JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. **43**: 30–52 (2022)
Published online 07 May 2021 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12590

ORIGINAL ARTICLE

WASSERSTEIN AUTOREGRESSIVE MODELS FOR DENSITY TIME SERIES

CHAO ZHANG[a] PIOTR KOKOSZKA[b*] AND ALEXANDER PETERSEN[a,c]

[a]*Department of Statistics and Applied Probability, University of California Santa Barbara, Santa Barbara, CA, USA*
[b]*Department of Statistics Colorado State University, Fort Collins, CO, USA*
[c]*Department of Statistics Brigham Young University, Provo, UT, USA*

Data consisting of time-indexed distributions of cross-sectional or intraday returns have been extensively studied in finance, and provide one example in which the data atoms consist of serially dependent probability distributions. Motivated by such data, we propose an autoregressive model for density time series by exploiting the tangent space structure on the space of distributions that is induced by the Wasserstein metric. The densities themselves are not assumed to have any specific parametric form, leading to flexible forecasting of future unobserved densities. The main estimation targets in the order-p Wasserstein autoregressive model are Wasserstein autocorrelations and the vector-valued autoregressive parameter. We propose suitable estimators and establish their asymptotic normality, which is verified in a simulation study. The new order-p Wasserstein autoregressive model leads to a prediction algorithm, which includes a data driven order selection procedure. Its performance is compared to existing prediction procedures via application to four financial return data sets, where a variety of metrics are used to quantify forecasting accuracy. For most metrics, the proposed model outperforms existing methods in two of the data sets, while the best empirical performance in the other two data sets is attained by existing methods based on functional transformations of the densities.

Received 10 November 2020; Accepted 09 April 2021

Keywords: Random densities; Wasserstein metric; time series; distributional forecasting

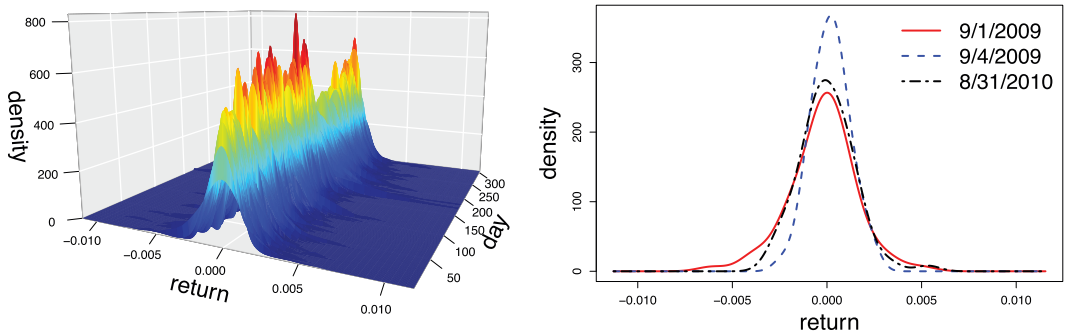
MOS subject classification: 62G05; 62G20; 62M10.

1. INTRODUCTION

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

An early approach to the analysis of distributional data by Kneip and Utikal (2001) used cross-sectional averaging and functional principal component analysis (FPCA) applied directly to yearly income densities. In a more recent work, Yang *et al.* (2020) represented the sample of distributions by their quantile functions, and applied a linear function-on-scalar regression model with quantile functions as response variables. These two approaches are principled alternatives to naively apply methods of functional data analysis (FDA) to density-valued data. Since there are a variety of functional representations that provide unique characterizations of the distributions, including densities, quantile functions, and cumulative distribution functions, one faces the need to choose a representation prior to applying the (typically linear) methods of functional data analysis. Further complicating this

*Correspondence to: Piotr Kokoszka, Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA.
E-mail: piotr.kokoszka@colostate.edu



(a) 9/1/2009 - 11/17/2010 (b) 9/1/2009, 9/4/2009 and 8/31/2010

Figure 1. Densities of XLK, the technology select sector SPDR Fund 5-minute intraday returns on selected dates

dilemma is the fact that these standard functional representations do not constitute linear spaces due to inherent nonlinear constraints (e.g., monotonicity for quantile functions or positivity and mass constraints for densities), so that outputs from models with linear underlying structures are generally inadequate. For this reason, methodological developments for the analysis of distributional data have taken a geometric approach over the last decade. Rather than choosing a functional form under which to analyze the data, one chooses a metric on the space of distributions to develop coherent models. Examples of suitable metrics that have been used successfully in the modeling of distributional data include the Fisher–Rao metric (Srivastava *et al.*, 2007), an infinite-dimensional version of the Aitchison metric (Egozcue *et al.*, 2006; Hron *et al.*, 2016), and the Wasserstein optimal transport metric (Panaretos and Zemel, 2016; Bigot *et al.*, 2017; Petersen and Müller, 2019).

In many cases, the distributions in a sample are indexed by time, for example annual income, fertility and mortality data, or financial returns or insurance claims at various time resolutions. In this article, we will assume that all such distributions possess a density with respect to the Lebesgue measure, and will refer to this type of data as a density time series. A motivating example is shown in Figure 1, depicting the distribution of 5-minute intraday returns of the XLK fund, which tracks the technology and telecommunication sectors within the S&P 500 index. The data we plot in Figure 1a covers 305 trading days, each with 78 records of 5-minute intraday return. Figure 1b demonstrates an alternative look at this dataset by plotting returns from three selected trading days. Kokoszka *et al.* (2019) considered various methods for forecasting density time series, most of which produced forecasts by first applying FPCA to the densities (or transformations of these), followed by fitting a multivariate time series models to the vectors of coefficients. Finally, the density forecasts were obtained by using the forecasts of the coefficients in the FPCA basis representation. Of these different methods, a modified version of the transformation of Petersen *et al.* (2016) gave superior forecasts in the majority of cases, and was also based on a sound theoretical justification in terms of explicitly controlling for the density constraints.

The main contribution of this article is to develop a geometric approach to density time series modeling under the Wasserstein metric. It is well-known that this geometry is intimately connected with quantile functions, and thus provides a flexible framework for modeling samples of densities that tend to exhibit “horizontal" variability, which can be thought of as variability of the quantiles. Examples of such variability in densities are given in Figure 1b. We develop theoretical foundations of autoregressive modeling in the space of densities equipped with the Wasserstein metric, followed by methodology for estimation and forecasting, including order selection. Since the Wasserstein geometry is not linear, care needs to be taken to ensure the model components and their restrictions are appropriately specified. Autoregressive models have been the backbone of time series analysis for scalar and vector-valued data for many decades, see for example, Lütkepohl (2006), among many other excellent textbooks. Autoregression has been extensively studied in the context of linear functional time series; most articles study or use order one autoregression, see Bosq (2000) and Horváth and Kokoszka (2012). This article thus merges two successful approaches: the Wasserstein geometry and time series autoregression.

DOI: 10.1111/jtsa.12590

In a very recent preprint, Chen *et al.* (2020) independently proposed a similar geometric approach to regression when distributions appear as both predictors and responses. As an extension of this formulation, they also developed an autoregressive model of order one for distribution-valued time series. Our AR(1) model proposed in Section 3.1 can be viewed as a special case of the model in Chen *et al.* (2020). However, the generalization, theory and methodology we subsequently pursue move in a completely different direction, so the two articles have little overlap. Even though we were not aware of the work of Chen *et al.* (2020), we did include their model, which is termed the fully functional Wasserstein autoregressive model in this work, as a one of the competing methods in our empirical analyses in Section 5. We also note that our focus on densities with respect to the Lebesgue measure is motivated by practical considerations, as such densities occur in applications. In particular, we formulate numerical algorithms applicable to this common setting. From the theoretical angle, our results related to existence and convergence could be extended to general probability measures. Working with densities actually introduces nontrivial complications. For example, the objects we want to predict must be densities, not general probability measures.

The remainder of the article is organized as follows. In Section 2 we provide the requisite background on Wasserstein geometry and introduce relevant definitions related to density time series. Section 3 is devoted to the development of the Wasserstein AR(p) model, including its estimation and forecasting, both in terms of theory and algorithms. Finite sample properties of our estimator are explored in Section 4, while Section 5 compares our forecasting algorithm to those currently available. We conclude the article with a discussion in Section 6. The Supporting Information contains proofs of the theorems stated in Section 3.

2. PRELIMINARIES

A density time series is a sequence of random densities $\{f_t, t \in \mathbb{Z}\}$. In the spirit of functional data analysis, no parametric form for the densities will be assumed. Furthermore, the models will be developed under the setting in which the densities are completely observed, although in practical situations they will need to be estimated from raw data that they generate. For example, the densities in Figure 1 are kernel density estimates with a Gaussian kernel.

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

2.1. Wasserstein Geometry and Tangent Space

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

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

DOI: 10.1111/jtsa.12590

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

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

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

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

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

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

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

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

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

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

where μ_f is the measure with density f and

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

DOI: 10.1111/jtsa.12590

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

Looking forward to building a Wasserstein autoregressive model, the logarithmic map will be used to lift the random densities into a linear tangent space, where the autoregressive model is imposed. An important point to keep in mind is that the image of $\mathcal{D}$ under Log_f is a convex cone, and thus a nonlinear subset of $\mathcal{T}_\mu \subset L^2(\mathbb{R}, f(u)\mathrm{d}u)$. We will deal with this technicality in the development of Wasserstein autoregressive models in Section 3. In particular, the forecasts produced by the model in the tangent space will not be constrained to lie in the image of the logarithmic map. This poses no practical problem since the forecasted densities are obtained through the exponential map, which is defined on the entirety of the tangent space.

2.2. Wasserstein Mean, Variance, and Covariance

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

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

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

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

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

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

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

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

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

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

DOI: 10.1111/jtsa.12590

as

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

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

2.3. Stationarity of Density Time Series

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

Definition 2.1. A density time series $\{f_t, t \in \mathbb{Z}\}$ is said to be (second-order) stationary if the following two conditions hold.

1. $\mathbb{E}_{\oplus}\left[f_t\right] = f_{\oplus}$ for all $t \in \mathbb{Z}$, so the f_t share a common Wasserstein mean. Denote $supp(f_{\oplus})$ by $D_{\oplus}$.
2. $\mathrm{Var}_{\oplus}(f_t) < \infty$.
3. For any $t, h \in \mathbb{Z}$, and almost all $u, v \in D_{\oplus}$, $C_{t,t+h}(u,v)$ does not depend on t.

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

Definition 2.2. A sequence $\{V_t\}$ of elements of a separable Hilbert space is said to be stationary if the following conditions hold: (i) $\mathbb{E}\left[\|V_t\|^2\right] < \infty$, (ii) $\mathbb{E}\left[V_t\right]$ does not depend on t, and (iii) the autocovariance operators defined by $\mathcal{G}_{t,t+h}(x) = \mathbb{E}\left[\langle (V_t - \mu), x\rangle (V_{t+h} - \mu)\right]$ do not depend on t ($\mu = \mathbb{E}V_0$).

Observe that Definition 2.2 clearly does not apply to the density time series $\{f_t, t \in \mathbb{Z}\}$ as densities do not form a vector space. The fact alone that differences $f_t - \mathbb{E}\left[f_{\oplus}\right]$ are not well-defined in a nonlinear space renders Definition 2.2 unsuitable for density time series. However, up taking $V_t = \mathrm{Log}_{f_{\oplus}}(f_t)$, Definition 2.1 implies Definition 2.2, with the separable Hilbert space in the latter being the tangent space $\mathcal{T}_{f_{\oplus}}$. As has been observed elsewhere (e.g., Panaretos and Zemel, 2016; Petersen *et al.*, 2016), the Wasserstein mean $f_{\oplus}$ (when it exists) is characterized by being the unique solution to $\mathbb{E}\left[\mathrm{Log}_{f_{\oplus}}(f_t)(u)\right] = 0$ for almost all u in the support of $f_{\oplus}$. Hence, condition (ii) is satisfied since $\mu = \mathbb{E}\left[V_0\right] = 0$, from which condition (i) follows as $\mathbb{E}\left[\|V_t\|^2\right] = \mathrm{Var}_{\oplus}(f_t) < \infty$. Lastly, condition (iii) holds since, for any element $x \in \mathcal{T}_{f_{\oplus}}$,

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

which is independent of t. Equivalently, if Q_t is the quantile function corresponding to f_t, Definition 2.1 implies that the optimal transport maps $T_t = Q_t \circ F_{\oplus} = X_t + \mathrm{id}$ form a stationary sequence in $\mathcal{T}_{f_{\oplus}}$ according to Definition 2.2 with $\mu = \mathrm{id}$.

Definition 2.3. A density time series $\{f_t, t \in \mathbb{Z}\}$ is said to be strictly stationary if the joint distributions on $\mathcal{D}^k$ of $(f_{t_1}, f_{t_2}, \ldots, f_{t_k})$ and $(f_{t_1+h}, f_{t_2+h}, \ldots, f_{t_k+h})$ are the same for any $k \in \mathbb{N}$ and choices $t_1, t_2, \ldots, t_k, h \in \mathbb{Z}$.

DOI: 10.1111/jtsa.12590

Note that, if the densities f_t share a common Wasserstein mean $f_\oplus$ and the joint distributions of $(T_{t_1}, T_{t_2}, \dots, T_{t_k})$ and $(T_{t_1+h}, T_{t_2+h}, \dots, T_{t_k+h})$ are the same for any $k \in \mathbb{N}$ and choices $t_1, t_2, \dots, t_k, h \in \mathbb{Z}$, then $\{f_t, t \in \mathbb{Z}\}$ is strictly stationary according to Definition 2.3. Since the existence and uniqueness of the Wasserstein mean implies that the Wasserstein variance is finite, it also follows that $\{f_t, t \in \mathbb{Z}\}$ is stationary according to Definition 2.1, provided the Wasserstein mean exists and is unique.

3. WASSERSTEIN AUTOREGRESSION

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

$$\gamma_h(u, v) := \mathrm{Cov}\left(f_t \ominus f_\oplus(u), f_{t+h} \ominus f_\oplus(v)\right). \tag{3.1}$$

3.1. Wasserstein AR Model of Order 1

From Definition 2.1, a useful path to pursue in developing an autoregressive model for density time series is to first establish a suitable primary model for a sequence $\{V_t\}$ on a tangent space $\mathcal{T}_{f_\oplus}$, for some $f_\oplus \in \mathcal{D}$. Recall that $\mathcal{T}_{f_\oplus}$ is a separable Hilbert space. The second step is to impose conditions on $\{V_t\}$ such that

(a) the measures $\mu_t = \mathrm{Exp}_{f_\oplus}(V_t)$ possess densities f_t that form a stationary density time series with Wasserstein mean $f_\oplus$, and
(b) the parameters in the primary model can still be estimated given observations of the f_t.

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

$$C_\epsilon(u, v) = \mathrm{Cov}\left[\epsilon_t(u), \epsilon_t(v)\right], \quad u, v \in \mathbb{R}. \tag{3.2}$$

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

$$V_t = \beta V_{t-1} + \epsilon_t, \quad t \in \mathbb{Z}. \tag{3.3}$$

As will be detailed in Theorem 3.1, (3.3) has a unique, suitably convergent, solution $V_t = \sum_{i=0}^{\infty} \beta^i \epsilon_{t-i}$ under the following conditions:

(A1) $|\beta| < 1$,
(A2) The innovations are i.i.d. elements of $\mathcal{T}_{f_\oplus}$, have mean zero, and $\int_{\mathbb{R}} C_\epsilon(u, u) f_\oplus(u) \mathrm{d}u < \infty$.

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

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

DOI: 10.1111/jtsa.12590

Denote the usual Hilbert norm on $L^2(\mathbb{R}, f_\oplus(u)\mathrm{d}u)$ by $\|\cdot\|$. We now state our first result associated with model (3.3), and its consequences for the density time series induced by the exponential map. Its proof, along with those of all other theoretical results, can be found in the Supporting Information.

Theorem 3.1. If (A1) and (A2) hold, then

$$V_t = \sum_{i=0}^{\infty} \beta^i \epsilon_{t-i} \tag{3.4}$$

defines a unique, strictly stationary solution in $\mathcal{T}_{f_\oplus}$ to model (3.3). This solution converges strongly,

$$\lim_{n\to\infty} \left\| V_t - \sum_{i=0}^{n} \beta^i \epsilon_{t-i} \right\| = 0 \text{ almost surely,} \tag{3.5}$$

and in mean square,

$$\lim_{n\to\infty} \mathbb{E} \left\| V_t - \sum_{i=0}^{n} \beta^i \epsilon_{t-i} \right\|^2 = 0. \tag{3.6}$$

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

In light of Theorem 3.1, we define the Wasserstein autoregressive model of order 1, or WAR(1) model, for a density time series $\{f_t, t \in \mathbb{Z}\}$ by

$$T_t - \mathrm{id} = \beta(T_{t-1} - \mathrm{id}) + \epsilon_t. \tag{3.7}$$

Under (A1)–(A3), we now know that a unique solution $f_t \ominus f_\oplus = T_t - \mathrm{id} = \sum_{i=0}^{\infty} \beta^i \epsilon_{t-i}$ exists such that $\{f_t, t \in \mathbb{Z}\}$ is strictly stationary according to Definition 2.3. Since they also share a common Wasserstein mean, the sequence is also stationary according to Definition 2.1.

In order for the results of Theorem 3.1 to not be vacuous, we will establish a set of innovation examples that satisfy (A2) and (A3). Given the structure of the tangent space in (2.3), consider innovations of the form $\epsilon_t(u) = \lambda_t(S_t(u) - u)$, where $\lambda_t > 0$ and S_t is an increasing map defined on $D_\oplus$ (and is thus an optimal transport map from $f_\oplus$ to some $\nu \in \mathcal{W}_2$). Both λ_t and S_t can be random. We now list specific examples for which (A2) and (A3) hold, where $|\beta| < 1$ throughout.

Example 3.1. Let η_t be i.i.d. random variables with mean zero and finite variance. The WAR(1) model admits constant innovations $\epsilon_t(u) \equiv \eta_t$, which can be identified as elements in $\mathcal{T}_{f_\oplus}$ by setting $S_t(u) = \eta_t \lambda_t^{-1} + u$ for any $\lambda_t > 0$.

Example 3.2. Let η_t be as in Example 3.1, and δ_t be i.i.d. random variables with mean zero such that $|\delta_t| < 1 - |\beta|$. Linear innovations $\epsilon_t(u) = \eta_t + \delta_t u$ are admissible under the WAR(1) model. The tangent space representation of $\epsilon_t(u)$ can be recovered by setting $S_t(u) = (1 + \delta_t \lambda_t^{-1})u + \eta_t \lambda_t^{-1}$ for any $\lambda_t > |\delta_t|$.

Example 3.3. Let η_t and δ_t be as in Example 3.2, with the additional constraint that the δ_t be symmetric about 0. The WAR(1) model admits periodic innovations $\epsilon_t(u) = \eta_t + \sin(\delta_t u)$, which can be viewed as tangent space elements by writing $S_t(u) = u + \eta_t \lambda_t^{-1} + \lambda_t^{-1} \sin(\delta_t u)$ for any $\lambda_t > |\delta_t|$.

DOI: 10.1111/jtsa.12590

In Examples 3.1–3.3, (A2) is clearly satisfied. Moreover, we have $\epsilon_t'(u) = 0$, $\epsilon_t'(u) = \delta_t$ and $\epsilon_t'(u) = \delta_t \cos(\delta_t u)$, respectively in each example. Thus, $\sup_{u\in D_\oplus} |\epsilon_t'(u)| \leq 1 - |\beta|$, so that differentiation and summation can be interchanged, yielding

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

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

Next, we express the autoregressive coefficient β in terms of the autocovariance functions γ_h defined in (3.1). Following the derivation of the Yule–Walker equations, it can be shown that

$$\beta = \frac{\int_{\mathbb{R}} \gamma_1(u,u) f_\oplus(u) \mathrm{d}u}{\int_{\mathbb{R}} \gamma_0(u,u) f_\oplus(u) \mathrm{d}u}. \tag{3.8}$$

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

3.1.1. Estimation and Forecasting

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

$$\rho_h = \frac{\int_{\mathbb{R}} \gamma_h(u,u) f_\oplus(u) \mathrm{d}u}{\int_{\mathbb{R}} \gamma_0(u,u) f_\oplus(u) \mathrm{d}u} = \frac{\int_{\mathbb{R}} \eta_h(u) f_\oplus(u) \mathrm{d}u}{\int_{\mathbb{R}} \eta_0(u) f_\oplus(u) \mathrm{d}u}, \quad \eta_h(u) = \gamma_h(u,u). \tag{3.9}$$

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

$$\hat{f}_\oplus(u) = \widehat{F}_\oplus'(u), \quad \widehat{F}_\oplus = \left(\frac{1}{n} \sum_{t=1}^{n} Q_t \right)^{-1}. \tag{3.10}$$

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

$$\hat{\rho}_h = \frac{\int_{\mathbb{R}} \hat{\eta}_h(u) \hat{f}_\oplus(u) \mathrm{d}u}{\int_{\mathbb{R}} \hat{\eta}_0(u) \hat{f}_\oplus(u) \mathrm{d}u}, \quad \hat{\eta}_h(u) = \frac{1}{n} \sum_{t=1}^{n-h} \left\{ \widehat{T}_t(u) - u \right\} \left\{ \widehat{T}_{t+h}(u) - u \right\}. \tag{3.11}$$

Then the natural estimator for β in (3.7) is

$$\hat{\beta} = \hat{\rho}_1. \tag{3.12}$$

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

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

The following result is a special case of Theorem 3.4 in Section 3.2; the proof of the more general result can be found in the Supporting Information.

DOI: 10.1111/jtsa.12590

Theorem 3.2. Suppose (A1)–(A4) hold. Then

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

where

$$\sigma_\epsilon^2 = \frac{\int_{\mathbb{R}^2} C_\epsilon^2(u,v) f_\oplus(u) f_\oplus(v) \mathrm{d}u\mathrm{d}v}{\left[\int_{\mathbb{R}} C_\epsilon(u,u) f_\oplus(u)\mathrm{d}u\right]^2} \tag{3.13}$$

is finite due to (A4).

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

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

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

$$\begin{aligned}\widehat{F}_{n+1}(u) &= \int_{\mathbb{R}} \mathbf{1}\left(\widehat{V}_{n+1}(v) + v \le u\right) \hat{f}_\oplus(v)\mathrm{d}v \\ &= \int_0^1 \mathbf{1}\left(\hat{\beta} Q_n(s) + (1-\hat{\beta})\widehat{Q}_\oplus(s) \le u\right) \mathrm{d}s,\end{aligned} \tag{3.14}$$

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

$$\hat{\beta} = \frac{\int_0^1 \hat{\lambda}_1(s)\mathrm{d}s}{\int_0^1 \hat{\lambda}_0(s)\mathrm{d}s}, \quad \hat{\lambda}_h(s) = \hat{\eta}_h(\widehat{Q}_\oplus(s)) = \frac{1}{n}\sum_{t=1}^{n-h}(Q_t(s) - \widehat{Q}_\oplus(s))(Q_{t+h}(s) - \widehat{Q}_\oplus(s)). \tag{3.15}$$

Since $\hat{\lambda}_h(s)$ is computed for $s \in [0,1]$, (3.15) emphasizes that the input densities f_t need not share the same support or be estimated over an identical grid, since all the critical calculations are carried out in terms of quantile functions. Only the quantile functions of the density time series need to be estimated over the same grid points, which extends the flexibility of the model.

The first step of the algorithm is to convert the available densities f_t into quantile functions. A simple approach to obtain these quantiles from densities is to first evaluate smooth cumulative distribution functions by integrating the estimated densities, followed by some form of numerical inversion. One such approach is readily implemented by the R function dens2quantile from package fdadensity, and this is the approach taken in our numerical

DOI: 10.1111/jtsa.12590

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

Input: densities $f_t, t = 1, 2, \ldots, n$; grid QSup spanning $[0, 1]$

`/* Quantities in steps 2-6 are evaluated for` $s \in$ QSup `*/`

Evaluate $Q_1(s), Q_2(s), \ldots, Q_n(s)$;

$\widehat{Q}_{\oplus}(s) \leftarrow n^{-1} \sum_{t=1}^{n} Q_t(s)$;

$\hat{\lambda}_h(s) \leftarrow n^{-1} \sum_{t=1}^{n-h} (Q_t(s) - \widehat{Q}_{\oplus}(s))(Q_{t+h}(s) - \widehat{Q}_{\oplus}(s))$, $h = 0, 1$;

$\hat{\beta} \leftarrow \int_0^1 \hat{\lambda}_1(s)\mathrm{d}s / \int_0^1 \hat{\lambda}_0(s)\mathrm{d}s$;

$\widehat{V}_{n+1}(\widehat{Q}_{\oplus}(s)) \leftarrow \hat{\beta}(Q_n(s) - \widehat{Q}_{\oplus}(s))$;

Generate grid dSup spanning $(\min_{s \in QSup} \widehat{V}_{n+1}(\widehat{Q}_{\oplus}(s)) + \widehat{Q}_{\oplus}(s), \max_{s \in QSup} \widehat{V}_{n+1}(\widehat{Q}_{\oplus}(s)) + \widehat{Q}_{\oplus}(s))$

`/* Quantities in steps 8-10 are evaluated for` $u \in$ dSup `*/`

Compute $\{[a_l, b_l]\}_{l=1}^{L(u)} \leftarrow \left\{ s \in [0, 1] : \widehat{V}_{n+1}(\widehat{Q}_{\oplus}(s)) + \widehat{Q}_{\oplus}(s) \le u \right\}$;

`/*` $\{[a_l, b_l]\}_{l=1}^{L(u)}$ `are disjoint subintervals of` $[0, 1]$. `*/`

$\widehat{F}(u)_{n+1} \leftarrow \sum_{l=1}^{L(u)} (b_l - a_l)$;

$\hat{f}(u)_{n+1} \leftarrow \widehat{F}'(u)_{n+1}$

experiments to achieve step 2 of the algorithm. Steps 7–9 demonstrate how to implement the exponential map defined in (2.5). From this definition, it is clear that the support of the forecasted density is given by the formula in step 7. Steps 8 and 9 then discover and evaluate the probabilities $\mathrm{Exp}_{\hat{f}_{\oplus}}(\widehat{V}_{n+1})((-\infty, u])$, for u in the support of the forecasted measure. Finally, step 10 can be executed by numerical integration, for example by computing differences.

3.2. Wasserstein AR Model of Order p

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

$$T_t - \mathrm{id} = \sum_{j=1}^{p} \beta_j (T_{t-j} - \mathrm{id}) + \epsilon_t, \tag{3.16}$$

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

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

The WAR(p) model in (3.16) can then be written as

$$\phi(B)\left(T_t - \mathrm{id}\right) = \epsilon_t, \tag{3.17}$$

where B is the backward shift operator, that is, for a discrete stochastic process $\{X_t, t \in \mathbb{Z}\}$, $B^i X_t = X_{t-i}$, $i \in \mathbb{Z}$. For the WAR(p) to have a causal solution, we make the following assumption as a generalization of (A1) in Section 3.1.

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

DOI: 10.1111/jtsa.12590

Under (A1′), $\frac{1}{\phi(z)} = \sum_{i=0}^{\infty} \psi_i z^i$, and the sequence $\{\psi_i\}_{i=0}^{\infty}$ satisfies $\sum_{i=0}^{\infty} |\psi_i| < \infty$. We will show that the solution to (3.17) can be written as

$$T_t - \mathrm{id} = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}. \tag{3.18}$$

Observe (3.18) is a strictly stationary and causal process. Similarly to the development of the WAR(1) model, $\{T_t - \mathrm{id}\}$ in (3.16) should be understood at this point as a general zero mean autoregressive process of order p in $\mathcal{T}_{f_\oplus}$. As shown below, (A1′) and (A2) together imply the existence of a unique, suitably convergent, solution $T_t - \mathrm{id} = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}(u)$ that is stationary in $\mathcal{T}_{f_\oplus}$ according to Definition 2.2. Once again, (A3) applied to $V_t = T_t - \mathrm{id}$ ensures that the application of the exponential map to $T_t - \mathrm{id}$ produces a stationary density time series with mean $f_\oplus$, as seen in the Theorem 3.3. We also remark that Examples 3.1–3.3 can be modified directly to guarantee the viability of the WAR(p) model; essentially $1 - |\beta|$ must be replaced with $\left(\sum_{i=0}^{\infty} |\psi_i|\right)^{-1}$.

Theorem 3.3. The following claims hold under Assumptions (A1′) and (A2).

(i) The series (3.18) is a strictly stationary solution in $\mathcal{T}_{f_\oplus}$ to the WAR(p) (3.16). This solution converges almost surely and in mean square, that is,

$$\lim_{n\to\infty} \left\| T_t - \mathrm{id} - \sum_{i=0}^{n} \psi_i \epsilon_{t-i} \right\| = 0 \quad a.s., \tag{3.19}$$

and

$$\lim_{n\to\infty} \mathbb{E} \left\| T_t - \mathrm{id} - \sum_{i=0}^{n} \psi_i \epsilon_{t-i} \right\|^2 = 0. \tag{3.20}$$

(ii) There is no other stationary solution (according to Definition 2.2) in $\mathcal{T}_{f_\oplus}$.

(iii) If, in addition, Assumption (A3) holds for $V_t = T_t - \mathrm{id}$, then T_t is strictly increasing, almost surely, and the measures $\mathrm{Exp}_{f_\oplus}(T_t - \mathrm{id})$ possess densities f_t that form a strictly stationary sequence according to Definition 2.1 with common Wasserstein mean $f_\oplus$.

Questions of the existence and uniqueness of solutions to ARMA equations are not obvious beyond the setting of scalar innovations, even though care must be exercised even in that standard case, as explained in Chapter 3 of Brockwell and Davis (1991). In the multivariate case, conditions on the spectral decomposition of the autoregressive matrices are needed, see Brockwell and Lindner (2010) and Brockwell *et al.* (2013) whose results were extended to Banach spaces by Spangenberg (2013). Simpler sufficient conditions in Hilbert spaces are given in Bosq (2000) (AR(p) case) and Klepsch *et al.* (2017) (ARMA(p, q) case). In our setting, the coefficients are scalars, but the innovations must conform to a nonlinear functional structure, so our conditions involve an interplay between the structure of the functional noise and the coefficients. The fully functional WAR(1) considered in Chen *et al.* (2020) is also constructed in the tangent space, so it is also subject to similar constraints as our WAR(p) model, namely that the solution must be restricted to image of the logarithmic map with probability one. We have addressed it through our assumption (A3) and suitable examples or error sequences. Assumption (B2) in Chen *et al.* (2020) is general, and it is, at this point, unclear whether concrete examples of innovations can be established that satisfy it for fully functional WAR models.

3.2.1. Estimation and Forecasting

Recall $\hat{f}_\oplus$, η_h and $\hat{\eta}_h$ as defined in (3.10), (3.9) and (3.11) respectively. Set $\{\mathbf{H}_p(u)\}_{jk} = \eta_{|j-k|}(u)$, $j, k = 1, \ldots, p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, and $\boldsymbol{\eta}_p(u) = (\eta_1(u), \ldots, \eta_p(u))^\top$. Following the derivation of the Yule–Walker equations, we obtain

DOI: 10.1111/jtsa.12590

$\mathbf{H}_p(u)\boldsymbol{\beta} = \boldsymbol{\eta}_p(u)$ as a characterization of the autoregressive parameters of the WAR(p) model, whence

$$\boldsymbol{\beta} = \left(\int_{\mathbb{R}} \mathbf{H}_p(u) f_{\oplus}(u) \mathrm{d}u \right)^{-1} \int_{\mathbb{R}} \boldsymbol{\eta}_p(u) f_{\oplus}(u) \mathrm{d}u, \tag{3.21}$$

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

$$\widehat{\boldsymbol{\beta}} = \left(\int_{\mathbb{R}} \widehat{\mathbf{H}}_p(u) \hat{f}_{\oplus}(u) \mathrm{d}u \right)^{-1} \int_{\mathbb{R}} \hat{\boldsymbol{\eta}}_p(u) \hat{f}_{\oplus}(u) \mathrm{d}u. \tag{3.22}$$

Set $\{\boldsymbol{\Psi}_p\}_{ij} = \sum_k \psi_k \psi_{k+|i-j|}$, $i,j = 1, \dots, p$. The following theorem establishes the asymptotic normality of the estimator (3.22).

Theorem 3.4. Suppose (A1′), (A2), (A3), and (A4) hold. Then

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \mathbf{N}(0, \boldsymbol{\Sigma}), \tag{3.23}$$

where $\Sigma_{ij} = \sigma_\epsilon^2 \left\{ \boldsymbol{\Psi}_p^{-1} \right\}_{ij}$, $i,j = 1, \dots, p$, and σ_ϵ^2 is the same as (3.13) in Theorem 3.2.

Indeed the above asymptotic covariance matrix is a generalization of the asymptotic variance in Theorem 3.2. The forecasting procedure is exactly the same as described in (3.14) with steps (4) and (5) of Algorithm 1 replaced by the above steps for estimating $\boldsymbol{\beta}$ and step (6) becoming

$$\widehat{V}_{n+1} = \sum_{i=1}^{p} \hat{\beta}_i (T_{n-i+1} - \mathrm{id}). \tag{3.24}$$

In addition to the autoregressive parameters, the autocorrelation functions are an important object in the study of time series. In our case, recall the lag-h Wasserstein autocorrelation functions are defined in (3.9). Denote $\boldsymbol{\rho}_h = \left(\rho_1, \rho_2, \dots, \rho_h\right)^{\intercal}$ and $\hat{\boldsymbol{\rho}}_h = \left(\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_h\right)^{\intercal}$, where $\hat{\rho}_i = \int_{\mathbb{R}} \hat{\eta}_i(u) \hat{f}_{\oplus}(u) \mathrm{d}u \big/ \int_{\mathbb{R}} \hat{\eta}_0(u) \hat{f}_{\oplus}(u) \mathrm{d}u$, $i = 1, \dots, h$.

Theorem 3.5. Suppose (A1′), (A2), (A3), and (A4) hold. Then

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

where

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

and the entries v_{jk}, $j,k = 1, \dots, n-1$, of $\mathbf{V}$ are defined in (A.9) and (A.10) in *Lemma* A.2 in the Supporting Information.

4. FINITE SAMPLE PROPERTIES OF AUTOREGRESSIVE PARAMETER ESTIMATORS

Simulations of the WAR(p) model were conducted to show that the autoregressive coefficients β_j can be accurately estimated, and to explore the normality of the estimators in finite samples. The simulation parameters included the

DOI: 10.1111/jtsa.12590

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

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

Wasserstein mean density $f_\oplus$ and quantile function $Q_\oplus$, the autoregressive parameters β_j, and a generative process for the innovations ϵ_t. The relation $Q_t(s) = T_t \circ Q_\oplus$ was used to obtain the quantile functions Q_t for use in our algorithms. Simulations were conducted using different Wasserstein mean densities and innovation processes to probe the sensitivity of estimators. In this section, results are presented for a setting in which the Wasserstein mean density corresponds to the uniform distribution on the unit interval, that is, $Q_\oplus(s) = s$, for $s \in [0, 1]$. Results under more complicated settings can be found in the Section A.5 of the Supporting Information.

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

$$T_t - \mathrm{id} = \beta_1 \left(T_{t-1} - \mathrm{id}\right) + \beta_2 \left(T_{t-2} - \mathrm{id}\right) + \beta_3 \left(T_{t-3} - \mathrm{id}\right) + \epsilon_t, \tag{4.1}$$

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

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

To begin, it is necessary to generate the initial maps T_1, T_2, and T_3. There exists a unique, stationary and causal solution to (4.1) in the form of (3.18). Hence, one can generate the initial signals purely based on past innovations. A burn-in period of $m = 1000$ was used to stabilize the simulated signals. Given a sequence of m burn-in innovations $\{\epsilon_{1-m}, \epsilon_{2-m}, \ldots, \epsilon_{-1}, \epsilon_0\}$ generated as above, based on (3.18), define

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

Then (4.1) can be applied recursively until $T_1 - \mathrm{id}$ through $T_3 - \mathrm{id}$ are obtained. One can then generate a time series of desired lengths with $T_1 - \mathrm{id}$ through $T_3 - \mathrm{id}$ and (4.1). This approach is equivalent to truncating the infinite sum in (3.18) but avoids the calculation of the coefficients ψ_i. In our numerical implementation, an equally spaced grid of length 100 on $[0, 1]$ was used for both u and s arguments, since the support of the Wasserstein mean and that of the quantile functions are both $[0, 1]$ in this setting. The autoregressive parameter estimates in (3.22) were computed using numerical integration.

The simulation was repeated 1000 times with sample sizes $n = 50, 100, 500, 1000, 2000$. The bias, standard deviation and root mean-square error (RMSE) are summarized in Table I, from which we can observe that they all trail off as sample size increases. For the purpose of demonstration, we only display histograms and QQ-plots for $n = 50, 100$ and 1000. The graphical evidence of the asymptotic marginal normality of the estimators $\hat{\beta}_i$, $i = 1, 2, 3$, is presented in Figures 2–4.

To investigate the joint normality, denote $\widehat{\boldsymbol{\beta}}_j = [\hat{\beta}_{1j}, \hat{\beta}_{2j}, \hat{\beta}_{3j}]^\intercal$, where $j = 1, 2, \ldots, 1000$ denotes the number of replicates. We randomly generate three pairs of 3×1, linearly independent unit vectors (v_1, v_2), (v_3, v_4) and (v_5, v_6). Calculate $X_{ij} = v_{ij}^\intercal \widehat{\boldsymbol{\beta}}_j$, $i = 1, 2, \ldots, 6$, $j = 1, 2, \ldots, 1000$. Scatter plots of X_{ij} v.s. $X_{(i+1)j}$, $i = 1, 3, 5$, are shown in Figure 5. The idea is that if a vector $[\beta_1, \beta_2, \beta_3]^\intercal$ is normal, then for any coefficients, the vectors $\sum_{j=1}^3 v_{1j}\beta_j$ and

DOI: 10.1111/jtsa.12590

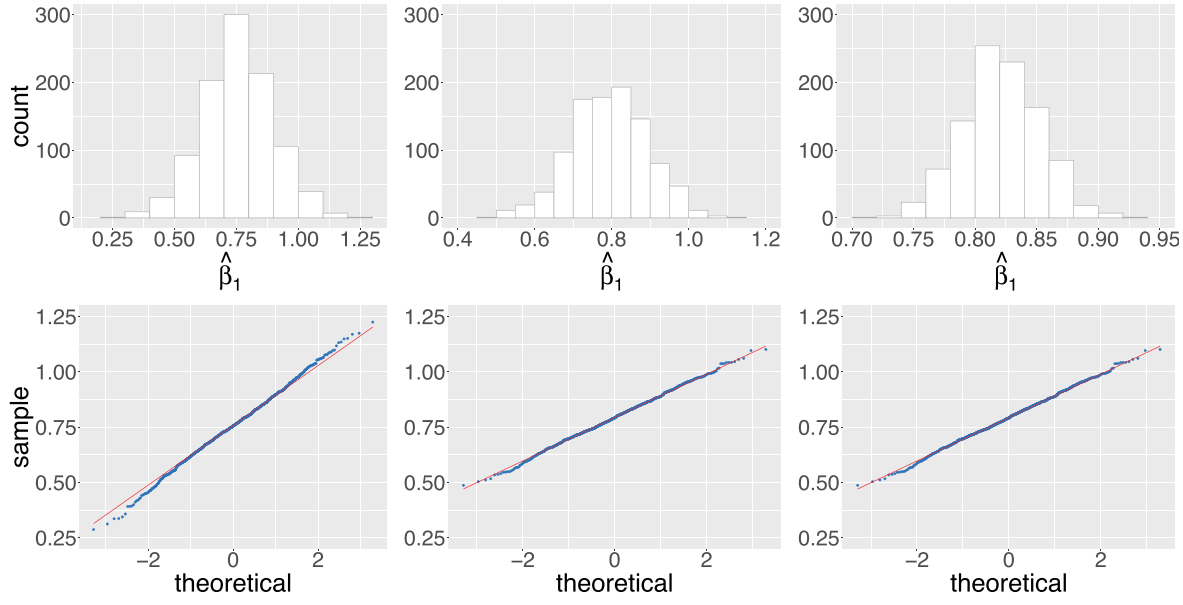


Figure 2. QQ plots and histograms of $\hat{\beta}_1$

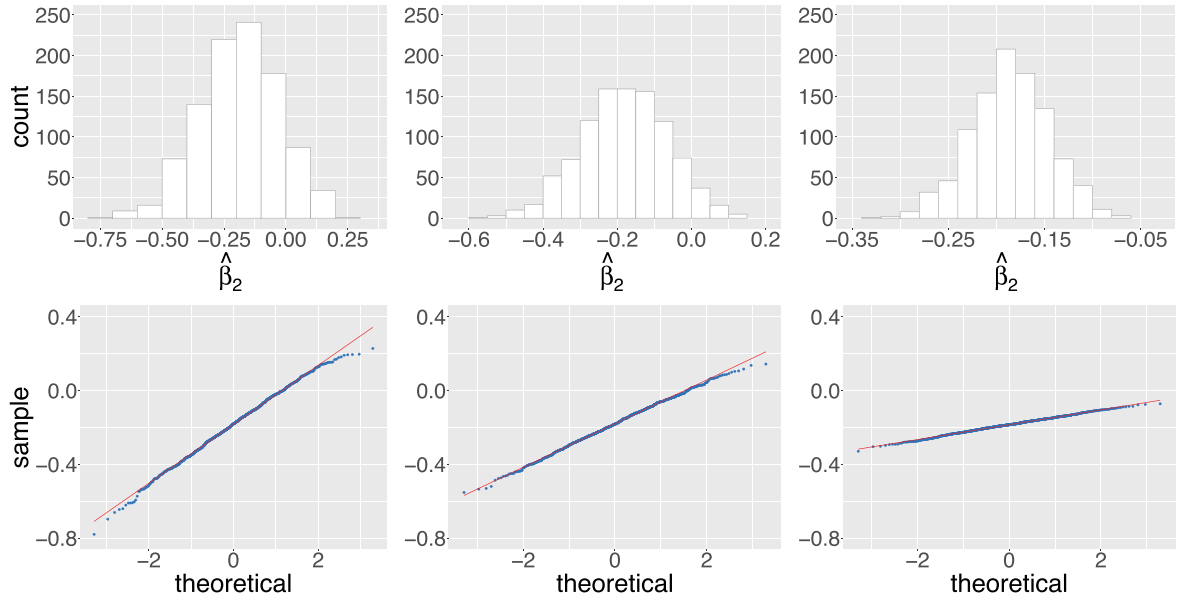
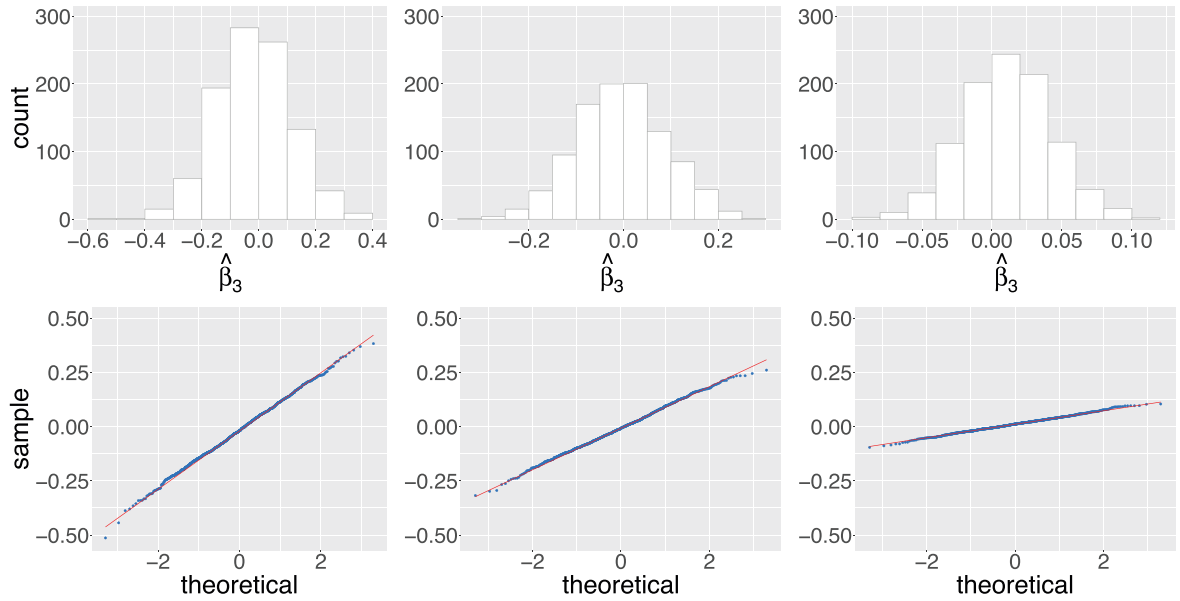


Figure 3. QQ plots and histogram of $\hat{\beta}_2$

J. Time Ser. Anal. **43**: 30–52 (2022)
DOI: 10.1111/jtsa.12590



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

Figure 4. QQ plots and histograms of $\hat{\beta}_3$

$\sum_{j=1}^{3} v_{2j}\beta_j$ have a joint bivariate normal distribution, which can be approximately verified by visual examination of scatter plots, if replications of $[\beta_1, \beta_2, \beta_3]^\intercal$ are available. As before, we only display the cases where $n = 50, 100$ and 1000 for demonstration. The elliptical patterns in Figure 5 suggest bivariate Gaussian distribution, which is what we expect. Moreover, for each n, we calculate $\widehat{\mathbf{\Sigma}}$, the sample covariance matrix of $\{\widehat{\boldsymbol{\beta}}_j, j = 1, 2, \ldots, 1000\}$, which is an estimator of the theoretical covariance matrix $\mathbf{\Sigma}$ in (3.23). Let $\|\cdot\|_F$ be the Frobenius norm, we use the relative Frobenius norm, $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F / \|\mathbf{\Sigma}\|_F$ to measure the differences between the sample covariance matrix and the theoretical asymptotic covariance matrix based on (3.23). Figure 6 shows that the relative difference approaches zero as sample size increases. All the aforementioned evidence supports the result of Theorem 3.4.

5. COMPARISON WITH OTHER FORECASTING METHODS

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

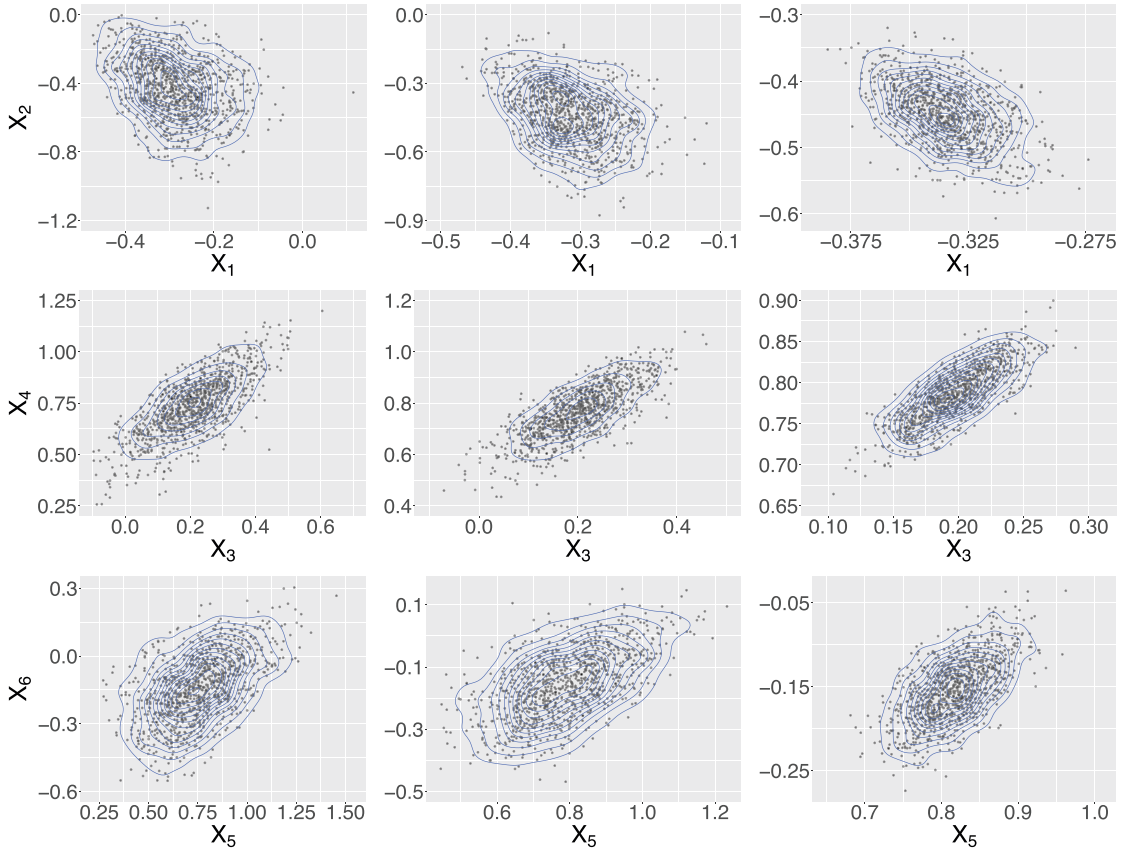
5.1. Benchmark Methods

We consider the following existing methods.

5.1.1. Compositional Data Analysis

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

DOI: 10.1111/jtsa.12590



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

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

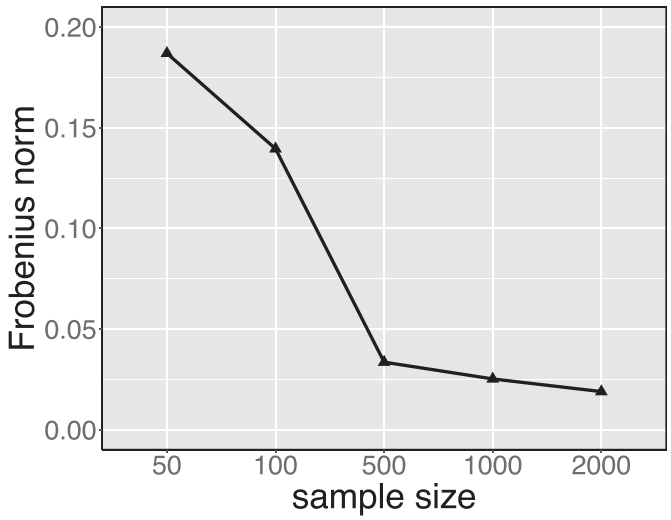


Figure 6. Difference between sample and theoretical covariance matrices

DOI: 10.1111/jtsa.12590

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

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

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

5.2. Data sets and Performance Metrics

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

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

1. The discrete version of Kullback–Leibler divergence (KLD; see Kullback and Leibler, 1951)
2. The square root of the Jensen–Shannon divergence (JSD; see Shannon 1948)
3. L_1 norm.

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

5.3. WAR(p) Models

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

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

Denote by $\hat{p}(n)$ and $\widehat{K}(n)$, the value of p and K which minimizes $R_p(n,K)$, we use WAR($\hat{p}(n)$) and the training window $[n-\widehat{K}(n)+1, n]$ to predict f_{n+1}. One way to implement this data-driven procedure is to select K and p simultaneously, which entails $|I_p| \times |I_K|$ runs of the forecasting algorithm. In our numerical experiments in this section, we observed that the choice of K has greater impact on the forecasting accuracy than the choice of p. In addition, within the data sets we investigated, the choice of K is relatively robust to the choice of p as the number of window sizes are small, that is, $|I_K| = 2$ for intra-day data sets and $|I_K| = 3$ for cross-sectional data sets (see Section 5.5). Therefore, to reduce the computational cost, we first use the WAR(1) model to determine K. After choosing training windows for each day, we then determine the order p.

DOI: 10.1111/jtsa.12590

5.4. Fully Functional WAR(p) Models

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

$$T_t(u) - u = \sum_{j=1}^{p} \int_{\mathbb{R}} \phi_j(u, v)(T_{t-j}(v) - v) f_{\oplus}(v) \mathrm{d}v + \epsilon_t(u), \tag{5.1}$$

where ϕ_j are the autoregressive parametric functions to be recovered. Thus, the key difference between the WAR(p) model proposed in this article and that of Chen *et al.* (2020) is in how the quantities $T_{t-j} - \mathrm{id}$ from previous timepoints are mapped to the tangent space prior to adding the innovations. In the WAR(p) model, these are simply multiplied by the autoregressive coefficients β_j. In contrast, the fully function WAR(p) applies an integral operator with kernel ϕ_j to these quantities. Note that, technically, the WAR(p) model is not a special case of the fully functional version, since the operation of multiplying by β_j is not compact, whereas the integral operators in (5.1) are compact. The estimation procedure follows by fitting the usual functional AR(p) model (see, e.g., Bosq, 2000) to the observed quantile functions Q_t, yielding estimates $\hat{\varphi}_j$ of the kernels $\varphi(s, s') = \phi_j(Q_{\oplus}(s), Q_{\oplus}(s'))$. In the case $p = 1$, this matches the estimation of Chen *et al.* (2020). Similarly to the WAR(p) model, forecasts are then constructed in the tangent space using the plug-in estimates $\hat{\phi}_j(u, v) = \hat{\varphi}_j(\hat{F}_{\oplus}(u), \hat{F}_{\oplus}(v))$, followed by application of the exponential map (2.5). Thus, in the presentation of our results, the method labeled "Fully Functional WAR(p)" can be considered as an extension of the model of Chen *et al.* (2020) to include orders $p \geq 1$.

In particular, we implement the same data-adaptable procedure as described in Section 5.3 with one additional component. The method used to fit the functional AR(p) model to the quantile functions performs functional principal component analysis as a first, which requires one to specify the number of components to retain. We thus introduce an additional tuning parameter R that represents proportion of variance required by the FPCA. Specifically, in the forecasting procedure, we reconstruct $\widehat{T}_t - \mathrm{id}$ with the smallest number of PCs that explain R percent of variance; see, for example, Section 3.3 of Horváth and Kokoszka (2012). We incorporate R into the data-driven procedure to determine its value for forecasting. Specifically, we compute

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

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

5.5. Results

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

DOI: 10.1111/jtsa.12590

Table II. Forecast accuracies of five methods, XLK intraday returns

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

Table III. Forecast accuracies of five methods, Bovespa intraday returns

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

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

The performance of WAR(p) model in the Bovespa and S&P500 data sets is not as competitive. Since our models rely on stationarity, we informally investigate the stationarity condition for each data set. In Figure 7, we plot the Wasserstein distance from all densities used in forecasting to their sample Wasserstein mean. These distances are larger in the Bovespa and S&P500 data sets, compared to those in XLK and DJI data sets. Indeed, the average Wassertein distance from these plots in Figure 7 are XLK: 4.045, Bovespa: 4.255, DJI: 421.25 and S&P500: 571.63. Hence stationarity could be a potential cause for a weaker performance of the WAR(p) model in the Bovespa and S&P500 data sets. Generally, no prediction method can be expected to be uniformly superior across all data sets and all time periods and according to all metrics. In our empirical study, The WAR(p) methods performs best for some data sets, and the LQDT and CoDa methods perform better for others.

6. DISCUSSION

The WAR(p) model provides an interpretable approach to model density time series by representing each density through its optimal transport map from the Wasserstein mean. Under this representation, stationarity of a density time series, whose elements reside in a nonlinear space, is defined according to the usual stationarity of the random transport maps in the tangent space, which is a separable Hilbert space. This article demonstrates how autoregressive models, built on the tangent space corresponding to the Wasserstein mean, possess stationary solutions that, in turn, define a stationary density time series. This link is not automatic, however, due to the fact that the logarithmic map lifting the densities to the tangent space is not surjective, and constraints are necessary to ensure the

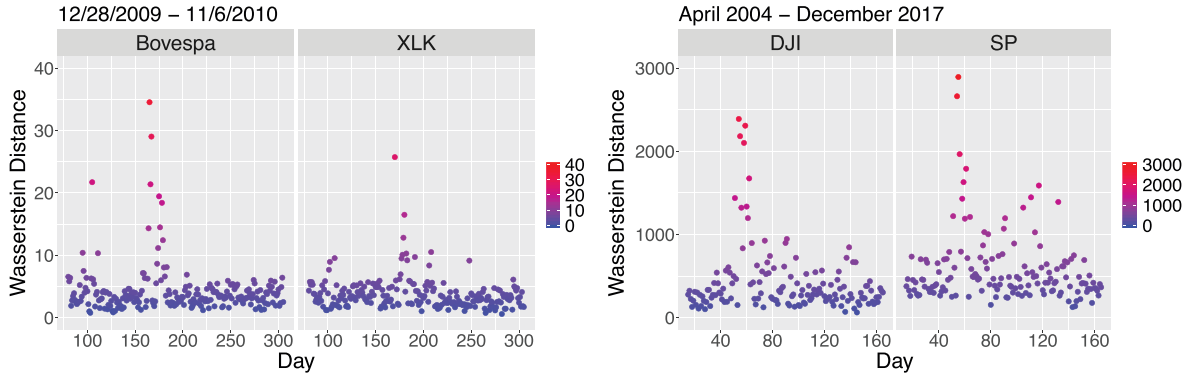
DOI: 10.1111/jtsa.12590

Table IV. Forecast accuracies of five methods, Dow–Jones cross-sectional returns

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

Table V. Forecast accuracies of five methods, S&P 500 cross-sectional returns

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



(a) Intraday Returns (b) Monthly Returns

Figure 7. Wasserstein distance between sample points to their Wasserstein mean

viability of the model. In our empirical analysis, the proposed WAR(p) model emerged as a competitive forecasting method for financial return densities when compared to various existing methods and using several different metrics for forecasting accuracy. The option of selecting the order p to suit a specific purpose is a useful feature of the model. We proposed a data-driven procedure that targets optimal prediction in terms of a specific metric, but other objectives, including a model fit in terms of information criteria could be used as well.

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

DOI: 10.1111/jtsa.12590

but truncated predictors, effectively equivalent to order selected AR(p) models, generally perform better, see for example, Section 3.5 of Shumway and Stoffer (2018).

We explored empirically the fully functional WAR(p) model, but we did not pursue its theoretical underpinnings because its predictive performance was not competitive; simpler models often provide better predictions. The theory of fully functional WAR(1) model was developed, independently and in parallel with our research, in Chen *et al.* (2020). It is also a model constructed in the tangent space and so it is subject to similar constraints as our WAR(p) models, namely that the solution must be restricted to image of the logarithmic map with probability one (see assumption (A3) in this article, and assumption (B2) in Chen *et al.* (2020)). It is unclear whether concrete examples of innovations can be established that satisfy this constraint for fully functional WAR(p) models, whereas we have established several concrete examples for WAR(p) models in this article. Still, fully functional WARMA(p, q) models might be useful in some settings, and their theory might then be developed.

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

ACKNOWLEDGEMENT

This research was partially supported by National Science Foundation grants DMS-1811888, DMS-1914882 and DMS-1923142.

DATA AVAILABILITY STATEMENT

The DJIA, S&P 500 and XLK data used in this research are publicly available from the CRSP database (Center for Research in Security Prices, crsp.org). They are available as supplementary files. The Bovespa data were provided by Capse Investimentos (capse.com.br), and can be requested from that company.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

REFERENCES

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

Bekierman J, Gribisch B. 2019. A mixed frequency stochastic volatility model for intraday stock market returns. *Journal of Financial Econometrics*, 10.1093/jjfinec/nbz021, (to appear in print).

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

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.

Bosq D. 2000. *Linear Processes in Function Spaces*. Berlin: Springer.

Brockwell P, Davis R. 1991. *Time Series: Theory and methods*. Berlin: Springer.

Brockwell PJ, Lindner A. 2010. Strictly stationary solutions of autoregressive moving average equations. *Biometrika* **97**: 765–772.

Brockwell PJ, Lindner A, Vollenbröker B. 2013. Strictly stationary solutions of multivariate ARMA equations with i.i.d. noise. *Annals of the Institute of Statistical Mathematics* **64**: 1089–1119.

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

DOI: 10.1111/jtsa.12590

Chen Y, Lin Z, Müller H-G. 2020. *Wasserstein regression*, arXiv preprint arXiv:2006.09660.
Egozcue JJ, Díaz-Barrero JL, Pawlowsky-Glahn V. 2006. Hilbert space of probability density functions based on aitchison geometry. *Acta Mathematica Sinica* **22**(4): 1175–1182.
Harvey CR, Liu Y, Zhu H. 2016. … and the cross-section of expected returns. *The Review of Financial Studies* **29**: 5–68.
Horta E, Ziegelmann F. 2018. Dynamics of financial returns densities: a functional approach applied to the Bovespa intraday index. *International Journal of Forecasting* **34**(1): 75–88.
Horváth L, Kokoszka P. 2012. *Inference for Functional Data with Applications*. Berlin: Springer.
Horváth L, Kokoszka P, Rice G. 2014. Testing stationarity of functional time series. *Journal of Econometrics* **179**: 66–82.
Hron K, Menafoglio A, Templ M, Hrůzová K, Filzmoser P. 2016. Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis* **94**: 330–350.
Klepsch J, Küppelberg C, Wei T. 2017. Prediction of functional ARMA processes with an application to traffic data. *Econometrics and Statistics* **1**: 128–149.
Kneip A, Utikal KJ. 2001. Inference for density families using functional principal component analysis. *Journal of the American Statistical Association* **96**(454): 519–542.
Kokoszka P, Miao H, Petersen A, Shang HL. 2019. Forecasting of density functions with an application to cross-sectional and intraday returns. *International Journal of Forecasting* **35**(4): 1304–1317.
Kokoszka P, Reimherr M. 2017. *Introduction to Functional Data Analysis*: Chapman and Hall/CRC.
Kullback S, Leibler R. 1951. On information and sufficiency. *The Annals of Mathematical statistics* **22**: 79–86.
Lütkepohl H. 2006. *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.
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 (A)* **178**: 187–203.
Panaretos VM, Zemel Y. 2016. Amplitude and phase variation of point processes. *The Annals of Statistics* **44**(2): 771–812.
Panaretos VM, Zemel Y. 2020. *An Invitation to Statistics in Wasserstein space*. Berlin: Springer Nature.
Pawlowsky-Glahn V, Egozcue J, Tolosana-Delgado R. 2015. *Modeling and Analysis of Compositional Data*. New York: Wiley.
Petersen A, Müller H-G. 2019. Wasserstein covariance for multiple random densities. *Biometrika* **106**: 339–351.
Petersen A, Liu X, Divani AA. 2020. Wasserstein *F*-tests and confidence bands for the Fréchet regression of density response curves. *Annals of Statistics* **49**(1): 590–611.
Petersen A, Müller H-G, et al. 2016. Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics* **44**(1): 183–218.
Salazar P, Napoli D, Mario J, Mostafa J, Alibay Z, Wendy P, Alexander M, Stephan A, Bershad EM, Damani R, Divani AA. 2019. Exploration of multiparameter hematoma 3d image analysis for predicting outcome after intracerebral hemorrhage. *Neurocritical Care*: 1–11.
Shang HL, Haberman S. 2020. Forecasting age distribution of death counts: an application to annuity pricing. *Annals of Actuarial Science* **14**: 150–169.
Shannon CE. 1948. A mathematical theory of communication. *Bell system Technical Journal* **27**(3): 379–423.
Shumway RH, Stoffer DS. 2018. *Time Series Analysis and Its Applications*. Berlin: Springer.
Spangenberg F. 2013. Strictly stationary solutions of ARMA equations in Banach spaces. *Journal of Multivariate Analysis* **121**: 127–138.
Srivastava A, Jermyn I, Joshi S. 2007. Riemannian analysis of probability density functions with applications in vision. In *2007 IEEE Conference on Computer Vision and Pattern Recognition* New York: IEEE; 1–8.
Villani C. 2003. *Topics in Optimal Transportation*: American Mathematical Society.
Wang J. 2012. *A state space model approach to functional time series and time series driven by differential equations*. Ph.D. thesis, Rutgers University-Graduate School-New Brunswick.
Yang H, Baladandayuthapani V, Rao AUK, Morris JS. 2020. Quantile function on scalar regression analysis for distributional data. *Journal of the American Statistical Association* **115**(529): 90–106.
Zhang X, Shao X. 2015. Two sample inference for the second-order property of temporally dependent functional data. *Bernoulli* **21**: 909–929.

DOI: 10.1111/jtsa.12590