserstein distance between T_{101} and its prediction was computed for different combinations of α_1, α_2 by treating the transport maps $\{T_i\}$ as quantile functions. For these comparisons, we modified LQD to operate on transport maps, rather than predictor distributions (as originally devised). The simulation results for 1000 Monte Carlo replications are in Table 1 (numbers multiplied by 100). The order of ATM_m was obtained by rolling-window validation based on a pre-sample of size 50. When $\alpha_2 = 0$, model (15) reduces to an autoregressive model of order 1. Overall, ATM was found to outperform WR and LQD. We also use this example with $h(x)$ chosen as natural cubic spline passing through (0, 0), (0.3, 0.5), (0.6, 0.8), (1, 1) to illustrate the empirical rate of convergence of the estimates for the parameters of $\text{ATM}_m(1)$. Figure 7 displays estimation error versus $\sqrt{n}$ based on 200 Monte Carlo repetitions, demonstrating that finite sample performance with increasing sample sizes matches the root-n convergence rate predicted by Theorem 2.

It is also of interest to consider a sequence of distributions that are not generated from any of the examined models. Starting with the sequence of square integrable functions

$$R_i(x) = \sin(\zeta_i x), \tag{16}$$

where $x \in [0, 1]$ and the $\{\zeta_i\}$ are generated from the AR(2) model, $\zeta_i = \alpha_1\zeta_{i-1} + \alpha_2\zeta_{i-1} + \alpha_3\zeta_{i-3} + \alpha_4\zeta_{i-4} + \epsilon_i$, $\epsilon_i i.i.d. \sim \text{Uniform}(-4\pi, 4\pi)$, we convert the $\{R_i\}$ to

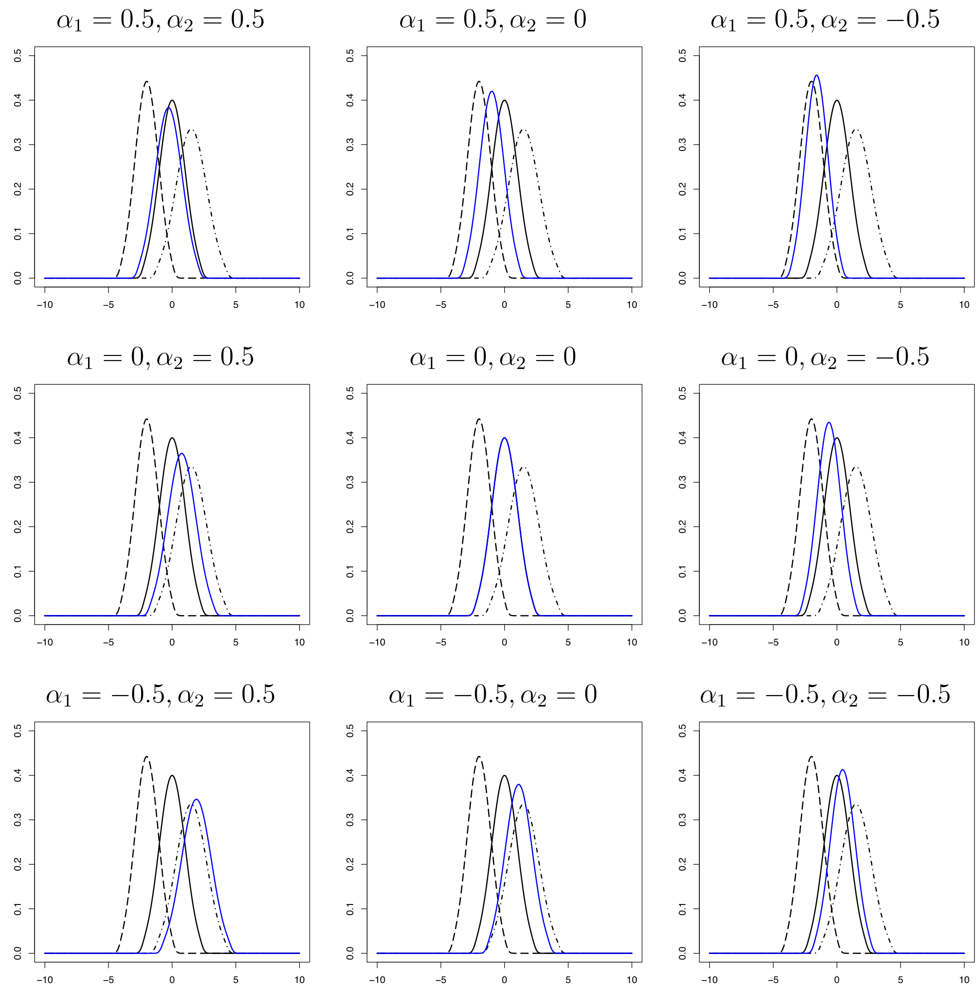

Figure 3. Illustrating $\text{ATM}_m(2)$ and $\text{ATM}_m(1)$ models for distributional time series. Each panel depicts the density functions for distributions μ_2 (dot-dashed), μ_3 (dashed) (these are the same across all panels) and the density of distribution μ_4 generated by ATM_m (solid), which varies across panels. For all panels the density of μ_0 (standard normal) is also included (solid black and always centered around the origin).

distributions by applying the inverse log quantile density transformation (Petersen & Müller, 2016), scaling the resulting distributions to be supported on [0, 1]; see Figure 6 (right panel) for an illustration. Again, we generate 100 distributions for training and report the results for 1,000 Monte Carlo replications. The simulation results are in the lower part of Table 1. For this case, we find that $\text{ATM}_m(1)$ is the overall preferred model.

6.4 Temperature data

One consequence of global warming may be an increasing frequency of warm summer nights in the Northern hemisphere. Inspired by the article of Bhatia and Katz (2021), we studied this with temperature data that were recorded at O'Hare international airport (available at https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND). The annual distributions of daily minimum temperatures, aggregating these temperatures over the period June 1 to September 30 over the summer months of each year, are illustrated in Figure 8 for the years from 1990 to 2019, where we use the distributions prior to 2019 as training data to predict the distribution for the year 2019.

For the ATM models we varied p from 1 to 3 and found that $p = 3$ yielded the best prediction. The observed and predicted densities for 2019 are shown in Figure 8. The Wasserstein distances

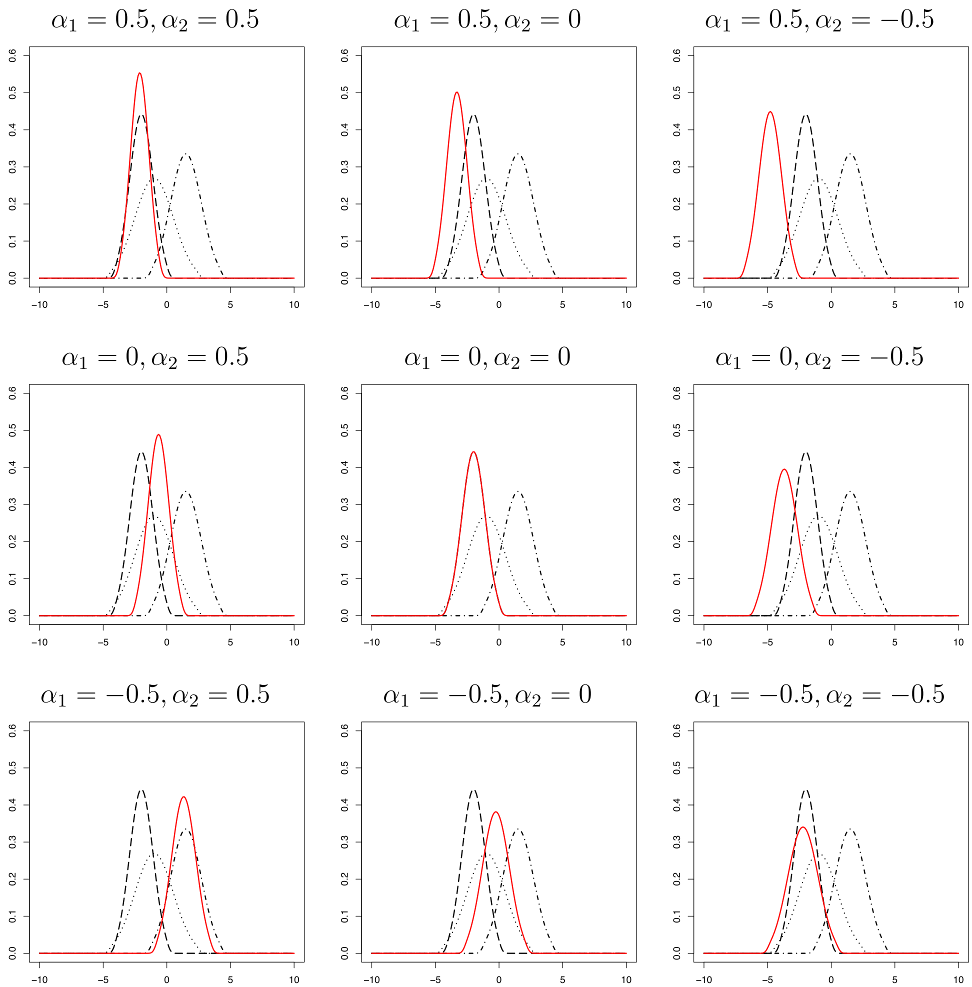

Figure 4. Illustrating ATM$_d$(2) and ATM$_d$(1) models for distributional time series. Each panel depicts the density functions for distributions μ_1 (dotted), μ_2 (dot-dashed) and μ_3 (dashed) (these are the same across all panels) and the density of distribution μ_4 generated by ATM$_d$ (solid), which varies across panels.

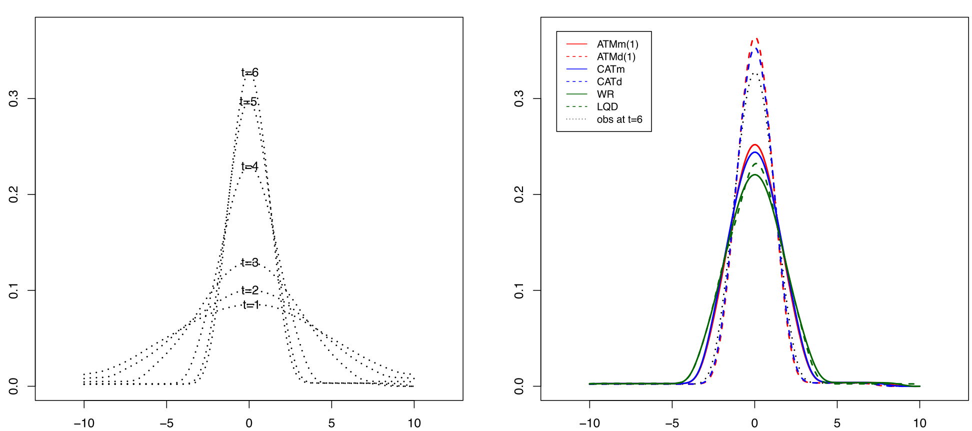

Figure 5. Left panel: The training sample introduced in Section 6.2. Right panel: The one-step forecasts obtained for different methods at $t = 7$, where only predictions obtained from ATM$_d$(1) and CAT$_d$ reflect the declining trend in variances, as only these two predictions have smaller variance compared to the last observed density at $t = 6$, which is also plotted on the right panel.

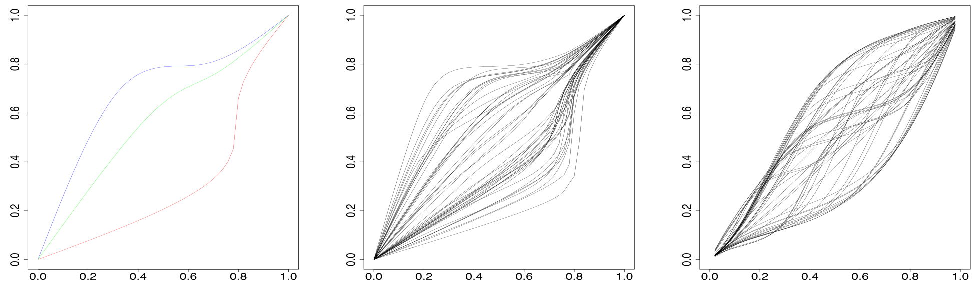

Figure 6. Auxiliary functions for the simulation. Left panel: The monotone function g (upper left curve), g^{-1} (lower right curve) and ϵ_i with $\xi_i = -0.7$ (middle curve) in the simulation (15). Middle panel: Sequence of generated transport maps T_i for simulation (15) with $\alpha_1 = -0.3$, $\alpha_2 = 0.2$. Right panel: Quantile functions generated for simulation (16) with $\alpha_1 = -0.3$, $\alpha_2 = 0.2$.

Table 1. Forecasting accuracy comparison for simulations (15) and (16)

	$(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$	(0.2, −0.5, 0.1, −0.3)	(0.5, 0, 0, 0)
Example (15)	ATM_m	**12.264**	**11.586**
	LQD	13.891	13.282
	WR	12.535	11.765
Example (16)	ATM_m	**9.841**	**9.644**
	LQD	10.079	9.836
	WR	10.082	9.838

Note. The minimizing values are bolded.

between observed and predicted distributions were found to be 0.334 for $\text{ATM}_d(3)$, 1.01 for $\text{ATM}_m(3)$, 0.462 for CAT_d, 1.477 for CAT_m, 1.134 for WR, and 1.255 for LQD. The fitted model coefficients for the best model, i.e., $\text{ATM}_d(3)$, are $\alpha_1 = -0.724$, $\alpha_2 = -0.5$, $\alpha_3 = -0.268$. Denote by $\mu_{2018}, \mu_{2017}, \mu_{2016}, \mu_{2015}$ the observed distributions for the years 2018, 2017, 2016, 2015, respectively, and by T_3 the optimal transport from μ_{2015} to μ_{2016}, by T_2 the optimal transport from μ_{2016} to μ_{2017} and by T_1 the optimal transport from μ_{2017} to μ_{2018}. The training set of distributions, i.e., the observed data, is illustrated in the form of densities in the left panel of Figure 8, predicted densities are in the middle panel and the densities of $\mu_{2015}, \cdots, \mu_{2018}$ in the right panel.

Comparing the densities of μ_{2017} and μ_{2018}, μ_{2016} and those of μ_{2017}, μ_{2015} and μ_{2016}, respectively, we find that T_1 corresponds to a shift to the right and a sharpening of the distribution, T_2 corresponds to a shift to the left and a smoothing of the distribution and T_3 corresponds to a shift to the right and a sharpening of the distribution. The proposed model applies deformations $\alpha_3 \odot T_3, \alpha_2 \odot T_2$ and $\alpha_1 \odot T_1$ sequentially to μ_{2018}. it is likely that ATM_d and CAT_d yield the best results because of the non-stationarity of this sequence, as the distributions shift to the right over the years, reflecting a warming trend.

6.5 U.S. house price data

Given the sequence of distributions $\{\mu_1, \mu_2, \cdots, \mu_n\}$, for a starting time $s_r \in \{k+1, k+2, \cdots, n-k\}$, we used the subset $\{\mu_{s_r}, \mu_{s_r+1}, \cdots, \mu_{s_r+k-1}\}$ to train models and to produce the prediction $\widehat{\mu}_{s_r+k}$ at time $s_r + k$. The autoregressive order p was selected so as to minimize $\sum_{i=s_r}^{s_r+k-1} d_{\mathcal{W}}(\mu_i, \widehat{\mu}_i)$, where $\widehat{\mu}_i$ is the predicted distribution at time i by ATM(p) trained on the sample $\{\mu_{i-k}, \cdots, \mu_{i-1}\}$. The candidate set for p was $\{1, 2, 3, 4, 5\}$ when $k = 8$ and $\{1, 2, 3, 4, 6, 8\}$ when $k > 8$. We adopted the rolling window approach (Zivot & Wang, 2007) and used the prediction loss $\sum_{s_r=k+1}^{n-k} d_{\mathcal{W}}(\mu_{s_r+k}, \widehat{\mu}_{s_r+k})/(n-2k)$.

The US house price data contain bimonthly median house prices for 306 U.S. cities and counties from June 1996 to August 2015 (available at http://www.zillow.com). We adjusted the data to

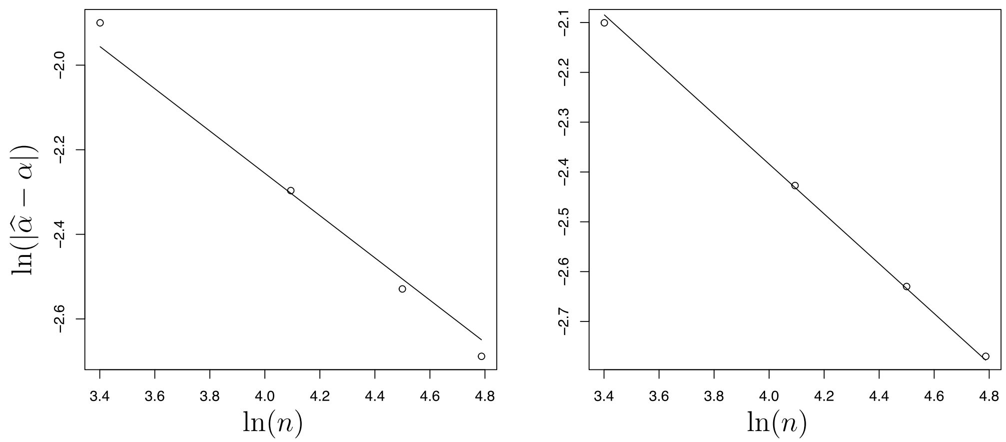

Figure 7. Log-estimation error of $\hat{\alpha}$ versus log sample size n for ATM(1), for $\alpha = 0.5$ (left) and $\alpha = -0.5$ (right). The solid black line in each panel is a line with slope -0.5 that is predicted by theory.

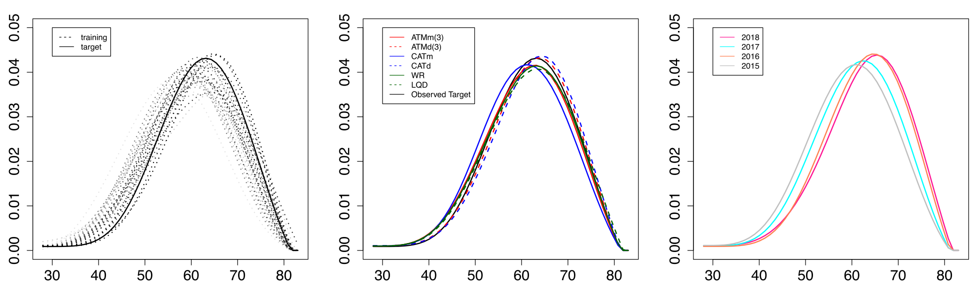

Figure 8. Left panel: Densities of the annual distributions of minimum summer night temperatures at O'Hare International Airport from 1990-2018. Middle panel: Observed density and predicted densities obtained from various models for the year 2019. Right figure shows the densities of μ_{2018}, μ_{2017}, μ_{2016}, μ_{2015} that are the observed distributions for years 2018, 2017, 2016, 2015, respectively.

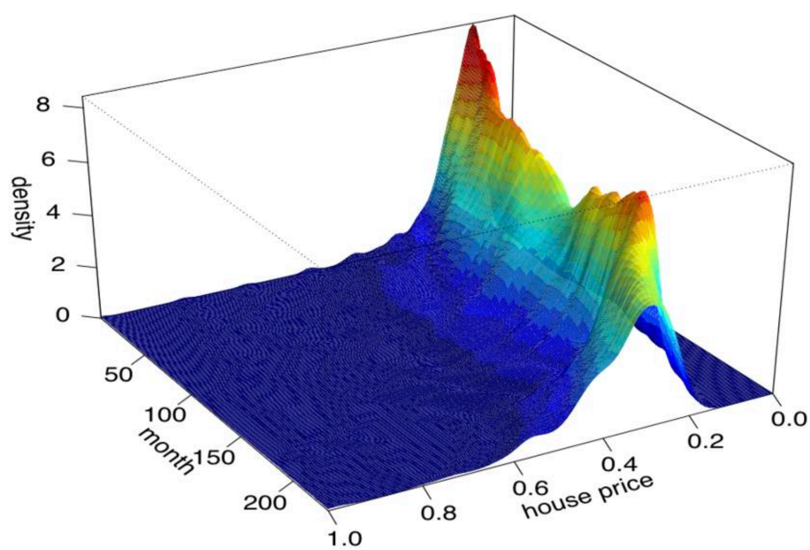

Figure 9. Distributions of US house prices across counties for 240 months between 1996 and 2015, shown as densities.

Table 2. Comparison of prediction errors for the US house price distributional time series, where k is the length of the training set

k	8	12	18	26	36
ATM_m	1.878	1.754	1.771	2.715	2.952
CAT_m	2.660	2.473	2.345	2.327	2.363
ATM_d	**1.647**	**1.611**	**1.652**	**1.708**	**1.778**
CAT_d	1.797	1.787	1.802	1.845	1.924
WR	4.052	3.986	4.074	4.045	4.322
LQD	3.405	3.079	2.927	2.730	2.860

Notes. Actual prediction errors to be multiplied by 10^{-3}. The minimizing values are bolded.

account for inflation by a monthly adjustment factor (deflator) and constructed the bimonthly house price distributions over the 306 cities/counties. The preprocessed distributions (equivalently density or quantile functions) were then scaled to be supported on [0, 1]. Figure 9 presents the house price densities over time. Setting the learning rate $\eta = 1$ in Algorithm 1, the prediction results are summarized in Table 2. In general, ATM_d emerged as the best performing model for these data, which is not surprising due to the non-stationarity of these data.

7 Concluding remarks

Distributional data analysis is challenged by the fact that distributions do not form a vector space and basic operations such as addition and multiplication are not available. This especially affects regression models, including distributional autoregressive models for time series analysis. At the same time, many time series data can be viewed as sequences of distributional data that are indexed by time and there is a need for more advanced statistical tools to model such time series. A key innovation of this paper is that it provides a novel class of regression models for distributional data that are intrinsic and enjoy geometric interpretations. These models result from adopting the point of view that predictors and responses are elements of a space of optimal transports that is equipped with basic algebraic operations. The existence of stationary solutions of the associated ATM models can be guaranteed if a geometric moment-contraction condition is satisfied.

The proposed models not only provide new ways of modelling distributional time series, but also shed light on the possibility of developing models for time series that take values in other geodesic metric spaces. The proposed approach is not limited to optimal transport, and other transports that correspond to geodesics with respect to relevant metrics in distribution spaces could similarly be considered, for example Fisher-Rao transports (Dai, 2022). Modelling time series that take values in the space of multivariate distributions will be a challenging future problem; see also the discussion of this case in Chen et al. (2022).

Acknowledgments

We express our gratitude to the Associate Editor and several referees for very helpful remarks that led to numerous improvements in the paper.

Conflict of interest: We declare no conflict of interest.

Funding

Research supported in part by NSF DMS-2014626 and NIH Echo UH3OD023313

Data availability

The data sets used in this research are openly available at sites https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND and http://www.zillow.com. The computer code to reproduce the results of numerical studies can be accessed at https://drive.google.com/drive/folders/1GCVJSNwgRN7FNYMEQt38RUxPlVYQI7xO?usp=sharing.

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series B* online.

References

Bekierman J., & Gribisch B. (2021). A mixed frequency stochastic volatility model for intraday stock market returns. *Journal of Financial Econometrics*, *19*(3), 496–530. https://doi.org/10.1093/jjfinec/nbz021

Bhatia A., & Katz J. (2021, September 16). *Why we are experiencing so many unusually hot summer nights*. The New York Times, A12.

Bigot J., Gouet R., Klein T., & López A. (2017). Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l'Institut Henri Poincaré B: Probability and Statistics*, *53*, 1–26. https://doi.org/10.1214/15-AIHP706

Bogin A., Doerner W., & Larson W. (2019). Local house price dynamics: New indices and stylized facts. *Real Estate Economics*, *47*(2), 365–398. https://doi.org/10.1111/1540-6229.12233

Bosq D. (2000). *Linear processes in function spaces: Theory and applications*. Springer-Verlag.

Chen Y., Lin Z., & Müller H.-G. (2022). Wasserstein regression. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2021.1956937

Cheng C., & Parzen E. (1997). Unified estimators of smooth quantile and quantile density functions. *Journal of Statistical Planning and Inference*, *59*(2), 291–307. https://doi.org/10.1016/S0378-3758(96)00110-3

Dai X. (2022). Statistical inference on the Hilbert sphere with application to random densities. *Electronic Journal of Statistics*, *16*(1), 700–736. https://doi.org/10.1214/21-EJS1942

Diaconis P., & Freedman D. (1999). Iterated random functions. *SIAM Review*, *41*(1), 45–76. https://doi.org/10.1137/S0036144598338446

Falk M. (1983). Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. *Statistica Neerlandica*, *37*(2), 73–83. https://doi.org/10.1111/j.1467-9574.1983.tb00802.x

Ghodrati L., & Panaretos V. M. (2022). Distribution-on-distribution regression via optimal transport maps. *Biometrika*, *109*(4), 957–974. https://doi.org/10.1093/biomet/asac005

Kloeckner B. (2010). A geometric study of Wasserstein spaces: Euclidean spaces. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, *9*(2), 297–323.

Kokoszka P., Miao H., Petersen A., & Shang H. L. (2019). Forecasting of density functions with an application to cross-sectional and intraday returns. *International Journal of Forecasting*, *35*(4), 1304–1317. https://doi.org/10.1016/j.ijforecast.2019.05.007

Leblanc A. (2012). On estimating distribution functions using Bernstein polynomials. *Annals of the Institute of Statistical Mathematics*, *64*(5), 919–943. https://doi.org/10.1007/s10463-011-0339-4

Matabuena M., & Petersen A. (2021). 'Distributional data analysis with accelerometer data in a NHANES database with nonparametric survey regression model', arXiv, arXiv:2104.01165, preprint: not peer reviewed.

Mazzuco S., & Scarpa B. (2015). Fitting age-specific fertility rates by a flexible generalized skew normal probability density function. *Journal of the Royal Statistical Society Series A*, *178*(1), 187–203. https://doi.org/10.1111/rssa.12053

McCann R. J. (1997). A convexity principle for interacting gases. *Advances in Mathematics*, *128*(1), 153–179. https://doi.org/10.1006/aima.1997.1634

Menafoglio A., Grasso M., Secchi P., & Colosimo B. M. (2018). Profile monitoring of probability density functions via simplicial functional PCA with application to image data. *Technometrics*, *60*(4), 497–510. https://doi.org/10.1080/00401706.2018.1437473

Newey W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, *59*(4), 1161–1167. https://doi.org/10.2307/2938179

Oikarinen E., Bourassa S. C., Hoesli M., & Engblom J. (2018). US metropolitan house price dynamics. *Journal of Urban Economics*, *105*(May 2018), 54–69. https://doi.org/10.1016/j.jue.2018.03.001

Ouellette N., & Bourbeau R. (2011). Changes in the age-at-death distribution in four low mortality countries: A nonparametric approach. *Demographic Research*, *25*(July–December 2011), 595–628. https://doi.org/10.4054/DemRes.2011.25.19

Panaretos V. M., & Zemel Y. (2016). Amplitude and phase variation of point processes. *The Annals of Statistics*, *44*(2), 771–812. https://doi.org/10.1214/15-AOS1387

Pegoraro M., & Beraha M. (2022). Projected statistical methods for distributional data on the real line with the Wasserstein metric. *Journal of Machine Learning Research*, *23*(37), 1–59.

Petersen A., & Müller H.-G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, *44*(1), 183–218. https://doi.org/10.1214/15-AOS1363

Petersen A., & Müller H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, *47*(2), 691–719. https://doi.org/10.1214/17-AOS1624

Shang H. L., & Hyndman R. J. (2017). Grouped functional time series forecasting: An application to age-specific mortality rates. *Journal of Computational and Graphical Statistics*, *26*(2), 330–343. https://doi.org/10.1080/10618600.2016.1237877

Shorack G. R., & Wellner J. A. (2009). *Empirical processes with applications to statistics*. SIAM.

Villani C. (2003). *Topics in optimal transportation*. American Mathematical Society.

Wu W. B., & Shao X. (2004). Limit theorems for iterated random functions. *Journal of Applied Probability*, *41*(2), 425–436. https://doi.org/10.1239/jap/1082999076

Zhang C., Kokoszka P., & Petersen A. (2022). Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis*, *43*(1), 30–52. https://doi.org/10.1111/jtsa.12590

Zivot E., & Wang J. (2007). *Modeling financial time series with S-PLUS®*. (Vol. *191*). Springer Science & Business Media.