Wasserstein Regression*

Yaqing Chen[1], Zhenhua Lin[2], and Hans-Georg Müller[1]

[1]Department of Statistics, University of California, Davis
[2]Department of Statistics and Applied Probability, National University of Singapore

Abstract

The analysis of samples of random objects that do not lie in a vector space is gaining increasing attention in statistics. An important class of such object data is univariate probability measures defined on the real line. Adopting the Wasserstein metric, we develop a class of regression models for such data, where random distributions serve as predictors and the responses are either also distributions or scalars. To define this regression model, we utilize the geometry of tangent bundles of the space of random measures endowed with the Wasserstein metric for mapping distributions to tangent spaces. The proposed distribution-to-distribution regression model provides an extension of multivariate linear regression for Euclidean data and function-to-function regression for Hilbert space valued data in functional data analysis. In simulations, it performs better than an alternative transformation approach where one maps distributions to a Hilbert space through the log quantile density transformation and then applies traditional functional regression. We derive asymptotic rates of convergence for the estimator of the regression operator and for predicted distributions and also study an extension to autoregressive models for distribution-valued time series. The proposed methods are illustrated with data on human mortality and distributional time series of house prices.
Keywords: Distribution regression; distributional time series; functional data analysis; parallel transport; tangent bundles.

1 Introduction

Regression analysis is one of the foundational tools of statistics to quantify the relationship between a response variable and predictors and there have been many extensions of simple models such as the multiple linear regression model to more complex data scenarios. These include linear models for function-to-function regression, where predictors and responses are both considered random elements

*This research was supported by NSF grants DMS-1712862 and DMS-2014626 and NUS Startup grant R-155-000-217-133. We wish to thank three anonymous referees, an Associate Editor, and the Editor for the helpful and constructive comments which led to numerous improvements in the paper.

in Hilbert space, with a variant where responses are scalars (**??**). Such linear functional regression models and their properties have been well studied (**?????**) and reviewed (**??**).

Samples that include random objects, which are random elements in general metric spaces that by default do not have a vector space structure, are increasingly common. Such data cannot be analyzed with methods devised for Euclidean or functional data, which are usually viewed as random elements of a Hilbert space (**??**). We focus here on the case where the random objects are random probability measures on the real line that satisfy certain regularity conditions. Specifically, at this time there are no in-depth studies with detailed statistical analysis of regression models that feature such random measures as predictors, in contrast to the situation where vector predictors are coupled with random distributions as responses (**?**).

Related work also includes a variety of methods that specifically target the case where Euclidean predictors are paired with responses that reside on a finite-dimensional Riemannian manifold (**??????**). Kernel and spline type methods have been proposed for the case where both predictors and responses are elements of finite-dimensional Riemannian manifolds (**???**). However, these methods do not cover spaces of probability measures under the Wasserstein metric, where the tangent spaces are subspaces of infinite-dimensional Hilbert spaces. Additionally, no comprehensive investigation of the statistical properties and asymptotic behavior of distribution-to-distribution regression models seems to exist. To develop the proposed model, we utilize tangent bundles in the space of probability distributions with the Wasserstein metric and parallel transport to obtain asymptotic results for regression operators and predicted measures.

A recent approach to including random distributions as predictors in complex regression models is to transform the densities of these distributions to unconstrained functions in the Hilbert space $\mathcal{L}^2$, e.g., by the log quantile density (LQD) transformation (**?**) and then to employ functional regression models where the transformed functions serve as predictors and the responses are either also the transformed functions or scalars (**???**), whence established methods for functional regression become applicable. However, the LQD transformation does not take into account the geometry of the space

of probability distributions and therefore the corresponding transformation map is not isometric and leads to deformations that change distances between pairs of objects. In contrast, the transformation method we develop here is closely adapted to the underlying geometry, leads to an isometric map and fully utilizes the geometric properties of the metric space of random measures equipped with the Wasserstein distance. We also found in implementations and simulations that the proposed geometric method that we refer to as Wasserstein regression works very well, especially when comparing it to a regression approach that is based on the LQD transformation. Other alternatives have been considered for regressing scalar responses on distribution-valued predictors (**?????**), but these are either Nadaraya–Watson type estimators that suffer from a severe curse of dimensionality, or kernel-based methods that rely on tuning parameters whose choice could be sensitive in real applications. **?** approximate input histograms by the closest weighted barycenters of a database of reference histograms with respect to Wasserstein distance, which work when input histograms are not far from the references, aiming at applications in image processing.

Our goal is to develop a regression model where the predictors and responses are both distributions in $\mathcal{W}$. A good starting point is linear regression in Euclidean spaces, where for a pair of random elements $(X, Y) \in (\mathbb{R}^p, \mathbb{R})$, $\mathbb{E}(Y \mid X) = \Gamma_{\mathrm{E}}(X) = \mathbb{E}Y + \beta^\top(X - \mathbb{E}X)$. The regression function Γ_{E} can be characterized by the following two properties: First, it maps the expectation of X to the expectation of Y; second, conditioning on X, it transports the line segment between $\mathbb{E}X$ and X to that between $\mathbb{E}Y$ and $\mathbb{E}(Y \mid X)$. Specifically,

$$\mathbb{E}Y = \Gamma_{\mathrm{E}}(\mathbb{E}X) \quad \text{and} \quad \mathbb{E}[\mathbb{E}Y + t(Y - \mathbb{E}Y) \mid X] = \Gamma_{\mathrm{E}}[\mathbb{E}X + t(X - \mathbb{E}X)], \text{ for all } t \in [0, 1]. \tag{1}$$

However, expectations and line segments are not well-defined for the space of distributions, since it is not a vector space. In this paper, we develop a distribution-to-distribution regression model that is analogous to traditional linear regression models for Euclidean and functional data, with the decisive difference that both predictors and responses are univariate probability measures. An example which we investigate later is to study the relationship of the age-at-death distributions of different countries

in 2013 to the distributions 30 years before. We also discuss an extension of our approach to an autoregressive model for distribution-valued time series. In our estimation procedures and theoretical analysis we cover the commonly encountered but more complex situation where neither predictor nor response distributions are directly observed and instead the available data consist of i.i.d. samples that are generated by each of these distributions. After we submitted this paper, a preprint reporting independently conducted but related work on autoregressive modeling of distributional time series was posted by **?**, where a simplified version of the distributional autoregressive model in (29) was studied.

The remainder of the paper is organized as follows. We first propose a distribution-to-distribution regression model based on the tangent bundle of the Wasserstein space of probability distributions in Section 2, with estimation and asymptotic theory in Section 3, and then describe an extension of the model to an autoregressive model for time series of distributions in Section 4. Simulation studies are illustrated in Section 5 to assess the finite-sample performance of the proposed estimators and a competing approach. The wide applicability of the proposed methods is demonstrated with applications to human mortality data and US house price data in Section 6.

2 Methodology

2.1 Tangent Bundle of the Wasserstein Space

Let D be $\mathbb{R}$ or a closed interval in $\mathbb{R}$, and $\mathscr{B}(D)$ be the Borel σ-algebra on D. We focus on the Wasserstein space $\mathcal{W} = \mathcal{W}(D)$ of probability distributions on $(D, \mathscr{B}(D))$ with finite second moments, endowed with the $\mathcal{L}^2$-Wasserstein distance

$$d_W(\mu_1, \mu_2) = \left\{ \int_0^1 [F_1^{-1}(p) - F_2^{-1}(p)]^2 \mathrm{d}p \right\}^{1/2}, \tag{2}$$

for $\mu_1, \mu_2 \in \mathcal{W}$, where F_1^{-1} and F_2^{-1} denote the quantile functions of μ_1 and μ_2, respectively; specifically, for any distribution $\mu = \mu(F) \in \mathcal{W}$ with cumulative distribution function (cdf) F, we consider the

quantile function F^{-1} to be the left continuous inverse of F, i.e.,

$$F^{-1}(p) = \inf\{r \in D : F(r) \geq p\}, \quad \text{for } p \in (0,1). \tag{3}$$

As demonstrated for example in **???**, basic concepts of Riemannian manifolds can be generalized to the Wasserstein space $\mathcal{W}$. We assume in the following that $\mu_* \in \mathcal{W}$ is an atomless reference probability measure, i.e., it possesses a continuous cdf F_*. For any $\mu \in \mathcal{W}$, the geodesic from μ_* to μ, $\gamma_{\mu_*,\mu}\colon [0,1] \to \mathcal{W}$, is given by

$$\gamma_{\mu_*,\mu}(t) = [t(F^{-1} \circ F_* - \mathrm{id}) + \mathrm{id}]\#\mu_*, \quad \text{for } t \in [0,1], \tag{4}$$

where for a measurable function $h\colon D \to D$, $h\#\mu_*$ is a push-forward measure such that $h\#\mu_*(A) = \mu_*(\{r \in D : h(r) \in A\})$ for any set $A \in \mathscr{B}(D)$. The tangent space at μ_* is defined as

$$T_{\mu_*} = \overline{\{t(F^{-1} \circ F_* - \mathrm{id}) : \mu = \mu(F) \in \mathcal{W},\, t > 0\}}^{\mathcal{L}^2_{\mu_*}},$$

where $\mathcal{L}^2_{\mu_*} = \mathcal{L}^2_{\mu_*}(D)$ is the Hilbert space of μ_*-square-integrable functions on $D \subset \mathbb{R}$, with inner product $\langle \cdot, \cdot \rangle_{\mu_*}$ and norm $\| \cdot \|_{\mu_*}$. The tangent space T_{μ_*} is a subspace of $\mathcal{L}^2_{\mu_*}$ equipped with the same inner product and induced norm (Theorem 8.5.1, **?**).

The exponential map Exp_{μ_*} is then defined by the push-forward measures, which maps functions of the form $g = t(F^{-1} \circ F_* - \mathrm{id})$ onto $\mathcal{W}$, with F^{-1} being the quantile function of an arbitrary distribution $\mu \in \mathcal{W}$,

$$\mathrm{Exp}_{\mu_*} g = (g + \mathrm{id})\#\mu_*. \tag{5}$$

While this exponential map is not a local homeomorphism (**?**), any $\mu \in \mathcal{W}$ can be recovered by $\mathrm{Exp}_{\mu_*}(F^{-1} \circ F_* - \mathrm{id})$ in the sense that $d_W(\mathrm{Exp}_{\mu_*}(F^{-1} \circ F_* - \mathrm{id}), \mu) = 0$, and the logarithmic map

$\mathrm{Log}_{\mu_*}\colon \mathcal{W} \to T_{\mu_*}$, as the right inverse of the exponential map, is given by

$$\mathrm{Log}_{\mu_*}\mu = F^{-1} \circ F_* - \mathrm{id}, \quad \text{for } \mu \in \mathcal{W}. \tag{6}$$

Furthermore, restricted to the log image, $\mathrm{Exp}_{\mu_*}|_{\mathrm{Log}_{\mu_*}(\mathcal{W})}$ is an isometric homeomorphism (e.g., Lemma 2.1, **?**).

2.2 Distribution-to-Distribution Regression

Let (ν_1, ν_2) be a pair of random elements with a joint distribution $\mathcal{F}$ on $\mathcal{W} \times \mathcal{W}$, assumed to be square integrable in the sense that $\mathbb{E}d_W^2(\mu, \nu_1) < \infty$ and $\mathbb{E}d_W^2(\mu, \nu_2) < \infty$ for some (and thus for all) $\mu \in \mathcal{W}$. Any element in $\mathcal{W}$ that minimizes $\mathbb{E}d_W^2(\cdot, \nu_1)$ is called a Fréchet mean of ν_1 (**?**). Since the Wasserstein space $\mathcal{W}$ is a Hadamard space (**?**), such minimizers uniquely exist (**?**) and are given by

$$\nu_{1\oplus} = \operatorname*{argmin}_{\mu \in \mathcal{W}} \mathbb{E}d_W^2(\mu, \nu_1) \quad \text{and} \quad \nu_{2\oplus} = \operatorname*{argmin}_{\mu \in \mathcal{W}} \mathbb{E}d_W^2(\mu, \nu_2). \tag{7}$$

It is well-known that for univariate distributions as we consider here, the quantile functions of the Fréchet means are simply

$$F_{1\oplus}^{-1}(\cdot) = \mathbb{E}F_1^{-1}(\cdot) \quad \text{and} \quad F_{2\oplus}^{-1}(\cdot) = \mathbb{E}F_2^{-1}(\cdot),$$

where $F_{1\oplus}^{-1}, F_{2\oplus}^{-1}, F_1^{-1}$ and F_2^{-1} are the quantile functions of $\nu_{1\oplus}, \nu_{2\oplus}, \nu_1$ and ν_2, respectively.

As suggested by the multiple linear regression as per (1), we replace expectations and line segments, which are not well-defined for the Wasserstein space, by Fréchet means and geodesics, respectively. Hence, a regression operator $\Gamma_{\mathrm{W}}\colon \mathcal{W} \to \mathcal{W}$ for the Wasserstein space would be expected to satisfy:

$$d_W(\nu_{2\oplus}, \Gamma_{\mathrm{W}}(\nu_{1\oplus})) = 0 \quad \text{and} \quad d_W(\mathbb{E}_\oplus\{\gamma_{\nu_{2\oplus},\nu_2}(t)|\nu_1\}, \Gamma_{\mathrm{W}}\{\gamma_{\nu_{1\oplus},\nu_1}(t)\}) = 0, \text{ for all } t \in [0,1], \tag{8}$$

where the conditional Fréchet mean $\mathbb{E}_\oplus\{\gamma_{\nu_{2\oplus},\nu_2}(t) \mid \nu_1\} \coloneqq \mathrm{argmin}_{\mu \in \mathcal{W}} \mathbb{E}[d_W^2(\mu, \gamma_{\nu_{2\oplus},\nu_2}(t)) \mid \nu_1]$.

We assume that the Fréchet means $\nu_{1\oplus}$ and $\nu_{2\oplus}$ are atomless so that they can be used as the reference probability measures as in Section 2.1. Note that $\mathrm{Log}_{\nu_{1\oplus}}\nu_{1\oplus} = 0$, $\nu_{1\oplus}$-a.e., and $\mathrm{Log}_{\nu_{2\oplus}}\nu_{2\oplus} = 0$, $\nu_{2\oplus}$-a.e., and that $\mathrm{Exp}_{\mu}(0) = \mu$ for any $\mu \in \mathcal{W}$. Furthermore, it follows from (4)–(6) and the isometry property of $\mathrm{Exp}_{\nu_{2\oplus}}|_{\mathrm{Log}_{\nu_{2\oplus}}\mathcal{W}}$ that $\gamma_{\nu_{1\oplus},\nu_1}(t) = \mathrm{Exp}_{\nu_{1\oplus}}(t\mathrm{Log}_{\nu_{1\oplus}}\nu_1)$ and that $\mathbb{E}_{\oplus}\{\gamma_{\nu_{2\oplus},\nu_2}(t) \mid \nu_1\} = \mathrm{argmin}_{\mu\in\mathcal{W}}\mathbb{E}(\|\mathrm{Log}_{\nu_{2\oplus}}\mu - t\mathrm{Log}_{\nu_{2\oplus}}\nu_2\|^2_{\nu_{2\oplus}} \mid \mathrm{Log}_{\nu_{1\oplus}}\nu_1) = \mathrm{Exp}_{\nu_{2\oplus}}[\mathbb{E}(t\mathrm{Log}_{\nu_{2\oplus}}\nu_2 \mid \mathrm{Log}_{\nu_{1\oplus}}\nu_1)]$. Hence, (8) can be rewritten as

$$\|\Gamma(0)\|_{\nu_{2\oplus}} = 0 \quad \text{and} \quad \|\mathbb{E}(t\mathrm{Log}_{\nu_{2\oplus}}\nu_2 \mid \mathrm{Log}_{\nu_{1\oplus}}\nu_1) - \Gamma(t\mathrm{Log}_{\nu_{1\oplus}}\nu_1)\|_{\nu_{2\oplus}} = 0, \text{ for all } t \in [0,1], \tag{9}$$

where $\Gamma\colon T_{\nu_{1\oplus}} \to T_{\nu_{2\oplus}}$, $\Gamma = \mathrm{Log}_{\nu_{2\oplus}} \circ \Gamma_{\mathrm{W}} \circ \mathrm{Exp}_{\nu_{1\oplus}}$, is a regression operator between tangent spaces $T_{\nu_{1\oplus}}$ and $T_{\nu_{2\oplus}}$.

As discussed in Section 2.1, $T_{\nu_{1\oplus}}$ and $T_{\nu_{2\oplus}}$ are subspaces of $\mathcal{L}^2_{\nu_{1\oplus}}$ and $\mathcal{L}^2_{\nu_{2\oplus}}$, respectively. Distribution-to-distribution regression can then be viewed as function-to-function regression, which has been well-studied in functional data analysis (see, e.g., **????**). Specifically, we assume that the random pair of distributions (ν_1, ν_2) satisfy the model

$$\mathbb{E}(\mathrm{Log}_{\nu_{2\oplus}}\nu_2 \mid \mathrm{Log}_{\nu_{1\oplus}}\nu_1) = \Gamma(\mathrm{Log}_{\nu_{1\oplus}}\nu_1), \tag{10}$$

where $\Gamma\colon T_{\nu_{1\oplus}} \to T_{\nu_{2\oplus}}$ is a linear operator defined as

$$\Gamma g(t) = \langle \beta(\cdot, t), g\rangle_{\nu_{1\oplus}}, \quad \text{for } t \in D \text{ and } g \in T_{\nu_{1\oplus}}. \tag{11}$$

Here, $\beta : D^2 \to \mathbb{R}$ is a coefficient function (i.e., the kernel of Γ) lying in $\mathcal{L}^2_{\nu_{1\oplus}\times\nu_{2\oplus}}$, and $\nu_{1\oplus} \times \nu_{2\oplus}$ is a product probability measure on the product measurable space $(D^2, \mathscr{B}(D^2))$ generated by $\nu_{1\oplus}$ and $\nu_{2\oplus}$. We note that our model satisfies (9). Furthermore, we assume

(A1) With probability 1, $\Gamma(\mathrm{Log}_{\nu_{1\oplus}}\nu_1) + \mathrm{id}$ is non-decreasing.

Assumption (A1) guarantees that $\Gamma(\mathrm{Log}_{\nu_{1\oplus}}\nu_1) \in \mathrm{Log}_{\nu_{2\oplus}}\mathcal{W}$ with probability 1. We demonstrate the

feasibility of the proposed model in (10) by providing a framework in Section 5 to construct explicit examples that satisfy the model requirements and (A1).

2.3 Covariance Structure, Regression Coefficient Function and Scalar Responses

Noting that $\mathbb{E}(\mathrm{Log}_{\nu_{1\oplus}}\nu_1) = 0$, $\nu_{1\oplus}$-a.e., and $\mathbb{E}(\mathrm{Log}_{\nu_{2\oplus}}\nu_2) = 0$, $\nu_{2\oplus}$-a.e., we denote the covariance operators of $\mathrm{Log}_{\nu_{1\oplus}}\nu_1$ and $\mathrm{Log}_{\nu_{2\oplus}}\nu_2$ by $\mathcal{C}_{\nu_1} = \mathbb{E}(\mathrm{Log}_{\nu_{1\oplus}}\nu_1 \otimes \mathrm{Log}_{\nu_{1\oplus}}\nu_1)$ and $\mathcal{C}_{\nu_2} = \mathbb{E}(\mathrm{Log}_{\nu_{2\oplus}}\nu_2 \otimes \mathrm{Log}_{\nu_{2\oplus}}\nu_2)$, respectively, and the cross-covariance operator by $\mathcal{C}_{\nu_1\nu_2} = \mathbb{E}(\mathrm{Log}_{\nu_{2\oplus}}\nu_2 \otimes \mathrm{Log}_{\nu_{1\oplus}}\nu_1)$. Since the two covariance operators $\mathcal{C}_{\nu_1}$ and $\mathcal{C}_{\nu_2}$ are trace-class, they have eigendecompositions (Theorem 7.2.6, **?**) as given below, which can be viewed as an analog to multivariate principal component analysis (**??**), yielding a corresponding decomposition for the cross-covariance operator $\mathcal{C}_{\nu_1\nu_2}$,

$$\mathcal{C}_{\nu_1} = \sum_{j=1}^{\infty} \lambda_j \phi_j \otimes \phi_j, \quad \mathcal{C}_{\nu_2} = \sum_{k=1}^{\infty} \varsigma_k \psi_k \otimes \psi_k, \quad \mathcal{C}_{\nu_1\nu_2} = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \xi_{jk} \psi_k \otimes \phi_j. \tag{12}$$

Here $\lambda_j = \mathbb{E}[\langle \mathrm{Log}_{\nu_{1\oplus}}\nu_1, \phi_j\rangle^2_{\nu_{1\oplus}}]$ and $\varsigma_k = \mathbb{E}[\langle \mathrm{Log}_{\nu_{2\oplus}}\nu_2, \psi_k\rangle^2_{\nu_{2\oplus}}]$ are eigenvalues such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and $\varsigma_1 \geq \varsigma_2 \geq \cdots \geq 0$, $\{\phi_j\}_{j=1}^{\infty}$ and $\{\psi_k\}_{k=1}^{\infty}$ are eigenfunctions that are orthonormal in $T_{\nu_{1\oplus}}$ and $T_{\nu_{2\oplus}}$, respectively, and $\xi_{jk} = \mathbb{E}[\langle \mathrm{Log}_{\nu_{1\oplus}}\nu_1, \phi_j\rangle_{\nu_{1\oplus}} \langle \mathrm{Log}_{\nu_{2\oplus}}\nu_2, \psi_k\rangle_{\nu_{2\oplus}}]$. With probability 1, the log transformations $\mathrm{Log}_{\nu_{1\oplus}}\nu_1$ and $\mathrm{Log}_{\nu_{2\oplus}}\nu_2$ admit the Karhunen–Loève expansions

$$\mathrm{Log}_{\nu_{1\oplus}}\nu_1 = \sum_{j=1}^{\infty} \langle \mathrm{Log}_{\nu_{1\oplus}}\nu_1, \phi_j\rangle_{\nu_{1\oplus}} \phi_j \quad \text{and} \quad \mathrm{Log}_{\nu_{2\oplus}}\nu_2 = \sum_{k=1}^{\infty} \langle \mathrm{Log}_{\nu_{2\oplus}}\nu_2, \psi_k\rangle_{\nu_{2\oplus}} \psi_k.$$

Then as in the classical functional regression (e.g., **???**), the regression coefficient function β can be expressed as

$$\beta = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} b_{jk} \psi_k \otimes \phi_j, \tag{13}$$

with $b_{jk} = \lambda_j^{-1}\xi_{jk}$. In order to guarantee that the right hand side of (13) converges in the sense that

$$\lim_{J,K\to\infty} \int_D \int_D \left[\sum_{k=1}^{K}\sum_{j=1}^{J} b_{jk}\phi_j(s)\psi_k(t) - \beta(s,t)\right]^2 \mathrm{d}\nu_{1\oplus}(s)\mathrm{d}\nu_{2\oplus}(t) = 0,$$

we assume (Lemma A.2, **?**)

$$\sum_{k=1}^{\infty}\sum_{j=1}^{\infty}\lambda_j^{-2}\xi_{jk}^2 < \infty. \tag{14}$$

To keep notations simple, we use the same notation $g_1 \otimes g_2$ for the operator and its kernel throughout this paper. Namely, for $g_1 \in \mathcal{L}^2_{\mu_1}$ and $g_2 \in \mathcal{L}^2_{\mu_2}$, $g_1 \otimes g_2$ can represent either an operator on $\mathcal{L}^2_{\mu_2}$ such that $(g_1 \otimes g_2)(g) = \langle g_2, g\rangle_{\mu_2} g_1$ for $g \in \mathcal{L}^2_{\mu_2}$ or its kernel, i.e., a bivariate function such that $(g_1 \otimes g_2)(s,t) = g_1(t)g_2(s)$ for all $s, t \in D$.

A variant of the proposed distribution-to-distribution regression in (10) is the pairing of distributions as predictors with scalar responses. For a pair of random elements (ν_1, Y) with a joint distribution on $\mathcal{W} \times \mathbb{R}$, a distribution-to-scalar regression model is

$$\mathbb{E}(Y \mid \mathrm{Log}_{\nu_{1\oplus}}\nu_1) = \mathbb{E}(Y) + \langle \beta_1, \mathrm{Log}_{\nu_{1\oplus}}\nu_1\rangle_{\nu_{1\oplus}}. \tag{15}$$

Here, $\nu_{1\oplus}$ is the Fréchet mean of ν_1 and $\beta_1 \colon D \to \mathbb{R}$ is a regression coefficient function in $\mathcal{L}^2_{\nu_{1\oplus}}$ which can be expressed as $\beta_1 = \sum_{j=1}^{\infty}\lambda_j^{-1}\langle \mathbb{E}(Y\mathrm{Log}_{\nu_{1\oplus}}\nu_1), \phi_j\rangle_{\nu_{1\oplus}}\phi_j$, where λ_j and ϕ_j are the eigenvalues and eigenfunctions of the covariance operator $\mathcal{C}_{\nu_1}$ of $\mathrm{Log}_{\nu_{1\oplus}}\nu_1$ as in (12), and we assume that $\sum_{j=1}^{\infty}\lambda_j^{-2}\langle \mathbb{E}(Y\mathrm{Log}_{\nu_{1\oplus}}\nu_1), \phi_j\rangle^2_{\nu_{1\oplus}} < \infty$. This model can also be viewed as function-to-scalar regression, which has been well studied in functional data analysis (**?????**).

3 Estimation

3.1 Distribution Estimation

While **?** assume distributions are fully observed, in reality this is usually not the case, and this creates an additional challenge for the implementation of the proposed distribution-to-distribution regression model. Options to address this include estimating cdfs (e.g., **????**), or estimating quantile functions (e.g., **????**) of the underlying distributions. Given an estimated quantile function $\widehat{F}^{-1}$ (resp. cdf $\widehat{F}$),

we convert it to a cdf (resp. a quantile function) by right (resp. left) continuous inversion,

$$\widehat{F}(r) = \sup\{p \in [0,1] : \widehat{F}^{-1}(p) \leq r\}, \quad \text{for } r \in \mathbb{R} \tag{16}$$

(resp. (3)). Alternatively, one can start with a density estimator to estimate densities (**??**) and then compute the cdfs and quantile functions by integration and inversion.

Suppose $\{(\nu_{1i}, \nu_{2i})\}_{i=1}^{n}$ are n independent realizations of (ν_1, ν_2). What we observe are collections of independent measurements $\{X_{il}\}_{l=1}^{m_{\nu_{1i}}}$ and $\{Y_{il}\}_{l=1}^{m_{\nu_{2i}}}$, sampled from ν_{1i} and ν_{2i}, respectively, where $m_{\nu_{1i}}$ and $m_{\nu_{2i}}$ are the sample sizes which may vary across distributions. Note that there are two independent layers of randomness in the data: The first generates independent pairs of distributions (ν_{1i}, ν_{2i}); the second generates independent observations according to each distribution, $X_{il} \sim \nu_{1i}$ and $Y_{il} \sim \nu_{2i}$.

For a distribution $\mu \in \mathcal{W}$, denote by $\widehat{\mu} = \mu(\widehat{F})$ the distribution associated with some cdf estimate $\widehat{F}$, based on a sample of measurements drawn according to μ. Using $\widehat{\nu}_{1i}$ and $\widehat{\nu}_{2i}$ as surrogates of ν_{1i} and ν_{2i}, the theoretical analysis of the estimation of the distribution-to-distribution regression operator requires the following assumptions that quantify the discrepancy of the estimated and true probability measures.

(A2) For any distribution $\mu \in \mathcal{W}$, with some nonnegative decreasing sequences $\tau_m = o(1)$ as $m \to \infty$, the corresponding estimate $\widehat{\mu}$ based on a sample of size m drawn according to μ satisfies

$$\sup_{\mu \in \mathcal{W}} \mathbb{E}[d_W^2(\widehat{\mu}, \mu)] = O(\tau_m) \quad \text{and} \quad \sup_{\mu \in \mathcal{W}} \mathbb{E}[d_W^4(\widehat{\mu}, \mu)] = O(\tau_m^2).$$

For example, for compactly supported distributions, the distribution estimator proposed by **?** satisfies (A2) with $\tau_m = m^{-1/2}$, while **?** consider a subset $\mathcal{W}_R^{\text{ac}}$ of $\mathcal{W}$ containing distributions that are absolutely continuous with respect to Lebesgue measure on a compact domain D such that

$$\sup_{\mu \in \mathcal{W}_R^{\text{ac}}} \sup_{r \in D_\mu} \max\{f_\mu(r), 1/f_\mu(r), |f'_\mu(r)|\} \leq R, \tag{17}$$

where f_μ is the density function of a distribution $\mu \in \mathcal{W}_R^{\mathrm{ac}}$, D_μ is the support of distribution μ and $R > 0$ is constant, and then obtain the rates $\sup_{\mu \in \mathcal{W}_R^{\mathrm{ac}}} \mathbb{E} d_W^2(\widehat{\mu}, \mu) = O(m^{-2/3})$ and $\sup_{\mu \in \mathcal{W}_R^{\mathrm{ac}}} \mathbb{E}[d_W^4(\widehat{\mu}, \mu)] = O(m^{-4/3})$ in (A2) (Proposition 1, **?**).

The following assumption on the numbers of measurements per distribution $m_{\nu_{1i}}$ and $m_{\nu_{2i}}$ facilitates our analysis:

(A3) There exists a sequence $m = m(n)$ such that $\min\{m_{\nu_{1i}}, m_{\nu_{2i}} : i = 1, \ldots, n\} \geq m$ and $m \to \infty$ as $n \to \infty$.

3.2 Regression Operator Estimation

We note that notations with "~" refer to estimators based on fully observed distributions, while those with "^" refer to estimators for which the distributions, ν_{1i} and ν_{2i}, are not fully observed and only samples of measurements drawn from the distributions are available.

Given independent realizations $\{(\nu_{1i}, \nu_{2i})\}_{i=1}^n$ of (ν_1, ν_2), we first consider an oracle estimator for the regression operator Γ, where we initially assume that $\{(\nu_{1i}, \nu_{2i})\}_{i=1}^n$ are fully observed. First of all, the empirical Fréchet means are well-defined and unique due to the fact that we work in Hadamard spaces. Specifically, replacing the expectation in (7) by that with respect to the empirical measure based on $\{(\nu_{1i}, \nu_{2i})\}_{i=1}^n$ gives

$$\widetilde{\nu}_{1\oplus} = \operatorname*{arg\,min}_{\mu \in \mathcal{W}} \sum_{i=1}^n d_W^2(\nu_{1i}, \mu) \quad \text{and} \quad \widetilde{\nu}_{2\oplus} = \operatorname*{arg\,min}_{\mu \in \mathcal{W}} \sum_{i=1}^\infty d_W^2(\nu_{2i}, \mu), \tag{18}$$

where the corresponding quantile functions are the empirical means of quantile functions across the sample,

$$\widetilde{F}_{1\oplus}^{-1}(\cdot) = \frac{1}{n} \sum_{i=1}^n F_{1i}^{-1}(\cdot) \quad \text{and} \quad \widetilde{F}_{2\oplus}^{-1}(\cdot) = \frac{1}{n} \sum_{i=1}^n F_{2i}^{-1}(\cdot), \tag{19}$$

and the corresponding distribution functions are given by right continuous inverses of the quantile functions as in (16). Then the log transforms $\mathrm{Log}_{\nu_{1\oplus}} \nu_{1i}$ and $\mathrm{Log}_{\nu_{2\oplus}} \nu_{2i}$ admit estimates $\mathrm{Log}_{\widetilde{\nu}_{1\oplus}} \nu_{1i}$ and $\mathrm{Log}_{\widetilde{\nu}_{2\oplus}} \nu_{2i}$. The covariance operators $\mathcal{C}_{\nu_1}$ and $\mathcal{C}_{\nu_2}$ can be estimated by $\widetilde{\mathcal{C}}_{\nu_1} = n^{-1} \sum_{i=1}^n \mathrm{Log}_{\widetilde{\nu}_{1\oplus}} \nu_{1i} \otimes$

$\mathrm{Log}_{\widetilde{\nu}_{1\oplus}}\nu_{1i}$ and $\widetilde{\mathcal{C}}_{\nu_2} = n^{-1}\sum_{i=1}^n \mathrm{Log}_{\widetilde{\nu}_{2\oplus}}\nu_{2i} \otimes \mathrm{Log}_{\widetilde{\nu}_{2\oplus}}\nu_{2i}$. We denote the eigenvalues and eigenfunctions of $\widetilde{\mathcal{C}}_{\nu_1}$ and $\widetilde{\mathcal{C}}_{\nu_2}$ by $\widetilde{\lambda}_j$ and $\widetilde{\phi}_j$, respectively by $\widetilde{\varsigma}_k$ and $\widetilde{\psi}_k$, where the eigenvalues are in non-ascending order. The cross-covariance operator $\mathcal{C}_{\nu_1\nu_2}$ can be estimated by $\widetilde{\mathcal{C}}_{\nu_1\nu_2} = n^{-1}\sum_{i=1}^n \mathrm{Log}_{\widetilde{\nu}_{2\oplus}}\nu_{2i} \otimes \mathrm{Log}_{\widetilde{\nu}_{1\oplus}}\nu_{1i}$.

Due to the compactness of $\mathcal{C}_{\nu_1}$, its inverse is not bounded, leading to an ill-posed problem (e.g., **??**). Regularization is thus needed and can be achieved through truncation. Oracle estimators for the regression coefficient function β and regression operator Γ are

$$\widetilde{\beta} = \sum_{k=1}^{K}\sum_{j=1}^{J} \widetilde{b}_{jk}\widetilde{\psi}_k \otimes \widetilde{\phi}_j \quad \text{and} \quad \widetilde{\Gamma}g(t) = \langle g, \widetilde{\beta}(\cdot,t)\rangle_{\widetilde{\nu}_{1\oplus}}, \text{ for } g \in \mathrm{Log}_{\widetilde{\nu}_{1\oplus}}\mathcal{W},\ t \in D, \tag{20}$$

where $\widetilde{b}_{jk} = \widetilde{\lambda}_j^{-1}\widetilde{\xi}_{jk}$, with $\widetilde{\xi}_{jk} = n^{-1}\sum_{i=1}^n \langle \mathrm{Log}_{\widetilde{\nu}_{1\oplus}}\nu_{1i}, \widetilde{\phi}_j\rangle_{\widetilde{\nu}_{1\oplus}} \langle \mathrm{Log}_{\widetilde{\nu}_{2\oplus}}\nu_{2i}, \widetilde{\psi}_k\rangle_{\widetilde{\nu}_{2\oplus}}$, and J and K are the truncation bounds, i.e., the numbers of included eigenfunctions.

Furthermore, we can construct an estimator based on the distribution estimation in Section 3.1 which will be applicable in practical situations, where typically ν_{1i} and ν_{2i} are observed in the form of samples generated from ν_{1i} and ν_{2i}. Denote the estimated quantile functions by $\widehat{F}_{1i}^{-1}$ and $\widehat{F}_{2i}^{-1}$, respectively. Then the quantile functions of the empirical Fréchet means $\widehat{\nu}_{1\oplus}$ and $\widehat{\nu}_{2\oplus}$ of $\widehat{\nu}_{1i}$ and $\widehat{\nu}_{2i}$ for $i = 1, \ldots, n$ are given by

$$\widehat{F}_{1\oplus}^{-1}(\cdot) = \frac{1}{n}\sum_{i=1}^n \widehat{F}_{1i}^{-1}(\cdot) \quad \text{and} \quad \widehat{F}_{2\oplus}^{-1}(\cdot) = \frac{1}{n}\sum_{i=1}^n \widehat{F}_{2i}^{-1}(\cdot), \tag{21}$$

and the corresponding distribution functions $\widehat{F}_{1\oplus}$ and $\widehat{F}_{2\oplus}$ can be obtained by right continuous inversion as per (16). Replacing ν_{1i} and ν_{2i} by the corresponding estimates $\widehat{\nu}_{1i}$ and $\widehat{\nu}_{2i}$, we can analogously obtain the estimates for the covariance operators, $\widehat{\mathcal{C}}_{\nu_1} = n^{-1}\sum_{i=1}^n \mathrm{Log}_{\widehat{\nu}_{1\oplus}}\widehat{\nu}_{1i} \otimes \mathrm{Log}_{\widehat{\nu}_{1\oplus}}\widehat{\nu}_{1i}$ and $\widehat{\mathcal{C}}_{\nu_2} = n^{-1}\sum_{i=1}^n \mathrm{Log}_{\widehat{\nu}_{2\oplus}}\widehat{\nu}_{2i} \otimes \mathrm{Log}_{\widehat{\nu}_{2\oplus}}\widehat{\nu}_{2i}$, as well as the estimate for the cross-covariance operator, $\widehat{\mathcal{C}}_{\nu_1\nu_2} = n^{-1}\sum_{i=1}^n \mathrm{Log}_{\widehat{\nu}_{2\oplus}}\widehat{\nu}_{2i} \otimes \mathrm{Log}_{\widehat{\nu}_{1\oplus}}\widehat{\nu}_{1i}$. We denote the eigenvalues and eigenfunctions of $\widehat{\mathcal{C}}_{\nu_1}$ and $\widehat{\mathcal{C}}_{\nu_2}$ by $\widehat{\lambda}_j$ and $\widehat{\phi}_j$, respectively by $\widehat{\varsigma}_k$ and $\widehat{\psi}_k$, where the eigenvalues are in non-ascending order. Data-based estimators

of the regression coefficient function β and regression operator Γ in (11) are then

$$\widehat{\beta} = \sum_{k=1}^{K} \sum_{j=1}^{J} \widehat{b}_{jk} \widehat{\psi}_k \otimes \widehat{\phi}_j, \quad \text{and} \quad \widehat{\Gamma} g(t) = \langle g, \widehat{\beta}(\cdot, t) \rangle_{\widehat{\nu}_{1\oplus}}, \text{ for } g \in \mathrm{Log}_{\widehat{\nu}_{1\oplus}} \mathcal{W}, \, t \in D, \tag{22}$$

where $\widehat{b}_{jk} = \widehat{\lambda}_j^{-1} \widehat{\xi}_{jk}$, and $\widehat{\xi}_{jk} = n^{-1} \sum_{i=1}^{n} \langle \mathrm{Log}_{\widehat{\nu}_{1\oplus}} \widehat{\nu}_{1i}, \widehat{\phi}_j \rangle_{\widehat{\nu}_{1\oplus}} \langle \mathrm{Log}_{\widehat{\nu}_{2\oplus}} \widehat{\nu}_{2i}, \widehat{\psi}_k \rangle_{\widehat{\nu}_{2\oplus}}$.

Regarding the numbers of eigenfunctions included, J and K, we note that larger values of J and K lead to smaller bias but larger variance and potential overfitting. We discuss the selection of J and K further in Section S.4.1 in the Supplementary Material.

While this paper focuses on univariate distributions, we note that the proposed method in principle can be extended to the multivariate setting, where however the optimal maps and hence the log maps in general do not have closed-form expressions and the estimation is completely different from the univariate setting. In addition, the required determination of the optimal transport maps is fraught with numerical difficulties (**?**). This is in contrast to the univariate case, where optimal transports just require the computation of quantile functions. Furthermore, the corresponding asymptotic analysis is also different from the univariate setting; in particular, the expression of the parallel transport does not hold in the multivariate case. See Section S.7 in the Supplementary Material for further discussion.

3.3 Parallel Transport

Note that the true regression operator, $\Gamma \colon T_{\nu_{1\oplus}} \to T_{\nu_{2\oplus}}$, and its estimators, $\widetilde{\Gamma} \colon T_{\widetilde{\nu}_{1\oplus}} \to T_{\widetilde{\nu}_{2\oplus}}$ and $\widehat{\Gamma} \colon T_{\widehat{\nu}_{1\oplus}} \to T_{\widehat{\nu}_{2\oplus}}$, are defined on different tangent spaces, which makes their comparison not so straightforward. For this, we employ parallel transport, which is a commonly used tool for data on manifolds (**???**). For two probability distributions $\mu_1, \mu_2 \in \mathcal{W}$, a parallel transport operator $\mathrm{P}_{\mu_1,\mu_2} \colon \mathcal{L}^2_{\mu_1} \to \mathcal{L}^2_{\mu_2}$ can be defined between the entire Hilbert spaces $\mathcal{L}^2_{\mu_1}$ and $\mathcal{L}^2_{\mu_2}$ by

$$\mathrm{P}_{\mu_1,\mu_2} g \coloneqq g \circ F_1^{-1} \circ F_2, \quad \text{for } g \in \mathcal{L}^2_{\mu_1}, \tag{23}$$

where F_1^{-1} and F_2 are the quantile function of μ_1 and cdf of μ_2, respectively. Assuming that μ_1 is atomless, restricted to the tangent space T_{μ_1}, the parallel transport operator $\mathrm{P}_{\mu_1,\mu_2}|_{T_{\mu_1}}$ defines the parallel transport from tangent space T_{μ_1} to T_{μ_2}.

Denote by $\mathcal{H}_{\mu_1,\mu_2}$ the space of all Hilbert–Schmidt operators from T_{μ_1} to T_{μ_2}, for $\mu_1, \mu_2 \in \mathcal{W}$. With $\mu_1, \mu_2, \mu_1', \mu_2' \in \mathcal{W}$ where μ_1' and μ_2 are atomless, we can define a parallel transport operator $\mathcal{P}_{(\mu_1,\mu_2),(\mu_1',\mu_2')}$ from $\mathcal{H}_{\mu_1,\mu_2}$ to $\mathcal{H}_{\mu_1',\mu_2'}$ by

$$(\mathcal{P}_{(\mu_1,\mu_2),(\mu_1',\mu_2')}\mathcal{A})g = \mathrm{P}_{\mu_2,\mu_2'}(\mathcal{A}(\mathrm{P}_{\mu_1',\mu_1}g)), \quad \text{for } g \in T_{\mu_1'} \text{ and } \mathcal{A} \in \mathcal{H}_{\mu_1,\mu_2}. \tag{24}$$

Denoting the Hilbert–Schmidt norm on $\mathcal{H}_{\mu_1,\mu_2}$ by $\|\cdot\|_{\mathcal{H}_{\mu_1,\mu_2}}$, for $\mu_1, \mu_2 \in \mathcal{W}$, properties of parallel transport operators P_{μ_1,μ_2} and $\mathcal{P}_{(\mu_1,\mu_2),(\mu_1',\mu_2')}$ that are relevant for the theory are listed in Proposition S1 in Section S.1.1 in the Supplementary Material. Given atomless distributions $\mu_1, \mu_2, \mu_1', \mu_2' \in \mathcal{W}$, applying Proposition S1 the discrepancy between operators $\mathcal{A} \in \mathcal{H}_{\mu_1,\mu_2}$ and $\mathcal{A}' \in \mathcal{H}_{\mu_1',\mu_2'}$ can be quantified in the space $\mathcal{H}_{\mu_1,\mu_2}$ by $\|\mathcal{P}_{(\mu_1',\mu_2'),(\mu_1,\mu_2)}\mathcal{A}' - \mathcal{A}\|_{\mathcal{H}_{\mu_1,\mu_2}}$.

3.4 Asymptotic Theory

Our goal for the theory is to evaluate the performance of the estimated regression operators, $\widetilde{\Gamma}$ and $\widehat{\Gamma}$ as per (20) and (22), respectively. According to the discussion in Section 3.3, if the true Fréchet means $\nu_{1\oplus}$ and $\nu_{2\oplus}$ and their estimators are atomless, the discrepancy between the estimated and true regression operators can be gauged by $\|\mathcal{P}_{(\widetilde{\nu}_{1\oplus},\widetilde{\nu}_{2\oplus}),(\nu_{1\oplus},\nu_{2\oplus})}\widetilde{\Gamma} - \Gamma\|_{\mathcal{H}_{\nu_{1\oplus},\nu_{2\oplus}}}$ and $\|\mathcal{P}_{(\widehat{\nu}_{1\oplus},\widehat{\nu}_{2\oplus}),(\nu_{1\oplus},\nu_{2\oplus})}\widehat{\Gamma} - \Gamma\|_{\mathcal{H}_{\nu_{1\oplus},\nu_{2\oplus}}}$, for $\widetilde{\Gamma}$ and $\widehat{\Gamma}$, respectively. To guarantee the atomlessness of $\nu_{1\oplus}$ and $\nu_{2\oplus}$ and their estimators $\widetilde{\nu}_{1\oplus}$ and $\widetilde{\nu}_{2\oplus}$, we assume

(A4) With probability equal to 1, the random distributions ν_1 and ν_2 are atomless.

Let $C > 1$ denote a constant. To derive the convergence rate of the estimators for the regression operator, $\widetilde{\Gamma}$ and $\widehat{\Gamma}$, we require the following conditions regarding the variability of ν_1 and ν_2, the spacing of the eigenvalues λ_j and ς_k, and the decay rates of the coefficients b_{jk}. Conditions of this type are

standard in traditional functional linear regression (e.g., **?**).

(A5) $\mathbb{E}(\|\mathrm{Log}_{\nu_{1\oplus}}\nu_1\|^4_{\nu_{1\oplus}}) < \infty$, and $\mathbb{E}(\langle \mathrm{Log}_{\nu_{1\oplus}}\nu_1, \phi_j\rangle^4_{\nu_{1\oplus}}) \le C\lambda_j^2$, for all $j \ge 1$; $\mathbb{E}(\|\mathrm{Log}_{\nu_{2\oplus}}\nu_2\|^4_{\nu_{2\oplus}}) < \infty$, and $\mathbb{E}(\langle \mathrm{Log}_{\nu_{2\oplus}}\nu_2, \psi_k\rangle^4_{\nu_{2\oplus}}) \le C\varsigma_k^2$, for all $k \ge 1$.

(A6) For $j \ge 1$, $\lambda_j - \lambda_{j+1} \ge C^{-1}j^{-\theta-1}$, where $\theta \ge 1$ is a constant.

(A7) For $k \ge 1$, $\varsigma_k - \varsigma_{k+1} \ge C^{-1}k^{-\vartheta-1}$, where $\vartheta > 0$ is a constant.

(A8) For $j, k \ge 1$, $|b_{jk}| \le Cj^{-\rho}k^{-\varrho}$, where $\rho > \theta + 1$ and $\varrho > 1$ are constants.

Note that (A8) implies (14). Furthermore, for J and K in (20) and (22), we assume

(A9) $n^{-1}J^{2\theta+2} \to 0$, $n^{-1}K^{2\vartheta+2} \to 0$, as $n \to \infty$.

Let $\mathscr{F} = \mathscr{F}(C, \theta, \vartheta, \rho, \varrho)$ denote the set of distributions $\mathcal{F}$ of (ν_1, ν_2) that satisfy (A1) and (A4)–(A8). Defining the sequence

$$\begin{aligned}\varkappa(n) &= \varkappa(n; \theta, \vartheta, \rho, \varrho)\\ &= \begin{cases} \min\left\{n^{\max\{2\rho/(2\vartheta+3),(4\rho-1)/(2\vartheta+2\varrho+2)\}/(\theta+2\rho)}, n^{1/(2\vartheta+3)}\right\}, & \text{if } \varrho - \vartheta \le 1,\\ \min\left\{n^{\max\{2\rho/(2\vartheta+3),(4\rho-1)/(2\vartheta+2\varrho+2)\}/(\theta+2\rho)}, (n/\log n)^{1/(2\vartheta+3)}\right\}, & \text{if } \varrho - \vartheta \in (1, 3/2],\\ \min\left\{n^{\max\{2\rho/(2\vartheta+3),(4\rho-1)/(2\vartheta+2\varrho+2)\}/(\theta+2\rho)}, n^{1/(2\varrho)}\right\}, & \text{if } \varrho - \vartheta > 3/2,\end{cases}\end{aligned}$$

then when distributions ν_{1i} and ν_{2i} are fully observed, we obtain

Theorem 1. *Assume (A1) and (A4)–(A9). If $J \sim n^{1/(\theta+2\rho)}$ and $K \sim \varkappa(n)$, as $n \to \infty$, then*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{\mathcal{F}\in\mathscr{F}} \mathbb{P}_{\mathcal{F}}\left(\|\mathcal{P}_{(\widetilde{\nu}_{1\oplus},\widetilde{\nu}_{2\oplus}),(\nu_{1\oplus},\nu_{2\oplus})}\widetilde{\Gamma} - \Gamma\|^2_{\mathcal{H}_{\nu_{1\oplus},\nu_{2\oplus}}} > M\alpha(n)\right) = 0, \tag{25}$$

where

$$\alpha(n) = \max\left\{n^{-(2\rho-1)/(\theta+2\rho)}, \varkappa(n)^{-(2\varrho-1)}\right\}. \tag{26}$$

We note that $\alpha(n) = n^{-(2\rho-1)/(\theta+2\rho)}$ in (26) if either of the following holds: $\varrho - \vartheta \le 1$ and $4\rho(\vartheta - \varrho + 2) \le 2\vartheta + 3 \le (2\varrho - 1)(\theta + 2\rho)/(2\rho - 1)$; or $1 < \varrho - \vartheta \le 3/2$ and $4\rho(\vartheta - \varrho + 2) \le 2\vartheta + 3 <$

$(2\varrho-1)(\theta+2\rho)/(2\rho-1)$; or $\varrho-\vartheta>3/2$ and $\varrho\geq\max\{\vartheta+2-(2\vartheta+3)/(4\rho),\,(\theta+2\rho)/(2\theta+2)\}$. In this case, $\widetilde{\Gamma}$ achieves the same rate as the minimax rate for function-to-scalar linear regression (**?**) and function-to-function linear regression (following similar arguments as in the proof of Theorem 3 of **?**).

Next, we consider the case where the distributions ν_{1i} and ν_{2i} are not fully observed. In addition, we require an assumption regarding the number of measurements per distribution and a uniform Lipschitz condition on the estimated cdfs to guarantee the atomlessness of the estimated Fréchet means $\widehat{\nu}_{1\oplus}$ and $\widehat{\nu}_{2\oplus}$ and hence to justify the use of $\|\mathcal{P}_{(\widehat{\nu}_{1\oplus},\widehat{\nu}_{2\oplus}),(\nu_{1\oplus},\nu_{2\oplus})}\widehat{\Gamma}-\Gamma\|_{\mathcal{H}_{\nu_{1\oplus},\nu_{2\oplus}}}$ as a measure of the estimation error.

(A10) For τ_m in (A2), $\tau_m\leq C\min\{n^{-1}J^{-\theta},n^{-1}K^{-1}\}$, for all n.

(A11) For any atomless distribution $\mu\in\mathcal{W}$, the corresponding estimate $\widehat{\mu}$ based on a sample of measurements drawn according to μ is also atomless.

For example, with $J\sim n^{1/(\theta+2\rho)}$ and $K\sim\varkappa(n)$ as in Theorem 1, (A10) holds with $m\sim\max\{n^{3(\theta+\rho)/(\theta+2\rho)},\,n^{3/2}\varkappa($ and $m\sim\max\{n^{4(\theta+\rho)/(\theta+2\rho)},\,n^2\varkappa(n)^2\}$ for the estimators proposed by **?** and **?**, respectively. We note that these two estimators also satisfy (A11). Then we find that the data-based estimator $\widehat{\Gamma}$ achieves the same rate as the estimator $\widetilde{\Gamma}$ based on fully observed distributions as shown in Theorem 1.

Theorem 2. *If (A1)–(A11) hold and choosing J and K as in Theorem 1, then*

$$\lim_{M\to\infty}\limsup_{n\to\infty}\sup_{\mathcal{F}\in\mathscr{F}}\mathbb{P}_{\mathcal{F}}\left(\|\mathcal{P}_{(\widehat{\nu}_{1\oplus},\widehat{\nu}_{2\oplus}),(\nu_{1\oplus},\nu_{2\oplus})}\widehat{\Gamma}-\Gamma\|^2_{\mathcal{H}_{\nu_{1\oplus},\nu_{2\oplus}}}>M\alpha(n)\right)=0. \tag{27}$$

We note that while the proposed method is based on function-to-function linear regression, the asymptotic analysis is more involved. The proofs of Theorems 1 and 2 are based on the geometry of the Wasserstein space, since we are not dealing with general functions in $\mathcal{L}^2$ space (with respect to the Lebesgue measure) as in functional data analysis but rather the log maps. In particular, we do not assume additive noise in the proposed model in (10). Furthermore, parallel transport maps are employed to quantify the estimation discrepancy of the estimators of the regression operator, Γ, the covariance and cross-covariance operators, $\mathcal{C}_{\nu_1}$, $\mathcal{C}_{\nu_2}$ and $\mathcal{C}_{\nu_1\nu_2}$, and the eigenfunctions, ϕ_j and ψ_k. All of

these create additional complexities for the theoretical derivations. For Theorem 2, the distributions ν_{1i} and ν_{2i} are not be fully observed and instead only data samples drawn from these distributions are available. Hence, we need to deal with two layers of stochastic mechanisms: The first layer generates random elements (ν_{1i}, ν_{2i}) taking values in $\mathcal{W}\times\mathcal{W}$; the second layer generates random samples according to ν_{1i} and ν_{2i}. Specifically, we need to tackle the discrepancy between the estimated distributions based on the observed data $\widehat{\nu}_{1i}$ and $\widehat{\nu}_{2i}$ and the actual underlying distributions ν_{1i} and ν_{2i}.

Theorems 1 and 2 entail the following corollaries on the prediction of ν_2 based on ν_1, where the target is the conditional Fréchet mean of ν_2 given ν_1, i.e., $\mathbb{E}_{\oplus}(\nu_2|\nu_1) \coloneqq \operatorname{argmin}_{\mu'\in\mathcal{W}} \mathbb{E}[d_W^2(\nu_2, \mu') \mid \nu_1] = \operatorname{Exp}_{\nu_{2\oplus}}[\mathbb{E}(\operatorname{Log}_{\nu_{2\oplus}}\nu_2 \mid \operatorname{Log}_{\nu_{1\oplus}}\nu_1)]$. In the following, for any given $\mu \in \mathcal{W}$, the corresponding estimate $\widehat{\mu}$ is assumed to be based on a sample of $m_\mu \geq m$ observations drawn from μ, where m is the lower bound of the number of observations per distribution as per (A3). We denote the prediction of ν_2 based on fully observed distributions by $\widetilde{\nu}_2(\mu) \coloneqq \operatorname{Exp}_{\widetilde{\nu}_{2\oplus}}[\widetilde{\Gamma}(\operatorname{Log}_{\widetilde{\nu}_{1\oplus}}\mu)]$, and the prediction based on samples generated from the distributions by $\widehat{\nu}_2(\widehat{\mu}) \coloneqq \operatorname{Exp}_{\widehat{\nu}_{2\oplus}}[\widehat{\Gamma}(\operatorname{Log}_{\widehat{\nu}_{1\oplus}}\widehat{\mu})]$, where $\widetilde{\Gamma}$ and $\widehat{\Gamma}$ are as per (20) and (22), respectively.

Corollary 1. *Under the assumptions of Theorem 1,*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{\mathcal{F}\in\mathscr{F}} \mathbb{P}_{\mathcal{F}} \left(d_W^2 \left(\widetilde{\nu}_2(\mu), \mathbb{E}_{\oplus}(\nu_2 \mid \nu_1 = \mu) \right) > M\alpha(n) \right) = 0. \tag{28}$$

Corollary 2. *Under the assumptions of Theorem 2,*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{\mathcal{F}\in\mathscr{F}} \mathbb{P}_{\mathcal{F}} \left(d_W^2 \left(\widehat{\nu}_2(\widehat{\mu}), \mathbb{E}_{\oplus}(\nu_2 \mid \nu_1 = \mu) \right) > M\alpha(n) \right) = 0.$$

For the proofs, see Section S.1.2 in the Supplementary Material. We further discuss the estimation and theoretical analysis for the distribution-to-scalar regression model as per (15) in Section S.2 in the Supplementary Material, where we show that the estimates of the regression coefficient function β_1 achieve the same rate as the minimax rate for the function-to-scalar linear regression based on fully observed predictor functions; see **?**.

4 Autoregressive Models for Distribution-Valued Time Series

Here we consider a distribution-valued time series $\{\mu_i\}_{i\in\mathbb{Z}}$, each element taking values in $\mathcal{W}$. We assume that the random process $\{\mu_i\}_{i\in\mathbb{Z}}$ is stationary in the sense that

(1) μ_i are square integrable, i.e., $\mathbb{E}d_W^2(\mu,\mu_i)<\infty$ for some (and thus for all) $\mu\in\mathcal{W}$;

(2) μ_i have a common Fréchet mean $\mu_\oplus$ that is atomless, i.e, $\mu_\oplus=\operatorname{argmin}_{\mu\in\mathcal{W}}\mathbb{E}d_W^2(\mu,\mu_i)$, for all $i\in\mathbb{Z}$;

(3) The autocovariance operators $\mathbb{E}(\operatorname{Log}_{\mu_\oplus}\mu_{i+r}\otimes\operatorname{Log}_{\mu_\oplus}\mu_i)$ do not depend on $i\in\mathbb{Z}$, which are hence denoted by $\mathcal{C}_r$, for all $r\in\mathbb{Z}$.

For $\{\mu_i\}_{i\in\mathbb{Z}}$, we assume a first order autoregressive model which is an extension of the distribution-to-distribution regression model in (10)

$$\operatorname{Log}_{\mu_\oplus}\mu_{i+1}=\Gamma(\operatorname{Log}_{\mu_\oplus}\mu_i)+\varepsilon_{i+1},\ \text{for } i\in\mathbb{Z}. \tag{29}$$

Here, $\Gamma\colon T_{\mu_\oplus}\to T_{\mu_\oplus}$ is a linear operator defined as

$$\Gamma g(t)=\langle\beta(\cdot,t),g\rangle_{\mu_\oplus},\quad \text{for } t\in D,\ \text{and } g\in T_{\mu_\oplus}, \tag{30}$$

where $\beta: D^2\to\mathbb{R}$ is the auto-regression coefficient kernel lying in $\mathcal{L}^2_{\mu_\oplus\times\mu_\oplus}$, and $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ are i.i.d. random elements taking values in the tangent space $T_{\mu_\oplus}$ such that $\mathbb{E}(\varepsilon_i)=0$ and $\mathbb{E}\|\varepsilon_i\|^2_{\mu_\oplus}<\infty$. Similar models have been previously studied in the seminal work of **?**. To ensure the existence and uniqueness of such a stationary process, we assume

(B1) There exists an integer $q\geq 1$ such that $\|\Gamma^q\|_{\mathcal{L}^2_{\mu_\oplus}}<1$.

Here, $\|\cdot\|_{\mathcal{L}^2_{\mu_\oplus}}$ denotes the sup norm for linear operators on $\mathcal{L}^2_{\mu_\oplus}$ and we define Γ^q by induction, $\Gamma^k(\cdot)=\Gamma[\Gamma^{k-1}(\cdot)]$, for any integer $k>1$. We note that under (B1), (29) has a unique stationary solution given

by

$$\mathrm{Log}_{\mu_\oplus}\mu_i = \sum_{r=0}^{\infty} \Gamma^r(\varepsilon_{i-r}), \tag{31}$$

where $\Gamma^0(\varepsilon_i) := \varepsilon_i$ and the right hand side converges in mean square, $\lim_{n\to\infty} \mathbb{E}\| \sum_{r=n}^{\infty} \Gamma^r(\varepsilon_{i-r})\|^2_{\mu_\oplus} = 0$, and also almost surely, i.e., $\lim_{n\to\infty} \| \sum_{r=n}^{\infty} \Gamma^r(\varepsilon_{i-r})\|_{\mu_\oplus} = 0$ with probability 1 (Theorem 3.1, **?**). Furthermore, we assume

(B2) With probability 1, $\sum_{r=0}^{\infty} \Gamma^r(\varepsilon_{-r}) + \mathrm{id}$ is non-decreasing.

Assumption (B2) guarantees that the right hand side of (31) lies in $\mathrm{Log}_{\mu_\oplus}\mathcal{W}$ a.s. We further provide a fully detailed example of a stationary process $\{\mu_i\}_{i\in\mathbb{Z}}$ that satisfies the autoregressive model as per (29) in Section S.3 in the Supplementary Material.

As in Section 3, we have $\mathbb{E}(\mathrm{Log}_{\mu_\oplus}\mu_1) = 0$, $\mu_\oplus$-almost surely. The operator $\mathcal{C}_0$ admits the eigendecomposition

$$\mathcal{C}_0 = \sum_{j=1}^{\infty} \lambda_j \phi_j \otimes \phi_j,$$

with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and orthonormal eigenfunctions $\{\phi_j\}_{j=1}^{\infty}$ in $T_{\mu_\oplus}$. With probability 1, the logarithmic transforms $\mathrm{Log}_{\mu_\oplus}\mu_i$ admit the expansion

$$\mathrm{Log}_{\mu_\oplus}\mu_i = \sum_{j=1}^{\infty} \langle \mathrm{Log}_{\mu_\oplus}\mu_i, \phi_j\rangle_{\mu_\oplus}\phi_j, \quad i \in \mathbb{Z},$$

and hence $\mathcal{C}_1 = \sum_{l=1}^{\infty}\sum_{j=1}^{\infty} \xi_{jl}\phi_l \otimes \phi_j$, where $\xi_{jl} = \mathbb{E}(\langle \mathrm{Log}_{\mu_\oplus}\mu_1, \phi_j\rangle_{\mu_\oplus}\langle \mathrm{Log}_{\mu_\oplus}\mu_2, \phi_l\rangle_{\mu_\oplus})$. With $b_{jl} = \lambda_j^{-1}\xi_{jl}$, the auto-regression coefficient function can then be expressed as

$$\beta = \sum_{l=1}^{\infty}\sum_{j=1}^{\infty} b_{jl}\phi_l \otimes \phi_j.$$

For the estimation of the operator Γ in (30), first considering a fully observed sequence of length n, $\mu_1, \mu_2, \ldots, \mu_n$, with the oracle estimator of the Fréchet mean $\widetilde{\mu}_\oplus$ defined analogously to (18), the autocovariance operators $\mathcal{C}_0$ and $\mathcal{C}_1$ can be estimated by their empirical counterparts $\widetilde{\mathcal{C}}_0 = n^{-1}\sum_{i=1}^{n} \mathrm{Log}^{\sim}_{\widetilde{\mu}_\oplus}\mu_i \otimes \mathrm{Log}^{\sim}_{\widetilde{\mu}_\oplus}\mu_i$ and $\widetilde{\mathcal{C}}_1 = (n-1)^{-1}\sum_{i=1}^{n-1} \mathrm{Log}^{\sim}_{\widetilde{\mu}_\oplus}\mu_{i+1} \otimes \mathrm{Log}^{\sim}_{\widetilde{\mu}_\oplus}\mu_i$. We denote the eigenvalues and eigenfunctions

of $\widetilde{\mathcal{C}}_0$ by $\widetilde{\lambda}_j$ and $\widetilde{\phi}_j$, respectively, where the eigenvalues $\widetilde{\lambda}_j$ are in non-ascending order. Then oracle estimators for the auto-regression coefficient function β and operator Γ in (30) are

$$\widetilde{\beta} = \sum_{l=1}^{J}\sum_{j=1}^{J} \widetilde{b}_{jl}\widetilde{\phi}_l \otimes \widetilde{\phi}_j, \quad \text{and} \quad \widetilde{\Gamma} g(t) = \langle g, \widetilde{\beta}(\cdot, t)\rangle_{\widetilde{\mu}_\oplus}, \text{ for } g \in \mathrm{Log}_{\widetilde{\mu}_\oplus}\mathcal{W}, \ t \in D, \tag{32}$$

where $\widetilde{b}_{jl} = \widetilde{\lambda}_j^{-1}\widetilde{\xi}_{jl}$, $\widetilde{\xi}_{jl} = (n-1)^{-1}\sum_{i=1}^{n-1}\langle \mathrm{Log}_{\widetilde{\mu}_\oplus}\mu_i, \widetilde{\phi}_j\rangle_{\widetilde{\mu}_\oplus}\langle \mathrm{Log}_{\widetilde{\mu}_\oplus}\mu_{i+1}, \widetilde{\phi}_l\rangle_{\widetilde{\mu}_\oplus}$, and J is the truncation bound.

As discussed for the independent case in Section 3.2, a realistic estimator $\widehat{\beta}$ for β based on the distribution estimation discussed in Section 3.1 can be obtained by replacing μ_i and $\mu_\oplus$ with the corresponding estimates $\widehat{\mu}_i$ and $\widehat{\mu}_\oplus$, the latter analogous to (21). Specifically, estimates for the autocovariance operators with corresponding decompositions are given by $\widehat{\mathcal{C}}_0 = n^{-1}\sum_{i=1}^{n} \mathrm{Log}_{\widehat{\mu}_\oplus}\widehat{\mu}_i \otimes \mathrm{Log}_{\widehat{\mu}_\oplus}\widehat{\mu}_i$ and $\widehat{\mathcal{C}}_1 = (n-1)^{-1}\sum_{i=1}^{n-1} \mathrm{Log}_{\widehat{\mu}_\oplus}\widehat{\mu}_{i+1} \otimes \mathrm{Log}_{\widehat{\mu}_\oplus}\widehat{\mu}_i$. We denote the eigenvalues and eigenfunctions of $\widehat{\mathcal{C}}_0$ by $\widehat{\lambda}_j$ and $\widehat{\phi}_j$, respectively, where the eigenvalues $\widehat{\lambda}_j$ are in non-ascending order. With $\widehat{b}_{jl} = \widehat{\lambda}_j^{-1}\widehat{\xi}_{jl}$ and $\widehat{\xi}_{jl} = (n-1)^{-1}\sum_{i=1}^{n-1}\langle \mathrm{Log}_{\widehat{\mu}_\oplus}\widehat{\mu}_i, \widehat{\phi}_j\rangle_{\widehat{\mu}_\oplus}\langle \mathrm{Log}_{\widehat{\mu}_\oplus}\widehat{\mu}_{i+1}, \widehat{\phi}_l\rangle_{\widehat{\mu}_\oplus}$, data-based estimators for the auto-regression coefficient function β and operator Γ in (30) are then given by

$$\widehat{\beta} = \sum_{l=1}^{J}\sum_{j=1}^{J} \widehat{b}_{jl}\widehat{\phi}_l \otimes \widehat{\phi}_j, \quad \text{and} \quad \widehat{\Gamma} g(t) = \langle g, \widehat{\beta}(\cdot, t)\rangle_{\widehat{\mu}_\oplus}, \text{ for } g \in \mathrm{Log}_{\widehat{\mu}_\oplus}\mathcal{W}, \ t \in D. \tag{33}$$

We first focus on the case where the distributions are fully observed. To derive the convergence rate of the estimator $\widetilde{\Gamma}$ in (32), we require the following assumptions analogous to the independent case in Section 3. Let $C > 1$ be a constant.

(B3) With probability 1, the distributions μ_i are all atomless.

(B4) $\mathbb{E}(\|\mathrm{Log}_{\mu_\oplus}\mu_i\|^4_{\mu_\oplus}) < \infty$, and $\mathbb{E}(\langle \mathrm{Log}_{\mu_\oplus}\mu_i, \phi_j\rangle^4_{\mu_\oplus}) \le C\lambda_j^2$, for all $j \ge 1$.

(B5) For $j \ge 1$, $\lambda_j - \lambda_{j+1} \ge C^{-1}j^{-\theta-1}$, where $\theta \ge 1/2$ is a constant.

(B6) For $j, l \ge 1$, $|b_{jl}| \le Cj^{-\rho}l^{-\varrho}$, where $\rho > \theta + 1$ and $\varrho > 1$ are constants.

(B7) $n^{-1}J^{2\theta+2} \to 0$, as $n \to \infty$.

Let $\mathscr{G} = \mathscr{G}(C, \theta, \rho, \varrho)$ denote the set of distributions $\mathcal{G}$ of the process $\{\mu_i\}$ that satisfy (B1)–(B6).

Then we obtain

Theorem 3. *Assume (B1)–(B7). If* $J \sim \min\{n^{1/(2\theta+2\rho+2\max\{2-\varrho,0\})}, n^{1/(2\theta+2\max\{\varrho,2\})}\}$, *then*

$$\lim_{M\to\infty}\limsup_{n\to\infty}\sup_{\mathcal{G}\in\mathscr{G}}\mathbb{P}_{\mathcal{G}}\left(\|\mathcal{P}_{(\widetilde{\mu}_\oplus,\widetilde{\mu}_\oplus),(\mu_\oplus,\mu_\oplus)}\widetilde{\Gamma}-\Gamma\|^2_{\mathcal{H}_{\mu_\oplus,\mu_\oplus}}>M\zeta(n)\right)=0,$$

where

$$\zeta(n)=\max\left\{n^{-(2\rho-1)/(2\theta+2\rho+2\max\{2-\varrho,0\})}, n^{-(2\varrho-1)/(2\theta+2\max\{\varrho,2\})}\right\}. \tag{34}$$

The convergence rate obtained for the estimator $\widetilde{\Gamma}$ in Theorem 3 is slower than the rate obtained for the independent case as per Theorem 1. This is due to the serial dependence among μ_i and with the special choice of J as above is manifested by the fact that $\alpha(n)$ as per (26) with $\theta=\vartheta$ is always smaller than $\zeta(n)$ as per (34).

Furthermore, regarding the estimator $\widehat{\Gamma}$ in (33) where only samples drawn from the distributions μ_i are available, we in addition make the following assumption of the numbers of measurements observed per distribution.

(B8) There exists a sequence $m=m(n)$ such that for the number of measurements per distribution m_{μ_i}, $\min\{m_{\mu_i}: i=1,2,\ldots,n\}\geq m$ and $m\to\infty$ as $n\to\infty$.

(B9) $\tau_m\leq Cn^{-1}$, for all n, where τ_m is as per (A2).

For example, if distributions μ_i are estimated via the methods used by **?** and **?**, in order to ensure (B9), it suffices to take $m\sim n^2$ and $m\sim n^{3/2}$, respectively. Then we show that the estimator $\widehat{\Gamma}$ in (33) converges with the same rate as $\widetilde{\Gamma}$, as shown in Theorem 3.

Theorem 4. *If (A2), (A11) and (B1)–(B9) hold and choosing* J *as in Theorem 3, then*

$$\lim_{M\to\infty}\limsup_{n\to\infty}\sup_{\mathcal{G}\in\mathscr{G}}\mathbb{P}_{\mathcal{G}}\left(\|\mathcal{P}_{(\widehat{\mu}_\oplus,\widehat{\mu}_\oplus),(\mu_\oplus,\mu_\oplus)}\widehat{\Gamma}-\Gamma\|^2_{\mathcal{H}_{\mu_\oplus,\mu_\oplus}}>M\zeta(n)\right)=0.$$

As for the independent case, Theorems 3 and 4 entail the following asymptotic results for the one-on-one prediction of μ_{n+1} given μ_n, where the target is the conditional Fréchet mean of μ_{n+1} given μ_n

by $\mathbb{E}_{\oplus}(\mu_{n+1} \mid \mu_n) \coloneqq \operatorname{argmin}_{\mu'} \mathbb{E}[d_W^2(\mu_{n+1}, \mu') \mid \mu_n] = \operatorname{Exp}_{\mu_{\oplus}}[\mathbb{E}(\operatorname{Log}_{\mu_{\oplus}} \mu_{n+1} \mid \operatorname{Log}_{\mu_{\oplus}} \mu_n)]$. For any given $\mu \in \mathcal{W}$, the corresponding estimate $\widehat{\mu}$ is assumed to be based on a sample of $m_\mu \geq m$ observations drawn from μ, where m is the lower bound of the number of observations per distribution as per (B8). The prediction of μ_{n+1} based on fully observed distributions is given by $\operatorname{Exp}_{\widetilde{\mu}_{\oplus}}[\widetilde{\Gamma}(\operatorname{Log}_{\widetilde{\mu}_{\oplus}} \mu)]$ and the prediction based on samples generated from the distributions by $\operatorname{Exp}_{\widehat{\mu}_{\oplus}}[\widehat{\Gamma}(\operatorname{Log}_{\widehat{\mu}_{\oplus}} \widehat{\mu})]$, where $\widetilde{\Gamma}$ and $\widehat{\Gamma}$ are as per (32) and (33), respectively. Then these predictions achieve the same rate as the estimates of the regression operators in Theorems 3 and 4.

Corollary 3. *Under the assumptions of Theorem 3,*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{\mathcal{G}\in\mathscr{G}} \mathbb{P}_{\mathcal{G}} \left(d_W^2(\operatorname{Exp}_{\widetilde{\mu}_{\oplus}}[\widetilde{\Gamma}(\operatorname{Log}_{\widetilde{\mu}_{\oplus}} \mu)], \mathbb{E}_{\oplus}(\nu_2 \mid \nu_1 = \mu)) > M\zeta(n) \right) = 0.$$

Corollary 4. *Under the assumptions of Theorem 4,*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{\mathcal{G}\in\mathscr{G}} \mathbb{P}_{\mathcal{G}} \left(d_W^2(\operatorname{Exp}_{\widehat{\mu}_{\oplus}}[\widehat{\Gamma}(\operatorname{Log}_{\widehat{\mu}_{\oplus}} \widehat{\mu})], \mathbb{E}_{\oplus}(\mu_{n+1} \mid \mu_n = \mu)) > M\zeta(n) \right) = 0.$$

Proofs and auxiliary lemmas for this section are in Section S.1.3 in the Supplementary Material.

5 Simulations

In practice, the fit of the logarithmic response may not fall in the logarithmic space with base point $\widehat{\nu}_{2\oplus}$, i.e.,

$$\widehat{\Gamma}(\operatorname{Log}_{\widehat{\nu}_{1\oplus}} \widehat{\nu}_{1i}) \notin \operatorname{Log}_{\widehat{\nu}_{2\oplus}} \mathcal{W}, \tag{35}$$

with $\widehat{\Gamma}$ given in (22). This problem was already recognized by **?**. If (35) happens, we employ a boundary projection method described in Section S.4.2 in the Supplementary Material. We compared the performance of the proposed method implemented with boundary projection (referred to as projection method) with two other approaches. The first of these is to employ an alternative to the proposed

boundary projection for those situations where the event (35) takes place, which was proposed by **?** in the context of principal component analysis (PCA). This alternative to handle the problem extends the domains of the distributions. We use this method by fitting the proposed distribution-to-distribution regression model with distributions on an extended domain when the event (35) happens, and then normalize the fitted distributions by restricting them back to the original domain. We refer to this as the domain-extension method in the following. The second alternative approach is the log quantile density (LQD) method (**?**), where we apply function-to-function linear regression to the LQD transformations of distributions and map the fitted responses back to the Wasserstein space $\mathcal{W}$ through the inverse LQD transformation (**?**). Specifically, we use the R package `fdadensity` (**?**) for implementations of the LQD transformations. To generate data for simulations, we provide the following framework to construct explicit examples, which also demonstrates the feasibility of the proposed model in (10).

Framework for Explicit Construction. For $D = [0, 1]$, we consider Fréchet mean distributions $\nu_{1\oplus}, \nu_{2\oplus} \in \mathcal{W}$ with bounded density functions, i.e., $\sup_{s\in D} f_{1\oplus}(s) < \infty$ and $\sup_{t\in D} f_{2\oplus}(t) < \infty$. We consider a set of orthonormal functions $\{\varphi_j\}_{j=1}^{\infty}$ in the Lebesgue-square-integrable function space on $[0, 1]$, $\mathcal{L}^2([0, 1])$, such that the φ_j are continuously differentiable with bounded derivatives, and $\varphi_j(0) = \varphi_j(1)$, for all $j \in \mathbb{N}_+$. In particular, φ_j can be taken as

$$\varphi_j(r) = \sqrt{2}\sin(2\pi jr), \quad \text{for } r \in [0, 1], \text{ and } j \in \mathbb{N}_+. \tag{36}$$

Suppose $\mathrm{Log}_{\nu_{1\oplus}}\nu_1$ admits the expansion $\mathrm{Log}_{\nu_{1\oplus}}\nu_1 = \sum_{j=1}^{\infty}\chi_j\varphi_j \circ F_{1\oplus}$, where χ_j are uncorrelated random variables with zero mean such that $\sum_{j=1}^{\infty}\chi_j^2 < \infty$ almost surely. We define the regression operator Γ as $\Gamma g = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} b_{jk}^*\langle g, \varphi_j \circ F_{1\oplus}\rangle_{\nu_{1\oplus}}\varphi_k \circ F_{2\oplus}$, for $g \in T_{\nu_{1\oplus}}$, with $b_{jk}^* \in \mathbb{R}$ such that $\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} {b_{jk}^*}^2 < \infty$. Hence, $\Gamma(\mathrm{Log}_{\nu_{1\oplus}}\nu_1) = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} b_{jk}^*\chi_j\varphi_k \circ F_{2\oplus}$. To guarantee $\sum_{j=1}^{\infty}\chi_j\varphi_j \circ F_{1\oplus} \in$

$\mathrm{Log}_{\nu_{1\oplus}}\mathcal{W}$ and $\sum_{k=1}^{\infty}\sum_{j=1}^{\infty} b_{jk}^{*}\chi_j\varphi_k \circ F_{2\oplus} \in \mathrm{Log}_{\nu_{2\oplus}}\mathcal{W}$, it suffices to require

$$\begin{cases} \sum_{j=1}^{\infty} \chi_j \varphi_j'(F_{1\oplus}(s)) f_{1\oplus}(s) + 1 \geq 0, \text{ for all } s \in D, \\ \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} b_{jk}^{*}\chi_j \varphi_k'(F_{2\oplus}(t)) f_{2\oplus}(t) + 1 \geq 0, \text{ for all } t \in D, \\ \sum_{j=1}^{\infty} \chi_j (\varphi_j \circ F_{1\oplus})' \text{ and } \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} b_{jk}^{*}\chi_j (\varphi_k \circ F_{2\oplus})' \text{ uniformly converge,} \end{cases} \quad \text{a.s.} \tag{37}$$

Requirement (37) is satisfied, e.g., when $|\chi_j| \leq \upsilon_{1j}/(\sup_{r\in[0,1]}|\varphi_j'(r)| \sup_{s\in D} f_{1\oplus}(s) \sum_{j'=1}^{\infty}\upsilon_{1j'})$ and $|b_{jk}^{*}\chi_j| \leq \upsilon_{1j}\upsilon_{2k}/(\sup_{r\in[0,1]}|\varphi_k'(r)| \sup_{t\in D} f_{2\oplus}(t) \sum_{j'=1}^{\infty}\upsilon_{1j'}\sum_{k'=1}^{\infty}\upsilon_{2k'})$, a.s., where $\{\upsilon_{1j}\}_{j=1}^{\infty}$ and $\{\upsilon_{2k}\}_{k=1}^{\infty}$ are two non-negative sequences such that $\sum_{j=1}^{\infty}\upsilon_{1j} < \infty$ and $\sum_{k=1}^{\infty}\upsilon_{2k} < \infty$, examples including $\{a^{-j}\}_{j=1}^{\infty}$ and $\{j^{-a}\}_{j=1}^{\infty}$, for any given $a > 1$.

With $\Gamma(\mathrm{Log}_{\nu_{1\oplus}}\nu_1)$ and $\nu_{2\oplus}$, the distributional response ν_2 can be generated by adding distortions to $\mathrm{Exp}_{\nu_{2\oplus}}(\Gamma(\mathrm{Log}_{\nu_{1\oplus}}\nu_1))$ through push-forward maps, i.e., $\nu_2 = g\#\mathrm{Exp}_{\nu_{2\oplus}}(\Gamma(\mathrm{Log}_{\nu_{1\oplus}}\nu_1))$, where $g\colon D \to D$ is a random distortion function independent of ν_1, such that g is non-decreasing almost surely, and that $\mathbb{E}[g(t)] = t$ almost everywhere on D. This is a valid method to provide random distortions for distributions (**?**) in the sense that the conditional Fréchet mean of ν_2 is on target, i.e., $\mathbb{E}_{\oplus}(\nu_2|\nu_1) := \mathrm{argmin}_{\mu\in\mathcal{W}}\mathbb{E}[d_W^2(\nu_2,\mu) \mid \nu_1] = \mathrm{Exp}_{\nu_{2\oplus}}(\Gamma(\mathrm{Log}_{\nu_{1\oplus}}\nu_1))$. Furthermore, the pair (ν_1,ν_2) generated in this way satisfies our model in (10). An example (**?**) of the random distortion function is $g = g_A$, where A is a random variable such that $\mathbb{P}(A \leq r) = \mathbb{P}(A \geq -r)$ for any $r \in \mathbb{R}$ and $\mathbb{P}(A = 0) = 0$, and g_a is defined as

$$g_a(r) = \begin{cases} r - |a|^{-1}\sin(ar), & \text{if } a \neq 0, \\ r, & \text{if } a = 0, \end{cases} \quad \text{for } r \in D. \tag{38}$$

Specifically, for our simulation studies, with $D = [0,1]$, we consider two cases with different choices of the Fréchet means $\nu_{1\oplus}$ and $\nu_{2\oplus}$:

Case 1. $\nu_{1\oplus} = TN_D(0.5, 0.2^2)$, and $\nu_{2\oplus} = TN_D(0.75, 0.3^2)$, where $TN_D(\mu,\sigma^2)$ denotes the Gaussian distribution $N(\mu,\sigma^2)$ truncated on D.

Case 2. $\nu_{1\oplus} = \mathrm{Beta}(6,2)$, and $\nu_{2\oplus} = \mathrm{Beta}(2,4)$.

Taking $J^* = K^* = 20$, for $j, k \in \mathbb{N}_+$, we set $b^*_{jk} = 2^{-k}\kappa_k^{-1}R_{2\oplus}^{-1}\kappa_j R_{1\oplus}$ if $j \leq J^*$ and $k \leq K^*$, and set $b^*_{jk} = 0$ otherwise, where $\kappa_l = \sup_{r\in[0,1]} |\varphi_l'(r)| = 2\sqrt{2}\pi l$, for $l \in \mathbb{N}_+$, $R_{1\oplus} = \sup_{s\in D} f_{1\oplus}(s)$ and $R_{2\oplus} = \sup_{t\in D} f_{2\oplus}(t)$. Taking $\upsilon_{1j} = 2^{-j}$, data were generated as follows:

Step 1: Generate $\chi_{ij} \sim \text{Unif}(-\upsilon_{1j}(\kappa_j R_{1\oplus}\sum_{l=1}^{\infty}\upsilon_{1l})^{-1}, \upsilon_{1j}(\kappa_j R_{1\oplus}\sum_{l=1}^{\infty}\upsilon_{1l})^{-1})$ independently for $i = 1, \ldots, n$ and $j = 1, \ldots, J^*$, whence $\text{Log}_{\nu_{1\oplus}}\nu_{1i} = \sum_{j=1}^{J^*}\chi_{ij}\varphi_j \circ F_{1\oplus}$, with the basis functions φ_j as per (36), $\Gamma(\text{Log}_{\nu_{1\oplus}}\nu_{1i}) = \sum_{k=1}^{K^*}\sum_{j=1}^{J^*} b^*_{jk}\chi_{ij}\varphi_k \circ F_{2\oplus}$, and $\nu_{1i} = \text{Exp}_{\nu_{1\oplus}}(\sum_{j=1}^{J^*}\chi_{ij}\varphi_j \circ F_{1\oplus})$.

Step 2: Generate ν_{2i} by adding distortion to $\Gamma(\text{Log}_{\nu_{1\oplus}}\nu_{1i})$: Sample $A_i \overset{\text{iid}}{\sim} \text{Unif}\{\pm\pi, \pm 2\pi, \pm 3\pi\}$; let $\nu_{2i} = g_{A_i}\#\text{Exp}_{\nu_{2\oplus}}[\Gamma(\text{Log}_{\nu_{1\oplus}}\nu_{1i})]$, with function g_a defined as per (38).

Step 3: Draw an i.i.d. sample of size m from each of the distributions $\{\nu_{1i}\}_{i=1}^n$ and $\{\nu_{2i}\}_{i=1}^n$.

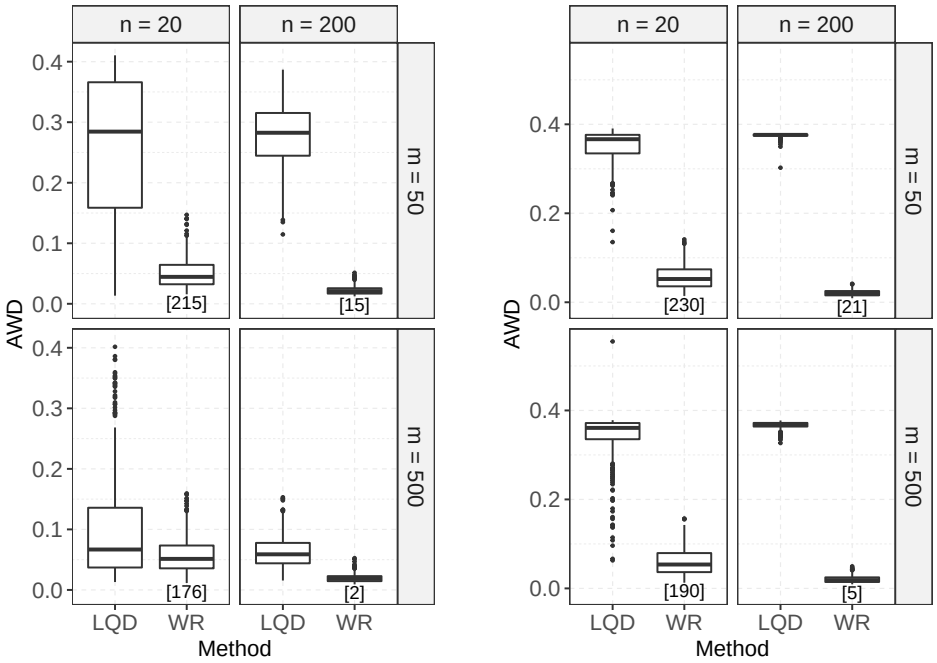
Four scenarios were considered with $n \in \{20, 200\}$ and $m \in \{50, 500\}$ for each case. We simulated 500 runs for each (n, m) pair. For the domain-extension method, the distribution domain is expanded from $[0, 1]$ to $[-0.5, 1.5]$ and $[-1, 2]$. To compare the three methods, we computed the out-of-sample average Wasserstein discrepancy (AWD) based on observations for 200 new predictors $\{\nu_{1i}\}_{i=n+1}^{n+200}$, for each Monte Carlo run. Denoting the fitted response distributions by $\nu_{2i}^\natural$, the out-of-sample AWD is given by

$$\text{AWD}(n, m) = \frac{1}{200}\sum_{i=n+1}^{n+200} d_W(\mathbb{E}_\oplus(\nu_{2i}|\nu_{1i}), \nu_{2i}^\natural), \tag{39}$$

with $\mathbb{E}_\oplus(\nu_{2i}|\nu_{1i})$ being the conditional Fréchet mean of ν_{2i} given ν_{1i} as defined above (38).

We found that the domain-extension method often failed to force the fit $\widehat{\Gamma}(\text{Log}_{\widehat{\nu}_{1\oplus}}\widehat{\nu}_{1i})$ to fall in the log space $\text{Log}_{\widehat{\nu}_{2\oplus}}\mathcal{W}$. In particular, this failure occurred in around 15–25% of the Monte Carlo runs where (35) happened when $n = 20$; therefore we do not report the results for this method. The results of the LQD method and the proposed Wasserstein regression method with boundary projection (WR) are summarized in the boxplots of Figure 1.

The proposed method outperforms the LQD method in all the scenarios considered. In fact, the log maps are isometries between the Wasserstein space and the log image spaces. This provides support for the proposed approach. In contrast, the LQD transformation is not an isometry and the ensuing



(a) Case 1. (b) Case 2.

Figure 1: Boxplots of the out-of-sample AWDs as per (39) for the four simulation setups with $(n, m) \in \{20, 200\} \times \{50, 500\}$, where "LQD" denotes the LQD method and "WR" denotes the proposed Wasserstein regression method. The numbers in brackets "[ ]" below the boxplots for WR indicate for how many runs event (35) happened and boundary projection became necessary.

distortions likely contribute to its inferior behavior. In particular, in Case 2 where the Fréchet mean distributions are beta distributions and the density functions are not bounded away from zero on D, the LQD method suffers from bias issues. When the number of distributions n increases, (35) is seen to happen less frequently and boundary projection is seldom needed when the sample size is large ($n = 200$).

Additional simulations illustrating the asymptotic result in Theorem 1, regarding the robustness of the proposed distribution-to-distribution regression method and comparing the proposed distribution-to-scalar regression method with a Gaussian process regression approach (**?**) can be found in Section S.5 in the Supplementary Material.

6 Applications

6.1 Mortality Data

There has been continuing interest in the nature of human longevity and the analysis of mortality data across countries and calendar years has provided some of the key data to study it (e.g., **????**). Of particular interest is how patterns of mortality of specific populations evolve over calendar time. Going beyond summary statistics such as life expectancy, viewing the entire age-at-death distributions as data objects is expected to lead to deeper insights into the secular evolution of human longevity and its dynamics. The Human Mortality Database (http://www.mortality.org) provides yearly life tables for 38 countries or areas, which yield histograms for the distributions of age-at-death. Smooth densities can then be obtained by applying local linear regression (**?**). We obtained these densities on the domain $[0, 100]$ (years of age).

In a first analysis, we focused on the $n = 32$ countries or areas for which data are available for the years 1983 and 2013. We applied the proposed distribution-to-distribution regression model with mortality distributions for an earlier year (1983) as the predictor and a later year (2013) as the response to compare the temporal evolution of age-at-death distributions among different countries. We show the leave-one-out prediction results together with the observed distributional predictors and responses for females in Figure 2 for Japan, Ukraine, Italy and the USA, which showcase different patterns of mortality change between 1983 and 2013. In addition to the graphical comparisons, Wasserstein discrepancies (WD) between the observed and leave-one-out predicted distributions are also listed. For all four countries, the observed and predicted distributions for 2013 are seen to be shifted to the right from the corresponding distributions in 1983, indicating increased longevity.

The top row of Figure 2 shows a comparison between the model anticipation and the actual observed distributions in 2013 in terms of density functions. Specifically, for Japan and the USA, the rightward mortality shift is seen to be more expressed than suggested by the leave-one-out prediction, indicating that longevity extension is more than anticipated, while the mortality distribution for Ukraine seems

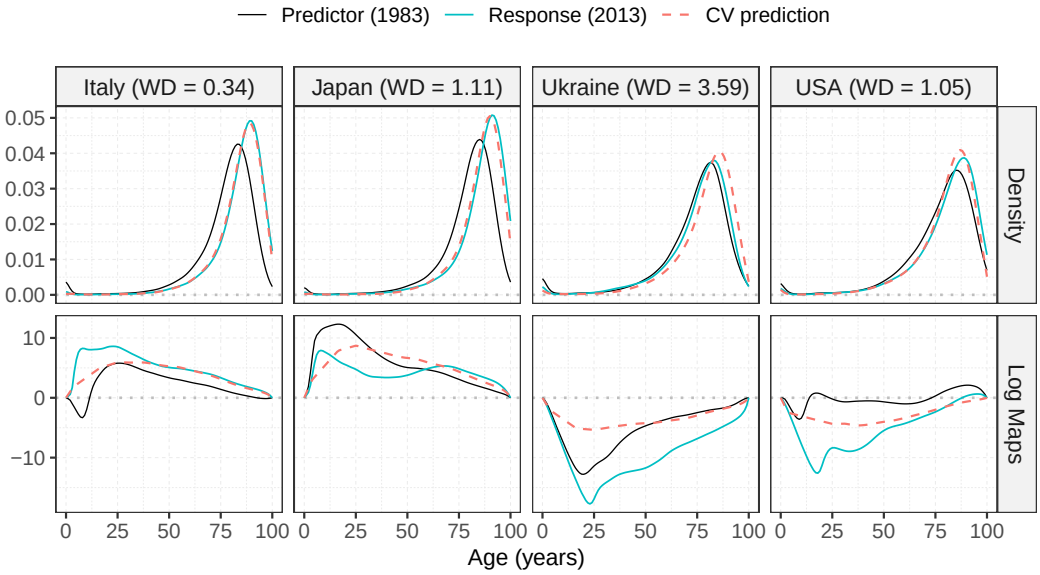


Figure 2: Age-at-death distributions of females in Italy, Japan, Ukraine, and the USA for 1983 and 2013, and the leave-one-out cross validation prediction based on the proposed distribution-to-distribution regression model, where the predictors are the distributions for 1983 and the responses are the distributions 30 years later. Top row: Observed densities for 1983 and 2013 and the leave-one-out predicted densities $\mathrm{Exp}_{\widehat{\nu}_{2\oplus}}(\widehat{\Gamma}(\mathrm{Log}_{\widehat{\nu}_{1\oplus}}\widehat{\nu}_{1i}))$ for 2013; Bottom row: Log-mapped predictors and responses, $\mathrm{Log}_{\widehat{\nu}_{1\oplus}}\widehat{\nu}_{1i}$ and $\mathrm{Log}_{\widehat{\nu}_{2\oplus}}\widehat{\nu}_{2i}$, and leave-one-out prediction for log responses $\widehat{\Gamma}(\mathrm{Log}_{\widehat{\nu}_{1\oplus}}\widehat{\nu}_{1i})$, where the estimated regression operator $\widehat{\Gamma}$ is defined in (22) and no boundary projection is needed for these four countries. The Wasserstein discrepancies (WDs) between the observed distributions and the corresponding leave-one-out prediction are indicated for each country.

to shift to the right at a slower pace than the model prediction would suggest, leading to a relatively large WD with a value of 3.59 between the observed and predicted response. In contrast, the regression fit for Italy almost perfectly matches the observed distribution in 2013.

The log maps shown in the bottom row of Figure 2 indicate the shifts of the distributions relative to the Fréchet mean across countries for the corresponding year. For Japan, the log maps for the observed predictors and responses and also the model prediction are all positive across the age domain, indicating that the distributions for Japan shift to the right from the Fréchet mean across countries, and Japanese females live longer compared to the average across countries at all the ages, while the magnitude of these log maps vary between 1983 and 2013 and also between observed and predicted distributions for 2013. The observed mortality distribution for 2013 has a bigger rightward shift relative

to the Fréchet mean distribution for older females and minors and a smaller one for younger adults than the model prediction. In contrast, Ukraine has a leftward shift from the Fréchet mean for females of all ages, and for 2013 the shift exceeds the model anticipation. For Italy, the log transformed predictor is negative before 15 and positive after, whence the predicted log response becomes positive throughout and also expands in size, meaning the relative standing of Italy in terms of longevity is anticipated to be improved in 2013 by the model prediction. The predicted distribution of Italy in 2013 is shifted to the right from the Fréchet mean for all ages, and such rightward shift is more expressed in the actual distribution in 2013. For the USA, the predicted log-mapped response for 2013 is entirely negative and consequently the mortality distribution moved to the left of the Fréchet mean, i.e., its relative standing in terms of longevity is anticipated to become worse, while the actual observation is a mixture of a rightward shift for more than 88 years of age and a leftward shift for the other ages.

We also illustrated the proposed autoregressive model for distribution-valued time series with the mortality data for Sweden, and the results are summarized in Section S.6 in the Supplementary Material.

6.2 House Price Data

A question of continuing interest to economists is how house prices change over time (e.g., **??**). We fitted the temporal evolution of house price distributions via the autoregressive distribution time series model described in Section 4, where we downloaded house price data from `http://www.zillow.com`. These data included bimonthly median house prices after inflation adjustment for $m = 306$ cities in the US from June 1996 to August 2015, for which the distribution of median house prices across the cities was constructed for every second month. The autoregressive model was trained on data up to April 2007 and predictions were computed for the remaining period, where we successively predicted the distribution of each month based on the prediction two months prior, i.e., by running the distribution time series model as estimated from the training period.

Figure 3 shows the fitting and prediction results for training and prediction periods, where selected months are ordered in time, while a five-number summary of the fitting and prediction WDs is given

in Table 1. The house price densities are found to be mostly uni-modal, and the peak shifts gradually to the right over time. Within the training period, the fitted densities are initially very close to the observed densities and then gradually are situated to the left of the observed densities, which means that the house price evolution overall accelerates during this period. For the prediction period, the predicted densities almost coincide with the observed distributions in 2007, fall behind the actual distribution in 2008, and then continue shifting to the right of the observed distributions. We find that the discrepancy between the predicted and observed house price distributions increases from 2007 to 2012 and then decreases afterwards. These findings are in line with the overheating of the housing market before 2006, the crash in 2007–2008, and the lingering effects of the financial crisis, followed by a recovery after 2012.

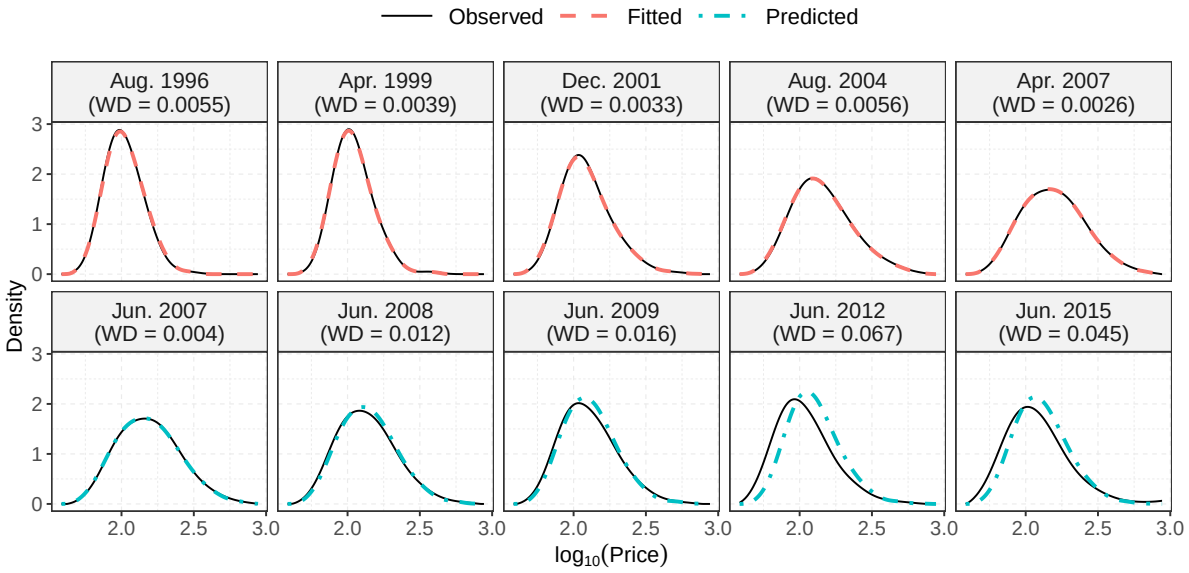


Figure 3: Observed and fitted (top row) / predicted (bottom row) densities of the house price distributions. Training period: August 1996 to April 2007. Prediction period: June 2007 to August 2015. Five representative months are depicted for each of the training and prediction periods in time order, where the Wasserstein discrepancies (WDs) are also listed.

Table 1: Five-number summary of the Wasserstein discrepancies in training and prediction periods.

	Min	$Q_{0.25}$	Median	$Q_{0.75}$	Max
Training	0.0020	0.0035	0.0047	0.0066	0.017
Prediction	0.0040	0.016	0.042	0.054	0.068