Journal of Machine Learning Research 27 (2026) 1-69 Submitted 10/24; Revised 1/26; Published 6/26

Sliced Wasserstein Regression

Han Chen* NAHCHEN@UCDAVIS.EDU
Department of Statistics
University of California, Davis
Davis, CA 95616, USA

Yidong Zhou* YDZHOU@UCDAVIS.EDU
Department of Statistics
University of California, Davis
Davis, CA 95616, USA

Hans-Georg Müller HGMUELLER@UCDAVIS.EDU
Department of Statistics
University of California, Davis
Davis, CA 95616, USA

Editor: Aryeh Kontorovich

Abstract

While statistical modeling of distributional data has gained increased attention, the case of multivariate distributions has been somewhat neglected despite its relevance in various applications. This is because the Wasserstein distance, commonly used in distributional data analysis, poses challenges for multivariate distributions. A promising alternative is the sliced Wasserstein distance, which offers a computationally simpler solution. We propose distributional regression models with multivariate distributions as responses paired with Euclidean vector predictors. The foundation of our methodology is a slicing transform from the multivariate distribution space to the sliced distribution space for which we establish a theoretical framework, with the Radon transform as a prominent example. We introduce and study the asymptotic properties of sample-based estimators for two regression approaches, one based on utilizing the sliced Wasserstein distance directly in the multivariate distribution space, and a second approach based on a new slice-wise distance, employing a univariate distribution regression for each slice. Both global and local Fréchet regression methods are deployed for these approaches and illustrated in simulations and through applications. These include the joint distribution of excess winter death rates and winter temperature anomalies in European countries as a function of base winter temperature, and also data from finance.

Keywords: Distributional data analysis, Multivariate distributional data, Radon transform, Slice-wise Wasserstein distance, Fréchet regression

1. Introduction

It is increasingly common for statisticians to encounter data that consist of samples of multivariate probability distributions. Examples include distributions of anthropometric measurements (Hron et al., 2023), joint distributions of stock price returns for multiple assets

∗. The first two authors contributed equally to this work.

or indices (Guégan and Iacopini, 2018), and joint distributions of systolic and diastolic blood pressure measurements (Fan and Müller, 2024). Distributional data differ fundamentally from functional data in that probability distributions do not form a linear vector space, limiting the direct applicability of classical functional data analysis tools.

Most existing methodology in distributional data analysis has focused on univariate distributions (Matabuena et al., 2021; Ghosal et al., 2023; Petersen et al., 2022). In contrast, regression methods for multivariate distributional responses remain comparatively underdeveloped (Dai, 2022), largely due to the computational and theoretical challenges associated with multivariate optimal transport and Wasserstein distances. In this paper, we study regression models that feature Euclidean predictors and multivariate distributional responses, a setting that naturally extends classical regression with scalar or vector outcomes and differs substantially from previously considered regression problems in which both predictors and responses are distributions.

For regression with Euclidean predictors and univariate distributional responses, several approaches embed distributions into Hilbert spaces, enabling the use of functional regression techniques. These include global transformations such as the log quantile density and log hazard transforms (Petersen and Müller, 2016), as well as constructions arising from Bayes Hilbert spaces in compositional data analysis (Hron et al., 2016; Menafoglio et al., 2018). When equipped with the Wasserstein metric, the space of univariate distributions also admits tangent space representations that have been used for principal component analysis and regression (Bigot et al., 2017; Chen et al., 2023a). However, such approaches rely on inverse mappings that are not globally well defined (Pegoraro and Beraha, 2022) and do not readily extend to general multivariate distributions.

A separate line of research considers regression problems in which both predictors and responses are distributions, typically aiming to learn optimal transport maps between random probability measures. Examples include transportation-based regression models (Ghodrati and Panaretos, 2022, 2023), autoregressive models for time series of distributions (Zhu and Müller, 2023) and distribution-on-distribution regression under multivariate Gaussian assumptions (Okano and Imaizumi, 2024). While these approaches address important problems, they target fundamentally different regression settings and are not directly applicable when predictors are Euclidean and only the responses are distribution-valued.

Existing approaches for regression with Euclidean predictors and multivariate distributional responses remain limited. Related work based on Bayes Hilbert space representations has primarily focused on the analysis and modeling of density functions rather than conditional regression. For example, Guégan and Iacopini (2018) considered functional representations of density-valued data in a Hilbert space setting, mainly in the context of temporal modeling, while Hron et al. (2023) proposed an orthogonal decomposition of bivariate densities into marginal and interaction components to study dependence structures. These approaches rely on explicit density estimation and linearization in Hilbert spaces and are not formulated as general regression frameworks with Euclidean predictors and multivariate distributional responses. Another related approach constructs conditional Wasserstein barycenters using entropy-regularized optimal transport (Fan and Müller, 2024), with a focus on interpolation and extrapolation of univariate and bivariate distributions. This approach relies on grid-based discretization of sample space, which suffers from the curse of dimensionality and becomes computationally infeasible for higher-dimensional distributions.

Fréchet regression (Petersen and Müller, 2019) provides a general and theoretically grounded framework for regression with Euclidean predictors and responses taking values in a metric space. When applied to distributional responses, Fréchet regression yields asymptotically consistent estimators for conditional Fréchet means. Existing theory and methodology (Zhou and Müller, 2024), however, are largely restricted to univariate distributions. Directly extending Fréchet regression to multivariate distributional responses would be challenging because it needs to overcome the computational and geometric difficulties inherent in multivariate optimal transport.

To address these challenges, the sliced Wasserstein distance (Bonneel et al., 2015) has emerged as a computationally efficient alternative and has gained widespread use in statistics and machine learning (Kolouri et al., 2016; Courty et al., 2017; Kolouri et al., 2019; Rustamov and Majumdar, 2023; Tanguy et al., 2024; Quellmalz et al., 2023; Zhang et al., 2024). Existing regression methods based on sliced Wasserstein distances have largely focused on different problem settings, such as regression with multivariate distributional inputs and scalar outputs (Meunier et al., 2022). These considerations motivate the development of a regression framework that combines the statistical foundations of Fréchet regression with the computational advantages of slicing-based Wasserstein distances. We propose a general slicing transform framework, with the Radon transform as a prominent example, that maps multivariate distributions into a collection of univariate distributions. Within this framework, we develop two regression approaches: a slice-averaged method that applies Fréchet regression in the multivariate distribution space equipped with a sliced Wasserstein distance, and a slice-wise method that performs Fréchet regression independently within each slice followed by reconstruction via a regularized inverse transform.

Both approaches come with theoretical guarantees on the convergence of the fitted regression functions under suitable regularity conditions. Our results extend univariate distributional regression theory to multivariate settings while explicitly accounting for the effects of the inverse slicing transform. We also provide practical algorithms, simulation studies, and real-data applications that demonstrate the utility of the proposed methods. The framework and its theoretical foundation that we provide in this paper has already been shown to be useful for applications beyond regression such as functional principal component analysis for multivariate distributions (Chen and Müller, 2025).

The remainder of the paper is organized as follows. Section 2 introduces slicing transforms with emphasis on the Radon transform, and Section 3 describes the associated slicing space. Section 4 presents the proposed regression models, with estimation procedures discussed in Section 5. Asymptotic convergence results are given in Section 6, and practical algorithms are described in Section 7. Simulation studies and data applications are reported in Sections 8 and 9, respectively, followed by a discussion in Section 10. Additional results and proofs are provided in the Appendix. The implementation details and code are publicly available at `https://github.com/yidongzhou/Sliced-Wasserstein-Regression`.

2. Slicing Transforms for Multivariate Distributions

2.1 Preliminaries

We denote by $\|\cdot\|_2$ the L^2 norm and by $\|\cdot\|_\infty$ the sup norm for vectors or functions and throughout use $C_0, C_1, \ldots$ to denote various constants and their dependence on relevant

quantities R will be indicated by writing $C_0(R), C_1(R), \ldots$. All notations are listed in Appendix A. We assume that the multivariate distributions under consideration possess density functions and share a common support D, assumed to be known or diligently chosen, usually from subject-matter considerations, where

(D1) The support $D \subset \mathbb{R}^p$ is compact and convex.

Denote the space of multivariate density functions on $\mathbb{R}^p$ with compact support D by

$$\mathcal{F} = \left\{ f \in L^1(\mathbb{R}^p) : f(z) \geqslant 0, \int_{\mathbb{R}^p} f(z)dz = 1, \text{ support } I(f) = D, f \text{ satisfies (F1)} \right\}, \quad (1)$$

where

(F1) There exists a constant $M_0 > 0$ and an integer $k \geqslant 2$ such that for all $f \in \mathcal{F}$, $\max\{\|f\|_\infty, \|1/f\|_\infty\} \leqslant M_0$ on D and f is continuously differentiable of order k on D and has uniformly bounded partial derivatives.

(F2) $k > p/2$.

(F3) $k \geqslant p + 1$.

Smoothness of higher order leads to faster decay rate of Fourier-transformed functions and Conditions (F2)–(F3) will be utilized to obtain rates of convergence. We also require a set

$$\mathcal{G} = \left\{ g \in L^1(\mathbb{R}) : g(u) \geqslant 0, \int_{\mathbb{R}} g(u)du = 1, g \text{ satisfies (D2) and (G1)} \right\}$$

of univariate density functions. Denoting the support of $g \in \mathcal{G}$ by $I(g)$, we require

(D2) For all $g \in \mathcal{G}$, the support $I(g)$ is compact and $\bigcup_{g\in\mathcal{G}} I(g)$ is bounded.

(G1) For an integer k as in (F1), there exists a constant $M_1 > 0$ such that for all $g \in \mathcal{G}$, $\max\{\|g\|_\infty, \|1/g\|_\infty\} \leqslant M_1$ on $I(g)$ and g is continuously differentiable of order k on $I(g)$ with uniformly bounded derivatives.

Throughout we utilize the unit sphere to denote a slicing parameter set in $\mathbb{R}^p$, $\Theta = \{z \in \mathbb{R}^p : \|z\|_2 = 1\}$. Defining the density slicing space Λ_Θ as a family of maps from Θ to $\mathcal{G}$,

$$\Lambda_\Theta = \left\{ \lambda : \Theta \mapsto \mathcal{G}, \int_\Theta \int_{\mathbb{R}} [\lambda(\theta)(u)]^2 \, dud\theta < \infty \right\},$$

Λ_Θ can be endowed with an L^2 metric,

$$d_2(\lambda_1, \lambda_2) = \left(\int_\Theta \int_{\mathbb{R}} (\lambda_1(\theta)(u) - \lambda_2(\theta)(u))^2 \, dud\theta \right)^{1/2}, \quad \text{for all } \lambda_1, \lambda_2 \in \Lambda_\Theta, \quad (2)$$

where we note that the integral is well defined because of the Cauchy-Schwarz inequality. Two maps λ_1, λ_2 will be considered to be identical if they coincide except on a set of measure zero.

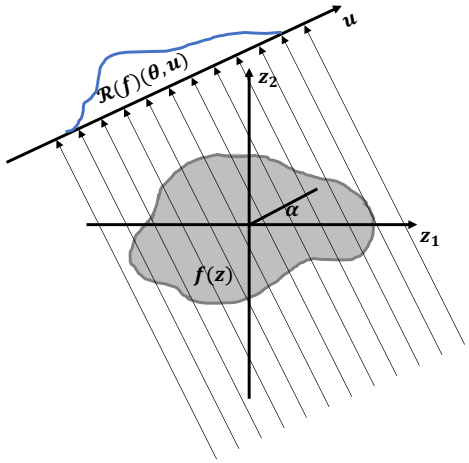


Figure 1: Schematic diagram for the two-dimensional Radon transform. For each unit vector $\theta(\alpha) = (\cos(\alpha), \sin(\alpha))$, the Radon transform integrates along the line $\langle z, \theta \rangle = u$ for each $u \in \mathbb{R}$.

2.2 Radon Transform and Slicing Transforms

The Radon transform $\mathcal{R}$ (Radon, 1917) is an integral transform, which maps an integrable p-dimensional function to the infinite set of its integrals over the hyperplanes of $\mathbb{R}^p$. Following the notation in Epstein (2007), let θ be a unit vector in Θ, u be an element in $\mathbb{R}$, and $l_{u,\theta}$ the affine hyperplane represented as $l_{u,\theta} = \{z \in \mathbb{R}^p : \langle z, \theta \rangle = u\}$. Employing an orthonormal basis $\{\theta, e_1, ..., e_{p-1}\}$ for $\mathbb{R}^p$ with $\langle \theta, e_j \rangle = 0$ and $\langle e_j, e_l \rangle = \delta_{jl}$, for $j, l = 1, ..., p-1$, the p-dimensional Radon transform $\mathcal{R} : \mathcal{F} \mapsto \Lambda_\Theta$ is defined through integrals over $l_{u,\theta}$,

$$\mathcal{R}(f)(\theta, u) = \int_{\mathbb{R}^{p-1}} f\left(u\theta + \sum_{j=1}^{p-1} s_j e_j\right) ds_1 \cdots ds_{p-1}, \quad \text{for } \theta \in \Theta \text{ and } u \in \mathbb{R}, \tag{3}$$

where we write $\mathcal{R}(f)(\theta, u)$ for $\mathcal{R}(f)(\theta)(u)$; see Figure 1 for a schematic illustration of the two-dimensional Radon transform. Since $l_{u,\theta}$ and $l_{-u,-\theta}$ are the same line, the Radon transform is an even function that satisfies $\mathcal{R}(f)(\theta, u) = \mathcal{R}(f)(-\theta, -u)$.

The inverse Radon transform is related to the Fourier transform. The p-dimensional Fourier transform $\mathcal{J}_p : L^1(\mathbb{R}^p) \mapsto \{f : \mathbb{R}^p \to \mathbb{C}\}$ is

$$\mathcal{J}_p(f)(\iota) = \int_{\mathbb{R}^p} e^{-i\langle z, \iota \rangle} f(z) dz, \quad \text{for all } \iota \in \mathbb{R}^p,$$

with the one-dimensional Fourier transform $\mathcal{J}_1 : L^1(\mathbb{R}) \mapsto \{f : \mathbb{R} \to \mathbb{C}\}$ given through the formula $\mathcal{J}_1(g)(r) = \int_{\mathbb{R}} g(u) e^{-iur} du$ for all $r \in \mathbb{R}$. The connection between the Fourier transform and the Radon transform is as follows.

Proposition 1 (Central Slicing Theorem (Bracewell, 1956)) *Assume $f \in \mathcal{F}$. For any real number r and a unit vector $\theta \in \Theta$, one has $\mathcal{J}_1(\mathcal{R}(f)(\theta))(r) = \mathcal{J}_p(f)(r\theta)$.*

The inverse Radon transform has been well investigated both theoretically and numerically (Abeida et al., 2012; Herman, 2009; Mersereau and Oppenheim, 1974). A common device to reconstruct the original multivariate function from its Radon transform is the filtered back-projection,

$$\mathcal{R}^{-1}(\lambda)(z) = \frac{1}{2(2\pi)^p} \int_\Theta \int_{\mathbb{R}} \mathcal{J}_1(\lambda(\theta))(r) e^{ir\langle\theta,z\rangle} |r|^{p-1} dr d\theta, \tag{4}$$

which satisfies $\mathcal{R}^{-1}(\mathcal{R}(f)) = f$ for any $f \in \mathcal{F}$ (Natterer, 2001; Epstein, 2007). The inversion formula can be decomposed into two steps. The first step acts as a high-pass filter, suppressing low-frequency components and amplifying high-frequency components. The second step implements an angular integral which can be interpreted as a back-projection of the filtered Radon transform (Epstein, 2007; Helgason, 2010). However, $\mathcal{R}^{-1}$ is not a continuous map, and small errors in the reconstruction of $\mathcal{R}(f)$ are amplified (Epstein, 2007). Therefore, regularization is usually applied when reconstructing the original function from its Radon transform (Horbelt et al., 2002; Qureshi et al., 2005; Shepp and Vardi, 1982), often through the use of a ramp filter that has a cut-off in the high-frequency domain (Epstein, 2007; Kak and Slaney, 2001).

Defining a regularizing function $\phi_\tau(r)$ with tuning parameter τ as $\phi_\tau(r) = 1$ for $|r| \leqslant \tau$ and $\phi_\tau(r) = 0$ for $|r| > \tau$, the regularized inverse map $\check{\mathcal{R}}_\tau^{-1} : \Lambda_\Theta \mapsto L^1(\mathbb{R}^p)$ is obtained by cutting off the high-frequency components in the filtered back-projection (4) through

$$\check{\mathcal{R}}_\tau^{-1}(\lambda)(z) = \frac{1}{2(2\pi)^p} \int_\Theta \int_{\mathbb{R}} \mathcal{J}_1(\lambda(\theta))(r) e^{ir\langle\theta,z\rangle} |r|^{p-1} \phi_\tau(r) dr d\theta. \tag{5}$$

As this regularized inverse is not guaranteed to be a multivariate density function, normalization is applied to map the resulting L^1 function into the multivariate density space $\mathcal{F}$ via

$$\mathcal{R}_\tau^{-1}(\lambda)(z) = \begin{cases} \left|\check{\mathcal{R}}_\tau^{-1}(\lambda)(z)\right| / \int_D \left|\check{\mathcal{R}}_\tau^{-1}(\lambda)(z)\right| dz & \text{if } \int_D \left|\check{\mathcal{R}}_\tau^{-1}(\lambda)(z)\right| dz > 0, \\ 1/|D| & \text{otherwise,} \end{cases} \tag{6}$$

where $|D|$ is the Lebesgue measure of the domain set D. Note that $\mathcal{R}_\tau^{-1}(\lambda)(z)$ satisfies the differentiability assumption in (F1) while the boundedness assumption can be enforced by projecting to the space where (F1) is satisfied.

If the Radon transform $\mathcal{R}(f)$ is approximated by $\widetilde{\mathcal{R}(f)}$, the corresponding reconstruction obtained via the regularized inverse Radon transform is $\widetilde{f}_\tau = \mathcal{R}_\tau^{-1} \circ \widetilde{\mathcal{R}(f)}$. We define the total reconstruction error as $\Delta_\tau = f - \widetilde{f}_\tau$. This error can be decomposed as

$$\Delta_\tau = \underbrace{f - \mathcal{R}_\tau^{-1} \circ \mathcal{R}(f)}_{\Delta_{\tau,1}} + \underbrace{\mathcal{R}_\tau^{-1} \circ \left(\mathcal{R}(f) - \widetilde{\mathcal{R}(f)}\right)}_{\Delta_{\tau,2}},$$

where $\Delta_{\tau,1}$ represents the error introduced by regularization of the inverse Radon transform, and $\Delta_{\tau,2}$ captures the effect of approximating the Radon transform.

Theorem 2 (Error bound for regularized Radon inversion) *Assume (D1), (F1), and (F3), and let $f \in \mathcal{F}$. As $d_2\left(\mathcal{R}(f), \widetilde{\mathcal{R}(f)}\right) \to 0$ and $\tau \to \infty$, the two error components satisfy*

$$\|\Delta_{\tau,1}\|_\infty = O\left(\tau^{-(k-p)}\right), \quad \|\Delta_{\tau,2}\|_\infty = O\left(\tau^p d_2\left(\mathcal{R}(f), \widetilde{\mathcal{R}(f)}\right)\right),$$

and consequently the total reconstruction error obeys

$$\|\Delta_\tau\|_\infty = O\left(\tau^{-(k-p)} + \tau^p d_2\left(\mathcal{R}(f), \widetilde{\mathcal{R}(f)}\right)\right).$$

The first error term $\Delta_{\tau,1}$ arises from the regularized inverse map and decreases with τ. The convergence of $\Delta_{\tau,1}$ relies on the smoothness assumption for densities, where higher order smoothness corresponds to a faster convergence rate of the Fourier transform in the frequency domain. The second error term $\Delta_{\tau,2}$ results from the approximation of $\mathcal{R}(f)$ and increases with τ. The value of the tuning parameter τ is then ideally chosen to minimize the total error Δ_τ. While the focus of this paper is primarily on the Radon transform, there are other transforms that may also be of interest such as the circular Radon transform (Kuchment, 2006) and the generalized Radon transform (Beylkin, 1984; Ehrenpreis, 2003); see Appendix F for further details.

Next, we consider a general *slicing transform* $\psi : \mathcal{F} \mapsto \Lambda_\Theta$ that maps a multivariate density function on D into a density slicing function indexed by the slicing parameter set Θ and satisfies

(T0) ψ is injective.

(T1) For a constant C_0, $d_2\left(\psi(f_1)(\theta), \psi(f_2)(\theta)\right) \leqslant C_0 d_2(f_1, f_2)$, for all $f_1, f_2 \in \mathcal{F}$ and $\theta \in \Theta$.

(T2) An inverse transform ψ^{-1} exists such that $\psi^{-1} \circ \psi(f) = f$, for all $f \in \mathcal{F}$.

(T3) There exists a sequence of approximating inverses ψ_τ^{-1} and constants $C_1(\tau)$, $C_2(\tau)$ such that

$$d_\infty\left(\psi_\tau^{-1} \circ \psi(f), f\right) \leqslant C_1(\tau) \text{ for all } f \in \mathcal{F}$$

and

$$d_\infty\left(\psi_\tau^{-1} \circ \psi(f), \psi_\tau^{-1}(\lambda)\right) \leqslant C_2(\tau) d_2\left(\psi(f), \lambda\right) \text{ for all } \lambda \in \Lambda_\Theta, f \in \mathcal{F} \text{ as } d_2\left(\psi(f), \lambda\right) \to 0,$$

where $C_1(\tau)$ and $C_2(\tau)$ depend only on τ, and $C_1(\tau)$ is decreasing to 0 while $C_2(\tau)$ is increasing as $\tau \to \infty$.

Assumption (T1) is concerned with the continuity of the forward transform, while Assumption (T2) provides the existence of an inverse transform from the image set $\psi(\mathcal{F})$ to $\mathcal{F}$. Note that $\psi(\mathcal{F})$ is not required to cover the entire space Λ_Θ and the inverse transform ψ^{-1} is only defined on the image space $\psi(\mathcal{F})$ of $\mathcal{F}$. The sequence of approximating inverses ψ_τ^{-1} in Assumption (T3) is required when mapping elements in Λ_Θ that are not in $\psi(\mathcal{F})$ back to $\mathcal{F}$. To be able to make use of a slicing transform for both forward and inverse transformations, we require an additional property and call a slicing transform *valid* if in addition to (T1)–(T3) it satisfies

(T4) Under (D1) and (F1), it holds that $\psi(\mathcal{F}) \subset \Lambda_\Theta$ and (D2), (G1) are satisfied for all $\lambda(\theta)$ for which $\lambda \in \psi(\mathcal{F})$.

Proposition 3 *The Radon transform $\mathcal{R}$ is a valid slicing transform, i.e., it satisfies Assumptions (T0)–(T4) where $C_1(\tau) = O(\tau^{-(k-p)})$ and $C_2(\tau) = O(\tau^p)$ as $\tau \to \infty$.*

3. Distribution Slicing and Sliced Distance

As a formal device, we introduce the map φ that assigns to any distribution its associated density function, and its inverse φ^{-1}, which maps a density function to the corresponding distribution. Recall from (1) that $\mathcal{F}$ denotes the space of multivariate density functions on $\mathbb{R}^p$ with compact support D, and $\mathcal{G}$ is the space of univariate density functions. We define $\mathscr{F} := \varphi^{-1}(\mathcal{F})$ and $\mathscr{G} := \varphi^{-1}(\mathcal{G})$ as the corresponding spaces of multivariate and univariate probability measures, respectively. Let G denote the map that assigns to each univariate distribution its quantile function, with G^{-1} as its inverse. We then define $\mathcal{Q} := G(\mathscr{G})$ as the space of quantile functions corresponding to univariate distributions in $\mathscr{G}$.

One has many choices for the metric d in the space $(\mathscr{F}, d)$ including the Fisher–Rao metric (Dai, 2022; Zhu and Müller, 2024) or the Wasserstein metric (Gibbs and Su, 2002), the latter being closely related to the optimal transport problem (Kantorovich, 2006; Villani, 2003). The L^2-Wasserstein metric between two distributions $\mu_1, \mu_2 \in \mathscr{F}$ is defined as

$$d_W^2(\mu_1, \mu_2) = \inf_{\substack{\pi \in \mathcal{P}(\mu_1, \mu_2) \\ (Z_1, Z_2) \sim \pi}} E\left(\|Z_1 - Z_2\|_2^2\right), \tag{7}$$

where $\mathcal{P}(\mu_1, \mu_2)$ is the set of joint probability measures on $D \times D$ with marginals μ_1 and μ_2, and Z_1, Z_2 are $\mathbb{R}^p$-valued random variables distributed according to μ_1 and μ_2, respectively.

For $p = 1$, the L^2-Wasserstein metric (7) can be equivalently expressed in terms of quantile functions,

$$d_W^2(\nu_1, \nu_2) = \int_0^1 \left(G^{-1}(\nu_1)(s) - G^{-1}(\nu_2)(s)\right)^2 ds, \quad \text{for all } \nu_1, \nu_2 \in \mathscr{G}.$$

When the dimension p of the random distribution is larger than 1, i.e., $p \geqslant 2$, one does not have an analytic form of the Wasserstein distance and the algorithms to obtain it are complex (Rabin et al., 2011; Peyré et al., 2019). The sliced Wasserstein distance (Bonneel et al., 2015) is a computationally more efficient alternative that utilizes the Radon transform. It has become increasingly popular for multivariate distributions due to its attractive topological and statistical properties (Kolouri et al., 2016; Nadjahi et al., 2019, 2020).

We define the quantile slicing space Γ_Θ as a family of maps from Θ to $\mathcal{Q}$,

$$\Gamma_\Theta = \left\{ \gamma : \Theta \mapsto \mathcal{Q}, \int_\Theta \int_{[0,1]} \gamma(\theta)(s)^2 ds d\theta < \infty \right\}$$

and introduce a map $\varrho : \Gamma_\Theta \mapsto \Lambda_\Theta$ as $\varrho(\gamma)(\theta) = \varphi \circ G^{-1}(\gamma(\theta)), \gamma \in \Gamma_\Theta$, which sends a quantile function $\gamma(\theta)$ to its corresponding density function. Similarly, ϱ^{-1} can be defined through $\varrho^{-1}(\lambda)(\theta) = G \circ \varphi^{-1}(\lambda(\theta))$, $\lambda \in \Lambda_\Theta$; see Figure 2 for a schematic illustration. A slicing transform ψ between $\mathcal{F}$ and Λ_Θ can be naturally extended to a transform $\widetilde{\psi}$ between $\mathscr{F}$ and Γ_Θ,

$$\widetilde{\psi}(\mu) = \varrho^{-1} \circ \psi \circ \varphi(\mu), \quad \mu \in \mathscr{F}. \tag{8}$$

The inverse transform $\widetilde{\psi}^{-1}$ and regularized inverse transform $\widetilde{\psi}_\tau^{-1}$ can be extended as

$$\widetilde{\psi}_\tau^{-1}(\gamma) = \varphi^{-1} \circ \psi_\tau^{-1} \circ \varrho(\gamma), \gamma \in \Gamma_\Theta \text{ and } \widetilde{\psi}^{-1}(\gamma) = \varphi^{-1} \circ \psi^{-1} \circ \varrho(\gamma), \gamma \in \widetilde{\psi}(\mathscr{F}). \tag{9}$$

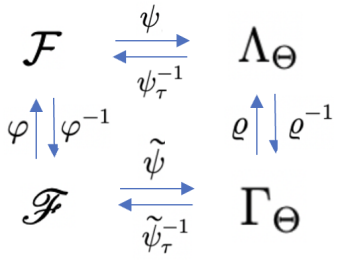


Figure 2: Schematic of maps involved in the slicing transform. Here φ maps the multivariate distribution space $\mathscr{F}$ to the multivariate density space $\mathcal{F}$ and ϱ maps the quantile slicing space Γ_Θ to the density slicing space Λ_Θ. A slicing transform ψ and its inverse ψ_τ^{-1} with a tuning parameter τ between $\mathcal{F}$ and Λ_Θ can be naturally extended to a slicing transform between $\mathscr{F}$ and Γ_Θ through the induced maps $\widetilde{\psi}$ and $\widetilde{\psi}_\tau^{-1}$.

Note that the space of Γ_Θ is closed and convex. Define the distribution-slicing Wasserstein metric on Γ_Θ through the aggregated Wasserstein distance across the slices

$$d_{DW}(\gamma_1, \gamma_2) = \left(\int_\Theta d_W^2 \left(G^{-1}(\gamma_1(\theta)), G^{-1}(\gamma_2(\theta)) \right) d\theta \right)^{1/2}$$
$$= \left(\int_\Theta \int_{[0,1]} \left(\gamma_1(\theta)(s) - \gamma_2(\theta)(s) \right)^2 ds d\theta \right)^{1/2}, \quad \gamma_1, \gamma_2 \in \Gamma_\Theta. \tag{10}$$

Here the integral is well defined because of the Cauchy-Schwarz inequality. We then define the slice-averaged Wasserstein distance as

$$d_{SW}(\mu_1, \mu_2) = d_{DW}(\widetilde{\psi}(\mu_1), \widetilde{\psi}(\mu_2)), \quad \text{for all } \mu_1, \mu_2 \in \mathscr{F}. \tag{11}$$

Proposition 4 *The slice-averaged Wasserstein distance is a distance over $\mathscr{F}$ if and only if (T0) is satisfied.*

4. Sliced Wasserstein Regression for Multivariate Distributions

A general approach for the regression of metric-valued responses and Euclidean predictors is Fréchet regression with both global and local versions (Petersen and Müller, 2019; Chen and Müller, 2022), where its application to distributional data has been extensively discussed in recent studies (Fan and Müller, 2022; Zhou and Müller, 2024). We extend Fréchet regression to the case of multivariate distributions by providing two types of sliced regression models. The first model applies Fréchet regression to the multivariate distribution space $\mathscr{F}$ equipped with the slice-averaged Wasserstein distance, while the second model applies a Fréchet regression step for each slice, followed by the regularized inverse transform.

Suppose $(X, \mu) \sim F$ is a random pair, where predictors X and responses μ take values in $\mathbb{R}^q$ and $\mathscr{F}$ and F is their joint distribution. Denote the marginal distributions of X and μ as F_X and F_μ respectively, and assume that mean $E(X)$ and variance $\mathrm{Var}(X)$ exist, where $\mathrm{Var}(X)$ is positive definite. The conditional distributions $F_{\mu|X}$ and $F_{X|\mu}$ are also

assumed to exist. Given any metric d on $\mathscr{F}$, Fréchet mean and Fréchet variance of the random distribution μ are defined as

$$\mu_{\oplus} = \underset{\omega \in \mathscr{F}}{\arg\min} E\left[d^2(\mu, \omega)\right], \quad V_{\oplus} = E\left[d^2(\mu, \mu_{\oplus})\right]. \tag{12}$$

The conditional Fréchet mean, i.e., the regression function of μ given $X = x$, targets

$$m(x) = \underset{\omega \in \mathscr{F}}{\arg\min}\, M(\omega, x), \quad M(\cdot, x) = E\left[d^2(\mu, \cdot)|X = x\right], \tag{13}$$

where $M(\cdot, x)$ is the conditional Fréchet function. The global Fréchet regression given $X = x$ is

$$m_G(x) = \underset{\omega \in \mathscr{F}}{\arg\min}\, M_G(\omega, x), \quad M_G(\cdot, x) = E\left[s_G(X, x) d^2(\mu, \cdot)\right], \tag{14}$$

where the weight $s_G(X, x) = 1 + (X - E(X))^\intercal \operatorname{Var}(X)^{-1}(x - E(X))$ is linear in x.

The proposed slice-averaged Wasserstein (SAW) regression employs Fréchet regression over the multivariate distribution space $\mathscr{F}$ equipped with the slice-averaged Wasserstein distance

$$m^{SAW}(x) = \underset{\omega \in \mathscr{F}}{\arg\min}\, M^{SAW}(\omega, x), \quad M^{SAW}(\cdot, x) = E\left[d^2_{SW}(\mu, \cdot)|X = x\right]. \tag{15}$$

The global slice-averaged Wasserstein (GSAW) regression given x is defined as

$$m_G^{SAW}(x) = \underset{\omega \in \mathscr{F}}{\arg\min}\, M_G^{SAW}(\omega, x), \quad M_G^{SAW}(\cdot, x) = E\left[s_G(X, x) d^2_{SW}(\mu, \cdot)\right]. \tag{16}$$

We note that for SAW based models the minimization is carried out over the space $\mathscr{F}$ and thus is automatically a multivariate distribution.

The second proposed model is the slice-wise Wasserstein (SWW) regression, where Fréchet regression is applied over the quantile slicing space Γ_Θ, equipped with distribution-slicing Wasserstein metric, followed by the regularized inverse transform,

$$m_\tau^{SWW}(x) = \widetilde{\psi}_\tau^{-1}\left[\underset{\gamma \in \Gamma_\Theta}{\arg\min}\, M^{SWW}(\gamma, x)\right], \quad M^{SWW}(\cdot, x) = E\left[d^2_{DW}(\widetilde{\psi}(\mu), \cdot)|X = x\right]. \tag{17}$$

The global slice-wise Wasserstein (GSWW) regression given $X = x$ is

$$m_{G,\tau}^{SWW}(x) = \widetilde{\psi}_\tau^{-1}\left[\underset{\gamma \in \Gamma_\Theta}{\arg\min}\, M_G^{SWW}(\gamma, x)\right], \quad M_G^{SWW}(\cdot, x) = E\left[s_G(X, x) d^2_{DW}(\widetilde{\psi}(\mu), \cdot)\right]. \tag{18}$$

Proposition 5 *Let $\gamma_{G,x} = \arg\min_{\gamma \in \Gamma_\Theta} M_G^{SWW}(\gamma, x)$, see (18). It can be characterized as*

$$\gamma_{G,x}(\theta) = \underset{\nu \in \mathscr{G}}{\arg\min} E\left[s_G(X, x) d^2_W\left(G^{-1}\left(\widetilde{\psi}(\mu)(\theta)\right), \nu\right)\right], \quad \textit{for almost all } \theta \in \Theta.$$

This result shows that the population SWW regression problem decouples across slices, allowing the minimization to be carried out independently for each projection direction θ. Consequently, SWW admits a practical slice-wise implementation while still targeting a population-level objective. Proposition 23 in Appendix G shows that the search space of SAW is a subset of that of SWW. In addition to the global Fréchet regression described above, we also consider local Fréchet regression variants for both SAW and SWW. In these local versions, the global weight function $s_G(X, x)$ is replaced by a localized weight function $s_L(X, x, h)$, which depends on a kernel function K and a bandwidth parameter $h > 0$. This modification allows the regression to adapt to local neighborhoods of the predictor space, giving higher weight to observations with predictors that are close to the target predictor level. The resulting methods are referred to as LSAW for the local version of SAW and LSWW for the local version of SWW. Further implementation details of these local variants are provided in Appendix D. Overall, the proposed methodology encompasses four regression approaches: GSWW, GSAW, LSWW, and LSAW.

5. Estimation

Suppose we have a sample of independent random pairs $\{(X_i, \mu_i)\}_{i=1}^n \sim F$. Define the sample mean and covariance as

$$\bar{X} = n^{-1} \sum_{i=1}^{n} X_i, \quad \hat{\Sigma} = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^\intercal.$$

When the random distributions μ_i are fully observed on domain D, sample-based estimators for GSAW and GSWW are given by

$$\check{m}_G^{SAW}(x) = \underset{\omega \in \mathscr{F}}{\arg\min}\, \check{M}_G^{SAW}(\omega, x), \quad \check{M}_G^{SAW}(\cdot, x) = n^{-1} \sum_{i=1}^{n} s_{iG}(x)\, d_{SW}^2(\mu_i, \cdot), \tag{19}$$

$$\check{m}_{G,\tau}^{SWW}(x) = \widetilde{\psi}_\tau^{-1}\left[\underset{\gamma \in \Gamma_\Theta}{\arg\min}\, \check{M}_G^{SWW}(\gamma, x) \right], \quad \check{M}_G^{SWW}(\cdot, x) = n^{-1} \sum_{i=1}^{n} s_{iG}(x)\, d_{DW}^2(\widetilde{\psi}(\mu_i), \cdot), \tag{20}$$

where $s_{iG}(x) = 1 + (X_i - \bar{X})^\intercal \hat{\Sigma}^{-1}(x - \bar{X})$ is the sample version of $s_G(X, x)$.

In many applications, the random distributions $\{\mu_i\}_{i=1}^n$ are not directly observed; instead, only samples generated from them are available. In such cases, we first estimate the densities $\hat{\mu}_i$ using kernel density estimation as described in Appendix C, which also lists the corresponding assumptions for this setting. The estimators in (19)–(20) then apply with μ_i replaced by $\hat{\mu}_i$, that is,

$$\hat{m}_G^{SAW}(x) = \check{m}_G^{SAW}(x)\big|_{\mu_i = \hat{\mu}_i}, \tag{21}$$

$$\hat{m}_{G,\tau}^{SWW}(x) = \check{m}_{G,\tau}^{SWW}(x)\big|_{\mu_i = \hat{\mu}_i}. \tag{22}$$

Sample-based estimators for the local smoothing models and bandwidth selection procedures are discussed in Appendix D.2.

A practical, data-driven method for selecting the tuning parameter τ in the presence of an i.i.d. sample of random pairs $\{(X_i, \mu_i)\}_{i=1}^n$ is to use leave-one-out cross-validation.

Specifically, we aim to minimize the discrepancy between predicted and observed distributions, given by

$$\hat{\tau} = \arg\min_{\tau} \sum_{i=1}^{n} d_{SW}^2(\mu_i, \hat{m}_{G,\tau,-i}^{SWW}(X_i)),$$

where $\hat{m}_{G,\tau,-i}^{SWW}(X_i)$ is the prediction at X_i from the GSWW regression of the ith-left-out sample $\{(X_{i'}, \mu_{i'})\}_{i' \neq i}$. When the sample size n exceeds 30, we substitute leave-one-out cross-validation with 5-fold cross-validation to strike a balance between computational efficiency and accuracy.

We define a slice-averaged Wasserstein R^2 coefficient to quantify the discrepancy between observed distributions and predicted distributions from the regression model as

$$R_{\oplus}^2 = 1 - \frac{E[d_{SW}^2(\mu, m(X))]}{E[d_{SW}^2(\mu, \mu_{\oplus})]},$$

where $m(x)$ is a regression function obtained through either SAW or SWW regression and $\mu_{\oplus}$ is the slice-averaged Wasserstein Fréchet mean $\mu_{\oplus} = \arg\min_{\omega \in \mathscr{F}} E\left[d_{SW}^2(\mu, \omega)\right]$, with empirical estimates

$$\hat{R}_{\oplus}^2 = 1 - \frac{\sum_{i=1}^{n} d_{SW}^2(\mu_i, \hat{m}(X_i))}{\sum_{i=1}^{n} d_{SW}^2(\mu_i, \hat{\mu}_{\oplus})}, \tag{23}$$

for a sample $\{(X_i, \mu_i)\}_{i=1}^n$. Here $\hat{\mu}_{\oplus} = \arg\min_{\omega \in \mathscr{F}} n^{-1} \sum_{i=1}^{n} d_{SW}^2(\mu_i, \omega)$ is the sample Fréchet mean and $\hat{m}(x)$ is the sample estimator of the regression function $m(x)$ for SAW or SWW regression.

6. Asymptotic Convergence

To study the convergence of SAW/SWW regression, we require a curvature condition at the true minimizer, as is commonly assumed for M-estimators. For Fréchet regression such a curvature condition has been established for the case of univariate distributions but not for multivariate distributions. Convexity assumptions have also been invoked in previous work (Fan and Müller, 2024; Boissard et al., 2015; Zhou and Müller, 2022). Specifically, we require

(A1) $\widetilde{\psi}(\mathscr{F})$ is closed and convex.

(A2) $\arg\min_{\gamma \in \Gamma_\Theta} M_G^{SWW}(\gamma, x)$ as per (18) is in $\widetilde{\psi}(\mathscr{F})$.

Assumption (A1) guarantees existence and uniqueness of the population minimizer for SAW regression, while Assumption (A2) ensures that the population SWW minimizer corresponds to a valid multivariate distribution under the inverse slicing transform. Although (A2) formalizes a compatibility requirement between slice-wise minimization and multivariate reconstruction, it is not unduly restrictive in common settings. In particular, Appendix I establishes that (A2) holds for multivariate Gaussian distributions, showing that the population SWW minimizer exactly recovers the true Gaussian distribution via its one-dimensional projections.

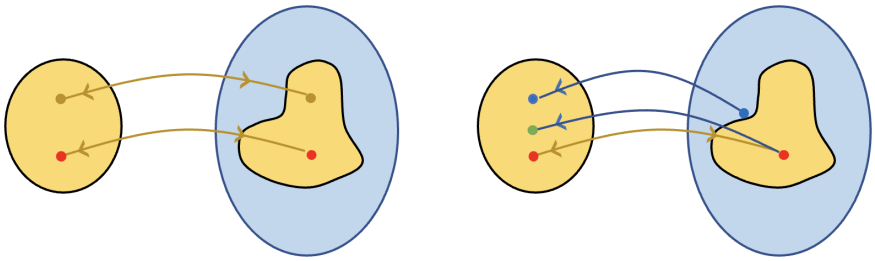


Figure 3: Diagram for SAW (left) and SWW (right) sample estimation. The red dots are population minimizers, the yellow dots are the sample SAW minimizer, the blue dots are the sample SWW minimizers and the green dot is the intermediate state used to facilitate the error analysis of SWW. Bidirectional yellow maps represent $\widetilde{\psi}$ and $\widetilde{\psi}^{-1}$ while directional blue maps represent the regularized inverse $\widetilde{\psi}_\tau^{-1}$.

The following results establish consistency and convergence rates for GSAW and GSWW under two scenarios: when densities are known and when they must be estimated from data. In the latter case, we additionally require assumptions (P1), (K1), (K2), and (F1′), which are stated in Appendix C together with details of the density estimation procedure. Corresponding results for LSAW and LSWW regression are provided in Theorems 14 and 16 for known densities, and in Theorems 15 and 17 when densities are estimated.

Theorem 6 (Convergence of GSAW regression) *Assume (D1), (F1)–(F2), (A1) and (T0)–(T2). For a fixed $x \in \mathbb{R}^q$, it holds for $m_G^{SAW}(x)$, $\check{m}_G^{SAW}(x)$ as per* (16), (19) *that*

$$d_{SW}\left(m_G^{SAW}(x), \check{m}_G^{SAW}(x)\right) = O_p(n^{-1/2}),$$
$$\sup_{\|x\|_E \leqslant B} d_{SW}\left(m_G^{SAW}(x), \check{m}_G^{SAW}(x)\right) = O_p(n^{-1/(2+\epsilon)}),$$

for a given $B > 0$ and any $\epsilon > 0$. When densities are estimated, the same rates hold for $\hat{m}_G^{SAW}(x)$ as per (21) *under (P1),* (F1′), *(K1)–(K2) in Appendix C.*

The GSAW thus yields a parametric convergence rate that does not depend on the dimension of the distribution space. Since the regularized inverse transform $\widetilde{\psi}_\tau^{-1}$ is continuous due to (T3), it is possible to obtain sup norm convergence for SWW. Under Assumption (A2), the population-level target m_G^{SWW} is

$$m_G^{SWW} = \widetilde{\psi}^{-1}\left[\operatorname*{arg\,min}_{\gamma \in \Gamma_\Theta} M_G^{SWW}(\gamma, x)\right], \tag{24}$$

where M_G^{SWW} is as in (18).

Theorem 7 (Convergence of GSWW regression) *Assume (D1), (F1), (F3), (A2), (T0)–(T4). For a fixed $x \in \mathbb{R}^q$, for $m_G^{SWW}(x)$, $\check{m}_{G,\tau}^{SWW}(x)$ as per* (24), (20), *$C_1(\tau)$ and $C_2(\tau)$ from (T3),*

$$d_\infty\left(m_G^{SWW}(x), \check{m}_{G,\tau}^{SWW}(x)\right) = O_p\left(C_1(\tau) + C_2(\tau)n^{-2/7}\right).$$

Furthermore, for a given $B > 0$ and any $\epsilon > 0$,

$$\sup_{\|x\|_E \leqslant B} d_\infty \left(m_G^{SWW}(x), \check{m}_{G,\tau}^{SWW}(x) \right) = O_p \left(C_1(\tau) + C_2(\tau) n^{-2/(7+\epsilon)} \right).$$

For estimated densities one obtains the same rates for $\hat{m}_{G,\tau}^{SWW}(x)$ as per (22) under (P1), (F1′), (K1)–(K2) in Appendix C.

The reconstruction error has two parts. The first term $C_1(\tau)$ is linked to the approximation bias in the reconstruction, while the second term $C_2(\tau)$ ensures that the approximation in the transformed space Γ_Θ is not excessively amplified (see Figure 3). If the slicing transform is the Radon transform, Corollary 8 shows that the curse of dimensionality is manifested in both $C_1(\tau)$ and $C_2(\tau)$, as the required order of smoothness increases with the dimension of the distribution to achieve the same convergence rate. We only provide the more intricate scenario when densities are not fully observed; analogous results hold for the case of known densities.

Corollary 8 *When taking the Radon transform $\mathcal{R}$ and the corresponding regularized inverse $\mathcal{R}_\tau^{-1}$ as per (3) and (6), under the assumptions of Theorem 7,*

$$d_\infty \left(m_G^{SWW}(x), \hat{m}_{G,\tau}^{SWW}(x) \right) = O_p \left(\tau^{-(k-p)} + \tau^p n^{-2/7} \right),$$
$$\sup_{\|x\|_E \leqslant B} d_\infty \left(m_G^{SWW}(x), \hat{m}_{G,\tau}^{SWW}(x) \right) = O_p \left(\tau^{-(k-p)} + \tau^p n^{-2/(7+\epsilon)} \right),$$

and with $\tau \sim n^{2/7k}$,

$$d_\infty \left(m_G^{SWW}(x), \hat{m}_{G,\tau}^{SWW}(x) \right) = O_p \left(n^{-2\frac{k-p}{7k}} \right),$$
$$\sup_{\|x\|_E \leqslant B} d_\infty \left(m_G^{SWW}(x), \hat{m}_{G,\tau}^{SWW}(x) \right) = O_p \left(n^{-2\frac{k-p}{7k+\epsilon}} \right).$$

7. Numerical Algorithm

7.1 SAW Regression

To solve problem (21) in practice, we use a numerical optimization process proposed in Bonneel et al. (2015) that involves parametrizing a probability measure with equal weights. Specifically, we use random observations $\mathbf{W} = \{\mathbf{W}_j\}_{j=1}^N \in \mathbb{R}^{p \times N}$ where each observation $\mathbf{W}_j \in \mathbb{R}^p$ follows a distribution characterized by a density function $f_{\mathbf{W}}$ in $\mathcal{F}$. We introduce the Radon slicing operation $\langle \cdot, \theta \rangle : \mathbb{R}^p \mapsto \mathbb{R}$ for each random observation, resulting in a univariate distribution with the density function $\mathcal{R} \circ f_{\mathbf{W}}(\theta)$. The multivariate distribution μ corresponding to the density $f_{\mathbf{W}}$ can be represented as a discrete input measure $\mu_{\mathbf{W}} = \frac{1}{N} \sum_{j=1}^N \delta_{\mathbf{W}_j}$. Similarly, we represent the i-th distribution μ_i as a discrete input measure $\mu_{\mathbf{W}^{(i)}} = \frac{1}{N} \sum_{j=1}^N \delta_{\mathbf{W}_{ij}}$ where $\mathbf{W}^{(i)} = \{\mathbf{W}_{ij}\}_{j=1}^{n_i} \in \mathbb{R}^{p \times N}$ and $\mathbf{W}_{ij} \sim \mu_i$. The GSAW regression of (21) given $X = x$ can be represented as

$$\underset{\mathbf{W} \in \mathbb{R}^{p \times N}}{\arg\min}\, \mathcal{M}_G(\mathbf{W}, x) = n^{-1} \sum_{i=1}^n \left[s_{iG}(x) d_{SW}^2(\mu_{\mathbf{W}^{(i)}}, \mu_{\mathbf{W}}) \right]. \tag{25}$$

It is known that the target function is smooth for the Radon transform, a result that we state below for reference. We use the notation $\mathbf{W}(\theta) = (\langle \mathbf{W}_j, \theta \rangle)_{j=1}^N \in \mathbb{R}^N$ and similarly $\mathbf{W}^{(i)}(\theta) = (\langle \mathbf{W}_{ij}, \theta \rangle)_{j=1}^N \in \mathbb{R}^N$. For any $A = \{A_j\}_{j=1}^N \in \mathbb{R}^N$ where $A_j \in \mathbb{R}$, Π_A is a permutation operator on A, such that $\Pi_A(A) = (A_{\Pi(1)}, A_{\Pi(2)}, ..., A_{\Pi(N)})^\intercal$ with $A_{\Pi(1)} \leqslant A_{\Pi(2)} \leqslant \cdots \leqslant A_{\Pi(N)}$.

Proposition 9 (Theorem 1 (Bonneel et al., 2015)) *For each fixed x and $N_i \equiv N$, we have $\mathcal{M}_G(\mathbf{W}, x) : \mathbb{R}^{p \times N} \mapsto \mathbb{R}$ is an L^1 function with a uniformly ρ_G-Lipschitz gradient for some $\rho_G > 0$ given by*

$$\nabla \mathcal{M}_G(\mathbf{W}, x) = n^{-1} \sum_{i=1}^n \left[s_{iG}(x) \int_\Theta \theta \left(\mathbf{W}(\theta) - \Pi^{-1}_{\mathbf{W}(\theta)} \circ \Pi_{\mathbf{W}^{(i)}(\theta)} \circ \mathbf{W}^{(i)}(\theta) \right)^\intercal d\theta \right].$$

The operation $\Pi^{-1}_{\mathbf{W}(\theta)} \circ \Pi_{\mathbf{W}^{(i)}(\theta)} \circ \mathbf{W}^{(i)}(\theta)$ aligns $\mathbf{W}^{(i)}(\theta)$ with $\mathbf{W}(\theta)$ in the sense of non-decreasing order for calculating the empirical Wasserstein distance. In practice, an equidistant grid $(\theta_1, \theta_2, ..., \theta_L)$ along the angular coordinate θ is used to approximate the integration over Θ. Note that (25) is non-convex and we use gradient descent to find a stationary point; see Algorithm 1.

Since the algorithm implied by Proposition 9 requires that the sample sizes N_i at which each distribution is sampled are identical, we set $N = \min_{i=1,...,n} N_i$ in step 2 of the following algorithm. In instances where $N_i > N$, we choose a randomly selected subsample of size N, referred to as downsampling in the following. For an additional discussion about the local modeling approach for SAW, we refer to Appendix D.4.

Algorithm 1 GSAW Algorithm when using the Radon Transform

1: Initialize a grid $(\theta_1, \theta_2, ..., \theta_L)$ along Θ
2: Set $N = \min_{i=1,...,n} N_i$, convergence threshold ε and learning rate η
3: For each $\mu_{\mathbf{W}^{(i)}}$, downsample $\mathbf{W}^{(i)}$ such that all $\mathbf{W}^{(i)} \in \mathbb{R}^{p \times N}$
4: Initialize $\mathbf{W}^{[0]} \in \mathbb{R}^{p \times N}$ arbitrarily and fix the output predictor $X = x$
5: **repeat**
6: Calculate $\nabla \mathcal{M}_G(\mathbf{W}^{[k]}, x)$

$$\nabla \mathcal{M}_G(\mathbf{W}^{[k]}, x) = (nL)^{-1} \sum_{i=1}^n \sum_{l=1}^L s_{iG}(x) \theta_l \left(\mathbf{W}(\theta_l) - \Pi^{-1}_{\mathbf{W}(\theta_l)} \circ \Pi_{\mathbf{W}^{(i)}(\theta_l)} \circ \mathbf{W}^{(i)}(\theta_l) \right)^\intercal$$

7: $\mathbf{W}^{[k+1]} = \mathbf{W}^{[k]} - \eta \nabla \mathcal{M}_G(\mathbf{W}^{[k]}, x)$
8: **until** Algorithm converges with $\left\| \mathbf{W}^{[k+1]} - \mathbf{W}^{[k]} \right\|_2 / \left\| \mathbf{W}^{[k]} \right\|_2 < \varepsilon$ to $\mathbf{W}^{[\infty]}$
9: Consider each column of $\mathbf{W}^{[\infty]}$ as a sample from $\hat{m}_G^{SAW}(x)$ and apply the kernel density estimator (27) in the Appendix to derive the density estimator $\hat{f}$

7.2 SWW Regression

Since SWW regression is split into completely separate slice-wise regression steps, one can leverage parallel computing to enhance computing efficiency; the permutation operator is

not needed. We note that the local model for the SWW approach can be implemented analogously simply by using local instead of global Fréchet regression in the algorithm.

Algorithm 2 GSWW Algorithm for the Radon Transform

1: Initialize a grid $(\theta_1, \theta_2, ..., \theta_L)$ along Θ and fix the output predictor $X = x$
2: Perform a Radon slicing operation over random observations, i.e. $\langle \mathbf{W}^{(i)}, \theta_l \rangle$, $i = 1, ..., n, l = 1, ..., L$
3: **for** $l = 1, ..., L$ **do** in parallel
4: Apply the Fréchet regression for sliced random pairs, $\big(X_i, \langle \mathbf{W}^{(i)}, \theta_l \rangle\big)$, $i = 1, ..., n$ (Petersen and Müller, 2019; Chen et al., 2023b)
5: Calculate the fitted density on the output $X = x$ as $\hat{\lambda}(\theta_l)$
6: **end for**
7: Apply the regularized Radon reconstruction $\hat{f} = \mathcal{R}_\tau^{-1}(\hat{\lambda})$ through (6)

8. Simulation Study

To evaluate the finite-sample performance of the proposed methods, we design a generative framework that produces random p-dimensional probability distributions μ, together with a scalar predictor X from a uniform distribution on $[-0.5, 0.5]$. We begin with multivariate Gaussian responses, where both the mean vector and covariance matrix depend on X. Specifically, given $X = x$, the mean vector is generated as $\mathcal{N}(\alpha(x), I_p)$, with $\alpha(x) \in \mathbb{R}^p$, and the covariance matrix is sampled from a Wishart distribution $\mathcal{W}_p(\mathscr{D}(x), p+1)$, with degrees of freedom $p + 1$ and scale matrix $\mathscr{D}(x) \in \mathbb{R}^{p \times p}$.

To generate more complex non-Gaussian responses, we apply a random transport map to the Gaussian distributions constructed above. Specifically, each coordinate of the multivariate distribution is pushed forward by a map T, chosen uniformly at random from $T_k(z) = z - \sin(kz)/|k|$ for $k \in \{-2, -1, 1, 2\}$ (Panaretos and Zemel, 2016). The map T is non-decreasing and has expectation equal to the identity. This transformation preserves the true regression function while significantly complicating the distributional structure, yielding non-Gaussian multivariate distributional responses.

We consider both Gaussian and non-Gaussian distributional settings, each with two different dimensions ($p = 2$ and $p = 5$), and further distinguish between global and local regression models. This results in a total of eight simulation settings. Table 1 summarizes all cases: Settings I–IV correspond to Gaussian responses with parameterizations depending on $X = x$, while Settings V–VIII represent their non-Gaussian counterparts, obtained by applying the random transport map T to the Gaussian distributions.

The proposed global (GSWW, GSAW) and local (LSWW, LSAW) regression methods were evaluated separately under each of the eight simulation settings. For comparison, we included two existing approaches for regression with distributional responses. The first comparison method, denoted as FM (Fan and Müller, 2024), is based on the Sinkhorn distance (Cuturi, 2013) and focuses on interpolation and extrapolation of univariate and bivariate distributions. The second comparison method, denoted as HMM (Hron et al., 2023), is based on a Bayes-Hilbert-space representation of bivariate densities using centered log-ratio transformations and spline approximations. Both FM and HMM are applicable only in the

Table 1: Simulation settings. Settings I–IV correspond to multivariate Gaussian distributional responses with distributional parameters depending on $X = x$ as indicated. Settings V–VIII are identical to I–IV, but with an additional random transport map T, sampled from $\{T_k(z) = z - \sin(kz)/k : k \in \{-2, -1, 1, 2\}\}$, applied to generate non-Gaussian multivariate distributional responses.

	Distribution	Dimension	Regression	Setting
I	Gaussian	$p = 2$	GSWW, GSAW	$\alpha(x) = (x, x)^\intercal$, $\mathscr{D}(x) = \mathrm{diag}(x+1, x+1)$
II			LSWW, LSAW	$\alpha(x) = \frac{1}{2}(\sin(\frac{\pi}{2}x), \frac{1}{2}\sin(\frac{\pi}{2}x))^\intercal$, $\mathscr{D}(x) = \mathrm{diag}(\cos(\frac{\pi}{2}x), \cos(\frac{\pi}{2}x))$
III		$p = 5$	GSWW, GSAW	$\alpha(x) = (x, \ldots, x)^\intercal$, $\mathscr{D}(x) = \mathrm{diag}(x+1, \ldots, x+1)$
IV			LSWW, LSAW	$\alpha(x) = (\frac{1}{2}\sin(\frac{\pi}{2}x), \ldots, \frac{1}{2}\sin(\frac{\pi}{2}x))^\intercal$, $\mathscr{D}(x) = \mathrm{diag}(\cos(\frac{\pi}{2}x), \ldots, \cos(\frac{\pi}{2}x))$
V–VIII	non-Gaussian	$p = 2, 5$	GSWW, GSAW; LSWW, LSAW	Same as Settings I–IV but with a random transport map T

bivariate setting ($p = 2$): FM becomes computationally infeasible in higher dimensions due to its grid-based discretization (e.g., a 50^5 grid for $p = 5$ would require excessive memory and runtime), while HMM is inherently designed for bivariate distributions. Accordingly, comparisons with these methods were restricted to the bivariate case.

Simulations were conducted for sample sizes $n \in \{50, 100, 200\}$. For each distribution, $N = 200$ observations were generated per distribution for $p = 2$ and $N = 500$ for $p = 5$. Each configuration was replicated 100 times, and estimation accuracy for the kth Monte Carlo run was quantified by the integrated squared error (ISE):

$$\mathrm{ISE}_k = \int_{-0.5}^{0.5} d_{SW}^2 \left(\hat{m}_k(x), m(x)\right) dx, \tag{26}$$

where $m(x)$ is the true regression function and $\hat{m}_k(x)$ the fitted regression function from the kth run. For the local methods, bandwidths were set to $0.25n^{-1/5}$.

The results for the proposed SWW and SAW approaches and the comparison methods are reported in Table 2. Across all considered settings, the proposed SWW and SAW methods yield considerably smaller mean ISE values than both comparison methods FM and HMM. This performance gain reflects the effectiveness of slicing in capturing distributional variation through one-dimensional projections, which is particularly advantageous for complex multivariate distributional responses. As the sample size n increases, both SWW and SAW exhibit clear convergence behavior, while FM and HMM show comparatively limited improvement. For FM, this behavior can be attributed to approximation errors induced by grid-based discretization and entropic regularization. In contrast, the reduced performance of HMM is likely due to its reliance on clr-based linear representations and spline smoothing, which do not explicitly capture the transport geometry underlying the distributional responses.

Table 2: Simulation results for three sample sizes across eight settings using the proposed regression methods (SWW and SAW) and comparison methods FM (Fan and Müller, 2024) and HMM (Hron et al., 2023). Each cell reports the mean (standard deviation) of the integrated squared error over 100 replications. FM is reported only for the bivariate case ($p = 2$) due to computational infeasibility in higher dimensions, while HMM is only applicable to bivariate distributions. Settings I–IV correspond to Gaussian responses, while Settings V–VIII correspond to non-Gaussian responses generated via a random transport map T.

		Distribution	Dimension	n	SWW	SAW	FM	HMM
I	global	Gaussian	$p = 2$	50	0.080 (0.035)	0.109 (0.032)	0.460 (0.070)	0.652 (0.071)
				100	0.053 (0.021)	0.086 (0.019)	0.438 (0.059)	0.637 (0.052)
				200	0.037 (0.012)	0.073 (0.011)	0.416 (0.031)	0.630 (0.037)
III			$p = 5$	50	0.076 (0.020)	0.115 (0.023)	—	—
				100	0.041 (0.010)	0.088 (0.016)	—	—
				200	0.027 (0.006)	0.076 (0.015)	—	—
V		non-Gaussian	$p = 2$	50	0.081 (0.037)	0.109 (0.034)	0.515 (0.126)	0.734 (0.097)
				100	0.047 (0.021)	0.081 (0.019)	0.481 (0.101)	0.709 (0.077)
				200	0.031 (0.012)	0.067 (0.010)	0.436 (0.049)	0.698 (0.055)
VII			$p = 5$	50	0.068 (0.019)	0.110 (0.020)	—	—
				100	0.033 (0.009)	0.081 (0.010)	—	—
				200	0.019 (0.005)	0.071 (0.010)	—	—
II	local	Gaussian	$p = 2$	50	0.150 (0.054)	0.126 (0.035)	0.448 (0.063)	0.588 (0.066)
				100	0.086 (0.024)	0.094 (0.019)	0.414 (0.045)	0.575 (0.050)
				200	0.056 (0.014)	0.076 (0.011)	0.393 (0.026)	0.567 (0.033)
IV			$p = 5$	50	0.144 (0.027)	0.131 (0.023)	—	—
				100	0.074 (0.016)	0.092 (0.016)	—	—
				200	0.046 (0.007)	0.077 (0.014)	—	—
VI		non-Gaussian	$p = 2$	50	0.163 (0.062)	0.131 (0.038)	0.505 (0.107)	0.644 (0.096)
				100	0.086 (0.026)	0.092 (0.020)	0.467 (0.093)	0.616 (0.084)
				200	0.055 (0.015)	0.073 (0.011)	0.422 (0.049)	0.594 (0.054)
VIII			$p = 5$	50	0.135 (0.024)	0.125 (0.019)	—	—
				100	0.068 (0.015)	0.086 (0.009)	—	—
				200	0.039 (0.006)	0.074 (0.009)	—	—

Between the two proposed methods, SWW consistently outperforms SAW, especially for larger sample sizes. This advantage can be attributed to the fact that SWW minimizes a convex objective function independently within each slice, ensuring a unique minimizer, whereas SAW relies on gradient-based optimization of a global non-convex objective over the multivariate distribution space. A visual comparison of the true and fitted bivariate densities obtained by SWW and SAW is provided in Figure 11 in Appendix J, where SWW produces density estimates that are visually more accurate and better aligned with the true distributions.

Runtime comparisons for the proposed methods (SWW and SAW) and the comparison methods FM and HMM are reported in Table 3. All experiments were conducted on a 16-inch MacBook Pro equipped with an Apple M3 Max chip (14 CPU cores and 36 GB unified memory) running macOS Tahoe 26.2, with parallelization configured to use 10 CPU cores. As shown in the table, FM is substantially more computationally demanding than all other methods in the bivariate case, reflecting the cost of repeated Sinkhorn iterations

Table 3: Runtime (in seconds) for the proposed regression methods (SWW and SAW) and comparison methods (FM) (Fan and Müller, 2024) and HMM (Hron et al., 2023) across different dimensions and sample sizes. FM is reported only for $p = 2$ due to computational constraints, while HMM applies only to bivariate distributions.

	Dimension	n	SWW	SAW	FM	HMM
global	$p = 2$	50	0.80	10.32	43.88	0.76
		100	1.04	16.94	107.76	1.28
		200	1.21	35.60	386.55	2.50
	$p = 5$	50	2.75	40.93	—	—
		100	3.31	78.64	—	—
		200	6.79	153.07	—	—
local	$p = 2$	50	0.71	9.37	45.64	0.64
		100	1.02	18.89	83.40	1.22
		200	1.29	35.20	399.69	2.40
	$p = 5$	50	2.05	41.27	—	—
		100	2.41	79.70	—	—
		200	4.75	153.83	—	—

on fine grids, and becomes infeasible for higher-dimensional distributions. In contrast, HMM exhibits comparatively low runtime in the bivariate setting, but this computational efficiency comes at the cost of reduced estimation accuracy, as observed in Table 2.

Among the proposed methods, SWW is consistently the most computationally efficient across all settings, while SAW incurs a substantially higher computational cost due to the need to solve a global non-convex optimization problem via gradient-based methods. The runtime of SAW increases rapidly with the sample size n and the distributional dimension p, whereas the runtime of SWW grows more moderately. These empirical findings are consistent with the theoretical and algorithmic differences between the two approaches and with existing analyses of computational complexity for bivariate Wasserstein barycenters based on Sinkhorn and sliced Wasserstein distances (Bonneel et al., 2015).

9. Data Analysis

9.1 Excess Winter Deaths in Europe

Excess winter deaths are a social and health challenge. For European countries, they have become an acute problem due to rapidly rising heating costs. It is known that in general Northern European countries have lower excess winter mortality rates compared to Southern European countries (Healy, 2003; Fowler et al., 2015). Our goal in this analysis was to follow up on this by modeling conditional distributions directly, rather than characterizing mortality effects through summary statistics such as sample mean or standard deviation, as was done in previous studies. Simply applying the conventional regression model to the scatter plot between excess winter mortality rates and the absolute winter temperature

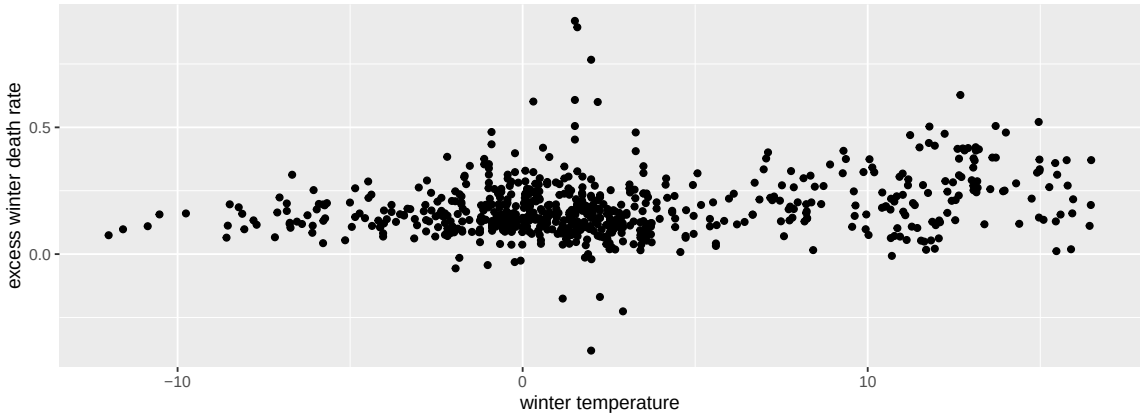


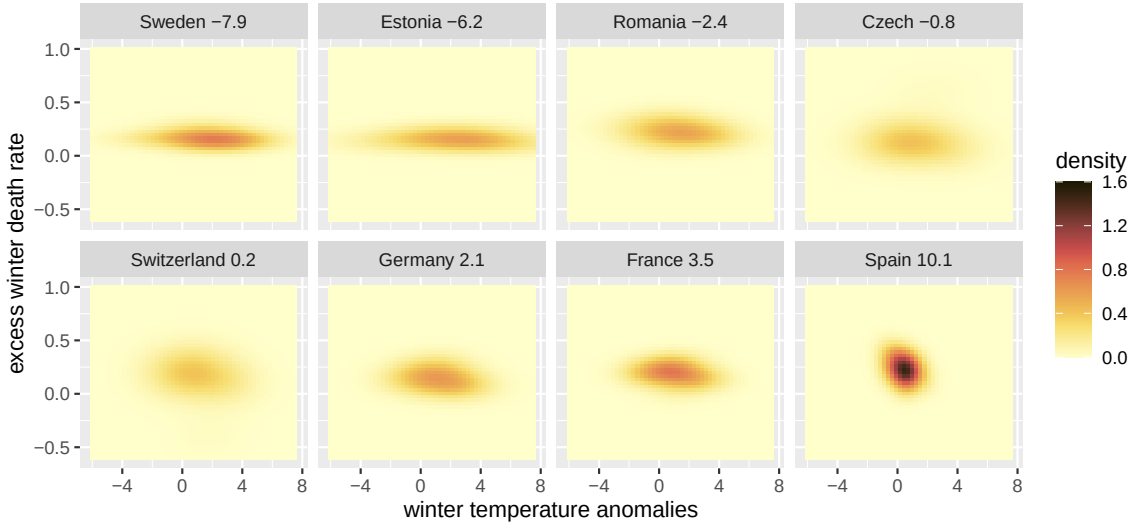
Figure 4: Scatterplot of excess winter death rates (y-axis) against winter temperature (x-axis)

(see Figure 4) makes the analysis susceptible to Simpson's paradox, creating a misleading perception of an association between increased excess winter death rates and higher winter temperature. A better approach is to separate the observations by country.

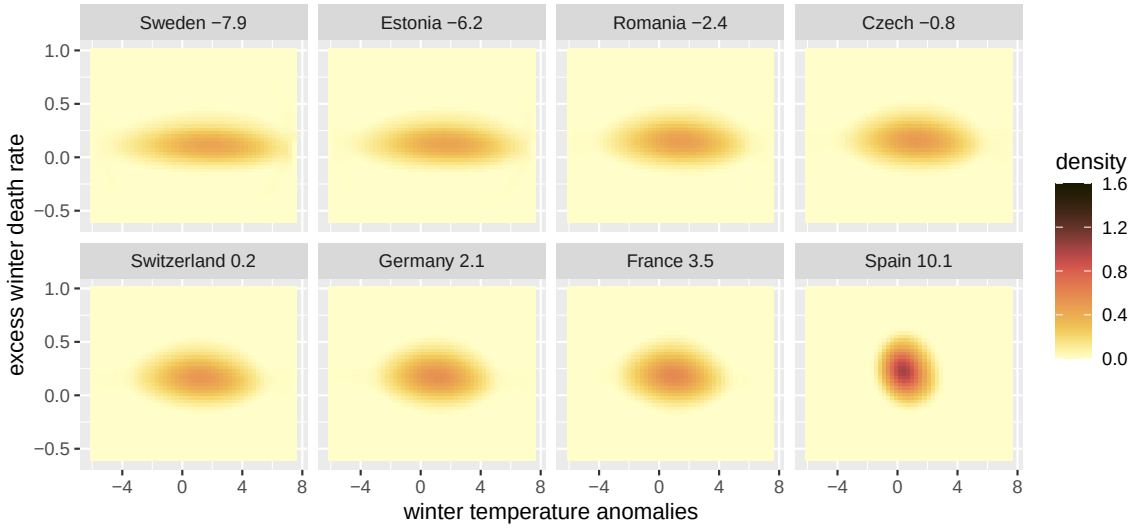
We focused on $n = 31$ countries with available data from 1999 to 2021, for which we adopted the proposed GSWW and GSAW models with the country-level base winter temperature averaged from 1961-90 as predictor and the country-level bivariate distribution between the excess rate of winter mortality relative to the previous summer and winter temperature anomaly relative to the base winter temperature from 1961-90 as the response. We chose standardized winter temperature anomalies to ensure better overlap of shared subdomains within country-specific bivariate distributions. The absolute number of deaths in each of the 31 European countries from 1999 to 2021 was obtained from the Eurostat database `https://ec.europa.eu/eurostat/web/main/data/database` and the winter temperature anomalies and base winter temperatures from 1961-90 from `https://www.uea.ac.uk/groups-and-centres/climatic-research-unit/data`. The observed smoothed bivariate distributions are in Figure 5(a) for a few selected countries.

The fitted densities obtained from the GSWW model are shown in Figure 5(b), where the corresponding sliced Wasserstein fraction of variance explained is 0.53; the fitted densities when fitting GSAW for the same countries can be found in Appendix J. Countries with warmer climates typically experience a higher rate of excess deaths during the winter season compared to colder countries for the same temperature anomaly. For instance, in Spain, a country with a warm climate, there is a roughly 30% increase in the number of deaths during the winter months compared to the preceding summer. However, in Sweden, a country with a colder climate, this increase is only around 15%.

Figure 6 illustrates the Fréchet regression steps for one-dimensional distributions for some representative projections. As the base temperature increases, the variance of the temperature anomalies decreases and the excess winter death rates shift to a higher level. The slice corresponding to the projection angle at 135 degrees illustrates the main effect of the regression, which is a shift to the right as winter base temperature goes up, leading to higher winter mortality, an effect that is most pronounced for Spain. Most likely, populations



(a)



(b)

Figure 5: (a) Observed smoothed and (b) fitted densities from global slice-wise Wasserstein regression (GSWW) of the joint distributions of excess winter death rates (y-axis) and winter temperature anomalies (x-axis) for selected European countries, ordered by their base winter temperature. Each panel is labeled with the country name, and the corresponding average base winter temperature (in °C) for 1961–1990 is shown. The sliced Wasserstein fraction of variance explained is 0.53.

used to a relatively warm winter contain a higher fraction of individuals who are more susceptible to cold temperatures than populations accustomed to a cold winter.

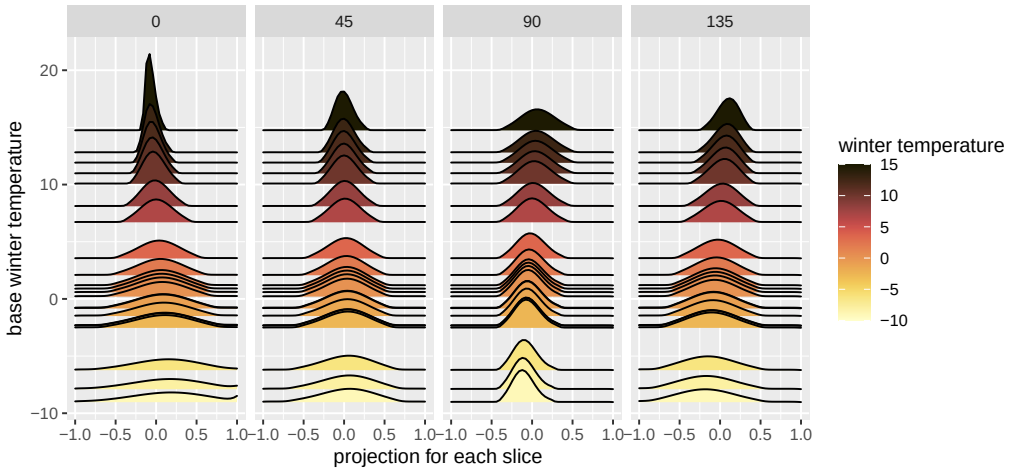
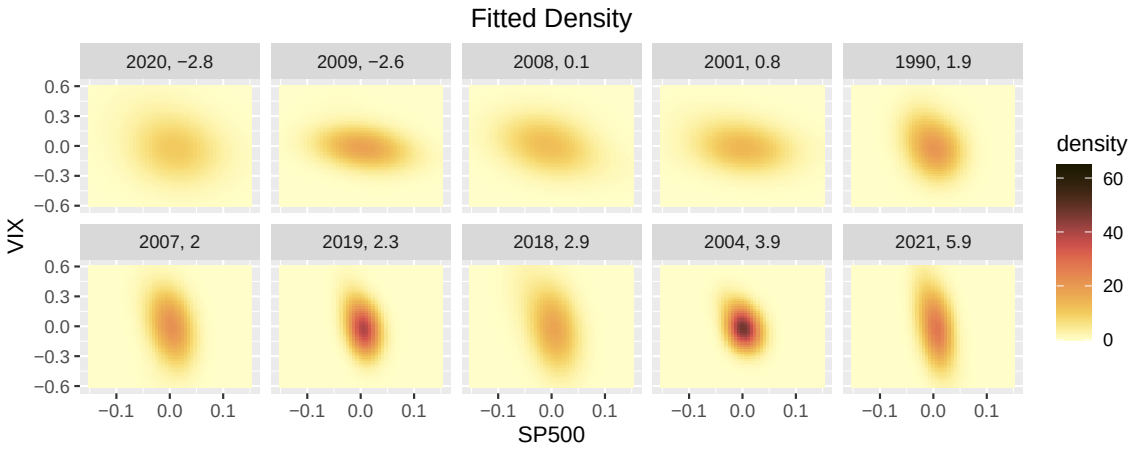


Figure 6: Fréchet regression between the base winter temperature (predictor, on the y-axis) and the slicing distributions (responses, on the x-axis) for various projections. The number at the top of each panel indicates the angle of the respective projection with the x-axis in Figure 5(a).

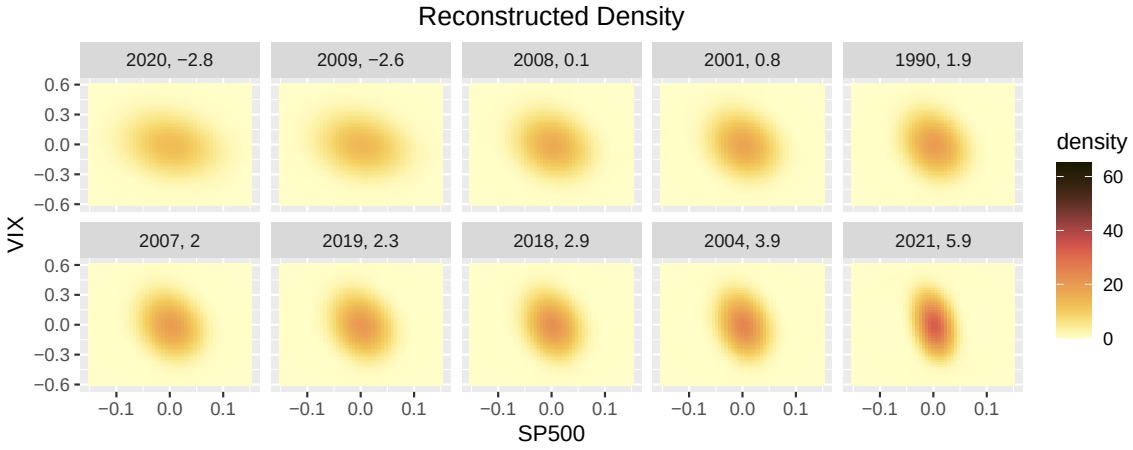
9.2 S&P 500 and VIX Index Joint Modeling

In the realm of financial markets, the application of distributional data modeling techniques holds promise, particularly concerning assets like stocks, exchange-traded funds or derivatives (Officer, 1972; Madan, 2020). One central challenge for financial market modeling stems from complex market dynamics featuring non-normality, low signal-to-noise ratio, and sudden, unpredictable changes in volatility (Madan and Wang, 2017). Traditional modeling of multivariate distributions in finance has centered on copula-based approaches that link marginal univariate distributions such as bilateral gamma marginals through a prespecified copula function (Cherubini et al., 2004; Küchler and Tappe, 2008; Madan, 2020). The promise of sliced Wasserstein regression is that it eliminates the need to decompose the bivariate distribution into marginal distributions and to specify a copula function by directly targeting the bivariate distributions and their relation with predictors. In this context, there is growing interest in exploring the joint distribution between weekly returns of two crucial financial indices, the S&P 500 index and the Volatility Index (VIX), where the latter tracks expected volatility in the stock market over the subsequent 30 days (Madan, 2020). The objective is to gain a deeper understanding of this joint distribution and its connection with the Gross Domestic Product (GDP) of the United States. The S&P 500 and VIX index data were sourced from Yahoo Finance (`https://finance.yahoo.com/`). The observed smoothed bivariate distributions are in Figure 7(a) for a few selected years, ordered by the associated GDP growth rate.

Consistent with the findings in the prior study by Madan (2020), the VIX returns are seen to exhibit more variation than the S&P 500 returns, along with an evident negative correlation between these two returns. This inverse relationship occurs because elevated VIX levels, indicating market turbulence, typically coincide with lower stock prices, reflecting



(a)



(b)

Figure 7: (a) Observed smoothed and (b) fitted densities from global slice-wise Wasserstein regression (GSWW) of the joint distributions of weekly returns of the VIX index, which reflects expected market volatility (y-axis) and S&P 500 index (x-axis) for selected years, arranged in order of the associated yearly GDP growth rate. Each panel is labeled by year; the corresponding yearly GDP growth rate (as a percentage) is also indicated.

investor anxiety. In contrast, periods of market stability or upward trends in stock prices are associated with lower VIX values. The fitted densities obtained from the GSWW model are shown in Figure 7(b). They are seen to generally track the observed trends but not the sharpness of the density peaks; the Wasserstein fraction of variance explained is low at 0.15. Details of how the sliced regression works can be seen in Figure 8, which illustrates the (global) Fréchet regression step for the univariate distributions corresponding to a few representative slices.

Further results for GSAW (instead of GSWW) and the local models (LSWW and LSAW) can be found in Appendix J. The slice-wise fits elucidate the association between the GDP

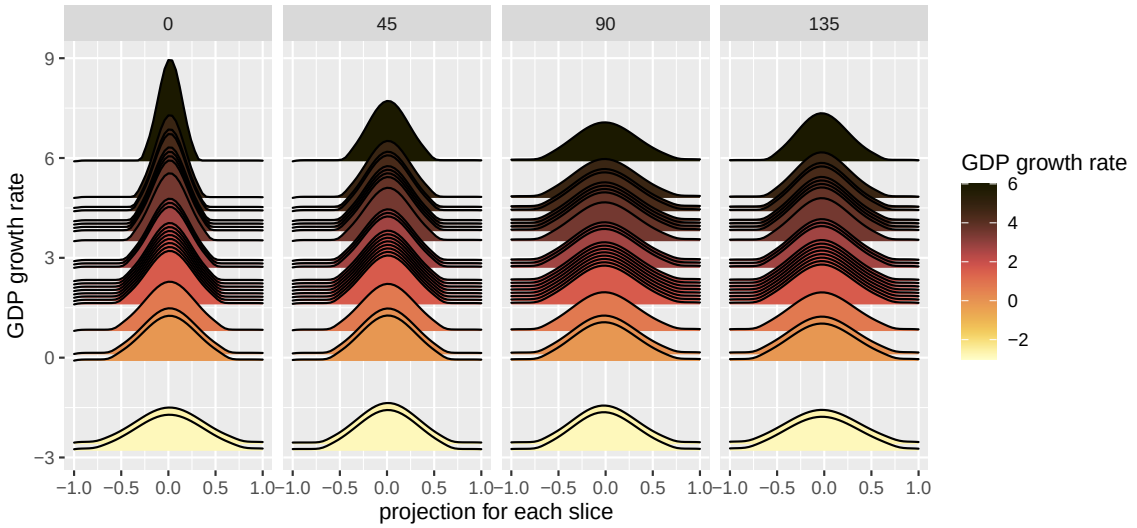


Figure 8: Fréchet regression between the GDP growth rate (predictor, on the y-axis) and the slicing distributions (responses, on the x-axis) for various projections. The number at the top of each panel indicates the angle of the respective projection with the x-axis in Figure 7(a).

growth rate and the bivariate distributions, showcasing the profound influence of GDP growth on the marginal distribution of the S&P 500 index's weekly return and its relationship with the VIX index. In Figure 8, an increase in the GDP growth rate coincides with a substantial decrease in the variance of the S&P 500 index return, contrasting with the relatively stable variance of the VIX index. Moreover, as depicted in Figure 7(b), a higher GDP growth rate intensifies the responsiveness of VIX index upticks to downticks in the S&P 500 index. In essence, during times of robust economic health (indicated by higher GDP growth rates), declines in the S&P 500 index are more likely to be associated with market turbulence, resulting in an increase in the VIX index.

For instance, during challenging periods such as 1990 (amidst the Gulf War), 2001 (following the September 11 terrorist attacks), 2008-2009 (amidst the financial crisis and substantial recession), and 2020 (amidst the COVID-19 pandemic), when the US economy faced significant hardships, a notable increase in the variance of the S&P 500 index was expected. In contrast, during more favorable economic phases like 2004 (highlighted by solid economic performance and low inflation rates), 2018-2019 (marked by sustained economic expansion and robust GDP growth), and 2021 (notably aided by fiscal stimulus to rebound from the pandemic's impact), when the US economy was thriving, downturns in the S&P 500 index were more prone to coincide with surges in the VIX index. This association potentially stems from fears regarding an impending economic downturn.

9.3 Exchange Traded Funds

An additional data illustration is presented in Appendix E to highlight the application of the local versions LSWW and LSAW of the sliced Wasserstein regression models.

10. Discussion and Outlook

The proposed SAW and SWW regression approaches are new tools for the statistical analysis of multivariate distributional data that come with theoretical guarantees and with the flexibility to employ both global and local Fréchet regression methods. The SAW regression model employs Fréchet regression over the multivariate distribution space equipped with the established slice-averaged Wasserstein distance, where the pointwise rates of convergence are optimal for both global and local SAW regression models under certain regularity conditions. The SWW approach offers a new perspective on how slicing transforms can be utilized for the analysis of data that consist of multivariate distributions. SWW regression is based on a regression step for each slice, thus allowing for parallel computation and avoiding the entanglement of the slices that is a characteristic of SAW regression. SWW is coupled with a regularized inverse transform from the sliced space to the original distribution space and is motivated by the idea that executing a regression step for each slice minimizes the slice-wise prediction error and when integrating this error it is smaller than the SAW error, as the latter seeks to minimize the aggregated error across all slices.

A downside of SWW regression is that it is associated with a slower rate of convergence. This is due to the fact that the slice-wise minimizers are not necessarily in the image space $\widetilde{\psi}(\mathscr{F})$, necessitating the use of a generalized version of the inverse slicing transform $\widetilde{\psi}_\tau^{-1}$ when transforming back to $\mathscr{F}$, which requires a regularization parameter τ and slows down the convergence of the final distribution estimates in the distribution space $\mathscr{F}$. Additionally, SWW requires a higher order of smoothness for the densities of the underlying distributions to achieve comparable convergence rates to SAW, since the convergence rates of SWW depend crucially on the convergence of the Fourier transform, where the undesirable assumption $k \geqslant p+1$ is required to ensure sufficiently fast convergence. However, these disadvantages are mitigated by the improved performance of SWW regression compared to SAW regression in both simulation settings and real data applications. This superior performance likely is due to smaller constants in the error rates for the slice-wise optimization in SWW as compared to the slice-averaged optimization in SAW, where the better rates of convergence of SAW cannot overcome this advantage for moderate sample sizes as typically encountered in statistical data analysis.

For situations where Assumption (A2) does not hold exactly, the theoretical arguments can be extended. In particular, if the slice-wise SWW population minimizer is sufficiently close to the image of the slicing transform $\widetilde{\psi}(\mathscr{F})$, one may redefine the population SWW target as the inverse image of its projection onto $\widetilde{\psi}(\mathscr{F})$, and analogous convergence results can still be obtained. We note this possibility here but do not pursue it further, as fully developing this alternative would substantially increase technical complexity without adding to the main statistical insights of the paper.

We investigated two cases, the theoretically ideal case where the underlying densities are all known, an assumption that has been routinely made in previous research on sliced Wasserstein methods, and a more realistic case where the densities are not known and must

be estimated from data generated from them. When densities need to be estimated, several additional issues arise. One of these is that uniform convergence of the estimators over the entire domain D is required, while the resulting estimates also need to be bona fide densities on D. While we offer a valid approach by truncating distributions to suitable subdomains, in order to keep the focus of the paper on the sliced Wasserstein approaches, a full resolution of this issue is left for future research. Commonly available estimators such as kernel estimators and their many variants do not achieve both properties over the entire domain; see Petersen and Müller (2016) for an in-depth discussion of the analogous issue for the one-dimensional case.

An ideal solution would lead to a bona fide density estimator that takes advantage of the additional order of smoothness k in the SWW case. However, our practical experience with kernel density estimation has shown that faster rates of convergence that are theoretically attainable for higher order smoothness of the underlying density functions coupled with higher order kernels rarely materialize for moderate sample sizes typically encountered in statistical data analysis. The reason for this is the same as mentioned above for the superiority of SWW over SAW in practical applications: The improvement in the rate of convergence is not strong enough to mitigate the effect of the larger constants associated with higher order kernels and one also incurs the problem of negative-valued density estimates (since higher order kernels cannot be non-negative).

In view of these considerations, we implemented a simple approach: We assume all distributions have sufficiently smooth densities f on domains $D_f \supset D_\epsilon \supset D$, where D is a compact convex set that is chosen by the user for a specific application. Then densities are estimated with positive kernels (of order 2) on D_ϵ and the truncated versions of these estimates on domain D are used as estimates of the densities on D, which are assumed to be the targets. This allows for boundary-effect-free estimation, yielding bona fide densities on D that satisfy the requisite assumptions. The individual domains D_f may be known in some settings, but are often unknown in practice. In such cases, the choice of the common domain D should be guided by subject-matter knowledge and standard statistical judgment.

Concurrent with our work, geometric properties associated with the so-called disintegrated distributional space have been recently studied (Kitagawa and Takatsu, 2026) and it was shown that the SAW representation and metric associated with the slice-averaged Wasserstein distance does not lead to a geodesic space (Park and Slepčev, 2025). In contrast, the SWW representation and associated distribution-slicing Wasserstein metric permits geodesics and thus is a geodesic space; this is a straightforward consequence of the fact that the Wasserstein space for univariate distributions is a geodesic space with the McCann interpolants serving as geodesics (Gangbo and McCann, 1996). A geodesic space is a prerequisite for adopting promising intrinsic approaches for principal component analysis (Bigot et al., 2017), regression (Zhu and Müller, 2026) and extrapolation (Fan and Müller, 2024). The geodesic structure of the novel slice-wise methodology that we introduced in this paper is thus a major advantage of this approach and exploiting this structure is a promising avenue for future research.

A natural direction for future research is to extend the proposed slicing-based framework to distribution-on-distribution regression, where both predictors and responses are probability distributions. Existing approaches in this setting include Ghodrati and Panaretos (2023) and Okano and Imaizumi (2024). Developing slicing-based regression methods

for distribution-valued predictors would require new methodological and theoretical tools, such as appropriate slicing transforms on predictor spaces and corresponding Fréchet regression formulations. This could be a promising avenue for broadening the scope of sliced Wasserstein regression methods.

For the common scenario where response distributions are not fully observed and must be reconstructed from random samples generated by the underlying distributions, a situation that was not considered in previous applications of the Radon transform for distribution slicing, the domain assumption requires a slight adjustment as discussed in Appendix C. This adjustment makes it possible to avoid an unproductive discussion of boundary effects when estimating densities from available data they generate while conforming with statistical practice, where invariably a user will choose a reasonable domain of interest. To guarantee that the density estimators are non-negative bona fide densities, we utilize the second-order smoothness of the multivariate density function. To take full advantage of the assumed smoothness of the multivariate density function to achieve a faster convergence rate, the option of employing a higher-order kernel κ of order k is available. However, the resulting density estimates are not guaranteed to be non-negative, violating the requirement of being bona fide densities. Further discussion of this issue can be found in Appendix C.

Acknowledgments

We wish to thank the editor and the reviewers for their valuable comments and suggestions, which have significantly improved the quality of the paper. This research was supported in part by NSF grants DMS-2014626 and DMS-2310450.

Appendix A. Abbreviations, Notations and Summary of Models

Table 4: List of notations (1). The bracketed explanation corresponds to the second half of the notation.

Notation	Explanation
D	known support of multivariate distribution
$\lvert D\rvert$	Lebesgue measure of the support D
p	dimension of the multivariate distribution
q	dimension of the predictor X
$f(z)$, $\mathcal{F}$	multivariate density functions (space)
$g(u)$, $\mathcal{G}$	univariate density functions (space)
$I(g)$	support of univariate distribution, $g \in \mathcal{G}$
$q(s)$, $\mathcal{Q}$	quantile functions (space)
$\mu, \mathscr{F}$	multivariate distributions (space)
$\nu, \mathscr{G}$	univariate distributions (space)
θ, Θ	slicing parameter (set)
λ, Λ_Θ	density slicing functions (space) from Θ to $\mathcal{G}$
γ, Γ_Θ	quantile slicing functions (space) from Θ to $\mathcal{Q}$
ψ, ψ_τ	slicing (regularized slicing) transform from $\mathcal{F}$ to Λ_Θ
$\mathcal{R}$, $\mathcal{R}_\tau$	Radon (regularized Radon) transform from $\mathcal{F}$ to Λ_Θ
$\phi_\tau(r)$	a regularizing function on Fourier domain
Δ_τ	reconstruction error
$\mathcal{J}_1, \mathcal{J}_p$	univariate (multivariate) Fourier transform
φ	map from $\mathscr{F}$ to $\mathcal{F}$ or from $\mathscr{G}$ to $\mathcal{G}$
ϱ	map from Γ_Θ to Λ_Θ
G	map from $\mathscr{G}$ to $\mathcal{Q}$
$\widetilde{\psi}, \widetilde{\psi}_\tau$	induced slicing (regularized slicing) transform from $\mathscr{F}$ to Γ_Θ

Appendix B. Additional Assumptions

(A3) $\arg\min_{\gamma\in\Gamma_\Theta} M_{L,h}^{SWW}(\gamma, x) \in \widetilde{\psi}(\mathscr{F})$ as per (32).

(L1) The kernel K used in the definition of the local sliced Wasserstein regression in Section D is a probability density function and is symmetric around zero. Furthermore, defining $K_{kj} = \int_{\mathbb{R}} K^k(u)u^j du$, $|K_{14}|$ and $|K_{26}|$ are both finite.

(L2) The marginal density of X denoted as $\mathcal{E}_X(x)$ and the conditional density of X given $\mu = \omega$, $\mathcal{E}_{X|\mu}(\cdot, \omega)$, exist and are twice continuously differentiable, the latter for all $\omega \in \mathscr{F}$, with $\sup_{x,\omega} \left| \left(\partial^2 \mathcal{E}_{X|\mu}/\partial x^2\right)(x, \omega)\right| < \infty$. Additionally, for any open $U \subset \mathscr{F}$, $\int_U dF_{\mu|X}(x, \omega)$ is continuous as a function of x.

(L3) The derivative K' exists and is bounded on the support of K, i.e., $\sup_{K(u)>0} |K'(u)| < \infty$; additionally, $\int_{\mathbb{R}} u^2 |K'(u)|(|u \log |u||)^{1/2} du < \infty$.

Table 5: List of notations (2). The bracketed explanation corresponds to the second half of the notation.

Notation	Explanation
d	metric on $\mathscr{F}$
d_W	Wasserstein distance on $\mathscr{F}$ or $\mathscr{G}$
d_{DW}	distribution-slicing Wasserstein metric on Γ_Θ
d_{SW}	slice-averaged Wasserstein distance on $\mathscr{F}$
d_2	L^2 norm
d_∞	L^∞ norm
Z_1, Z_2	random variables on $\mathbb{R}^p$
$\mathcal{P}(\mu_1, \mu_2)$	probability measure with marginal distributions μ_1, μ_2
μ, X	random pair with $\mu \in \mathscr{F}, X \in \mathbb{R}^q$
F	joint distribution of X and μ
F_X, F_μ	marginal distribution of X (of μ)
$m(x)$	Fréchet minimizer
$M(\cdot, x)$	conditional Fréchet function
$s_G(X, x), s_{i,G}(x)$	(sample) weight function of global Fréchet regression
SAW, GSAW	(global) slice-averaged Wasserstein
SWW, GSWW	(global) slice-wise Wasserstein
$m_G^{SAW}(x), m_{G,\tau}^{SWW}$	GSAW (GSWW) regression minimizer
$M_G^{SAW}(\cdot, x), M_G^{SWW}(\cdot, x)$	GSAW (GSWW) conditional Fréchet function
$\gamma_{G,x}$	minimizer of $M_G^{SWW}(\cdot, x)$
$\{(X_i, \mu_i)\}_{i=1}^n$	a sample of random pairs of predictors and measures
$\hat{\mu}_i, \hat{f}(x)$	estimated distribution (density function)
$\bar{X}, \hat{\Sigma}$	sample mean (variance) of $\{X_i\}_{i=1}^n$
$R^2_\oplus$	slice-averaged Wasserstein R^2 coefficient
$\mathbf{W} = \{\mathbf{W}_j\}_{j=1}^N$	random observations
$\mathbf{W}^{(i)} = \{\mathbf{W}_{ij}\}_{j=1}^{n_i}$	random observations from μ_i
$\mathbf{W}(\theta), \mathbf{W}^{(i)}(\theta)$	sliced observation (observations from μ_i)
$\mathcal{M}_G(\mathbf{W}, x)$	GSAW objective function
$\langle \cdot, \theta \rangle$	Radon slicing operation on random observation in $\mathbb{R}^p$
Π_A	permutation operator on $A \in \mathbb{R}^N$
η	step parameter in the gradient descent

(L4) Let $\mathcal{T}$ be a closed interval in $\mathbb{R}$ and $\mathcal{T}^o$ be its interior. Denote $\mathcal{E}_X(s)$ and $\mathcal{E}_{X|\mu}(\cdot, \omega)$ as per (L2), which exist and are twice continuously differentiable on $\mathcal{T}^o$, the latter for all $\omega \in \mathscr{F}$. The marginal density $\mathcal{E}_X(x)$ is bounded away from zero on $\mathcal{T}$, $\inf_{x \in \mathcal{T}} \mathcal{E}_X(x) > 0$. The second-order derivative $\mathcal{E}''_X(x)$ is bounded, $\sup_{x \in \mathcal{T}^o} |\mathcal{E}''_X(x)| < \infty$. The second-order partial derivatives $(\partial^2 \mathcal{E}_{X|\mu}(x, \omega)/\partial x^2)(x, \omega)$ are uniformly bounded, i.e., $\sup_{x \in \mathcal{T}^o, \omega \in \mathscr{F}} |(\partial^2 \mathcal{E}_{X|\mu}/\partial x^2)(x, \omega)| < \infty$. Additionally, for any open set $U \subset \mathscr{F}$, $\int_U dF_{\mu|X}(x, \omega)$

Table 6: List of notations for the Appendix. The bracketed explanation corresponds to the second half of the notation.

Notation	Explanation
D_ϵ	neighborhood surrounding the common support D
D_f	support of the density function f
D_i	support of the i-th distribution
f_{i,D_i}, μ_{i,D_i}	the i-th sample density (distribution) on support D_i
$f_{i,D_i\|D_\epsilon}$	the i-th sample density on the subdomain D_ϵ
$N_i, \tilde{N}_i$	number of observations on D_i (D_ϵ)
κ, b	kernel function (bandwidth) for density estimation
$\mathcal{E}_{X\|\mu}, \mathcal{E}_X$	conditional (marginal) density of X
K, h	kernel function (bandwidth) for local regression
$\mathcal{T}$	domain of the predictor X when $q = 1$
$s_L(X, x, h), s_{iL,h}(x)$	(sample) weight function of local Fréchet regression
LSAW	local slice-averaged Wasserstein
LSWW	local slice-wise Wasserstein
$m_{L,h}^{SAW}(x), m_{L,h,\tau}^{SWW}$	LSAW (LSWW) regression minimizer
$M_{L,h}^{SAW}(\cdot, x), M_{L,h}^{SWW}(\cdot, x)$	LSAW (LSWW) conditional Fréchet function
$\gamma_{L,h,x}$	minimizer of $M_{L,h}^{SWW}(\cdot, x)$
$\mathcal{M}_{L,h}(\mathbf{W}, x)$	LSAW objective function
$\mathcal{D}^{\mathbf{k}}$	differentiation operator on $\mathcal{F}$
$\mathcal{K}, a$	kernel function (bandwidth) used for a proof
$N(\epsilon, \mathscr{F}, d)$	covering number using balls of size ϵ
$B_\delta[m_G^{SAW}(x)]$	δ-ball in $\mathscr{F}$centered at $m_G^{SAW}(x)$
$B_G(x)$	best approximation
Ψ	map from quantile functions to distribution functions
$\eta_0, \eta_1, \eta_2, \beta_0, \beta_1, \beta_2$	constants
A_1, A_2, A_3, B_0, B_1	constants
$\tau_0, \tau_1, \tau_2, C_0, C_1, C_2, C, \zeta$	constants

is continuous as a function of x; for any $x \in \mathcal{T}$, $M(\cdot, x)$ is equicontinuous, i.e.,

$$\limsup_{y \to x} \sup_{\omega \in \mathscr{F}} |M(\omega, y) - M(\omega, x)| = 0.$$

Appendix C. Additional Preliminary Density Estimation Step

We assume for our main results that the multivariate distributions under consideration are well defined and possess density functions on a common domain D, satisfying (D1) and (F1). In practice, the common domain D will either be pre-specified in statistical applications or will be chosen by the analyst based on subject-matter or practical considerations. This is exemplified in our data examples. Once D has been selected, distributions truncated on D are the targets of the analysis and serve as responses for the SAW or SWW regression

Table 7: Summary of the global and local regression models.

Type	Regression	Object	Result
Global		weight function	$s_G(X,x) = 1 + (X - E(X))^\intercal \operatorname{Var}(X)^{-1}(x - E(X))$
	SAW	target function	$m_G^{SAW}(x) = \arg\min_{\omega\in\mathcal{F}} M_G^{SAW}(\omega, x)$ $M_G^{SAW}(\cdot, x) = E[s_G(X,x) d_{SW}^2(\mu, \cdot)]$
		convergence	$d_{SW}(m_G^{SAW}(x), \hat{m}_G^{SAW}(x)) = O_p(n^{-1/2})$
	SWW	target function	$m_{G,\tau}^{SWW}(x) = \widetilde{\psi}_\tau^{-1}\left[\arg\min_{\gamma\in\Gamma_\Theta} M_G^{SWW}(\gamma, x)\right]$ $M_G^{SWW}(\cdot, x) = E\left[s_G(X,x) d_{DW}^2(\widetilde{\psi}(\mu), \cdot)\right]$
		convergence	$d_\infty\left(m_G^{SWW}(x), \hat{m}_{G,\tau}^{SWW}(x)\right) = O_p\left(C_1(\tau) + C_2(\tau) n^{-2/7}\right)$
Local		weight function	$s_L(X,x,h) = K_h(X - x)[\upsilon_2 - \upsilon_1(X - x)]/\sigma_0^2$
	SAW	target function	$m^{SAW}(x) = \arg\min_{\omega\in\mathscr{F}} M^{SAW}(\omega, x)$ $M^{SAW}(\cdot, x) = E\left[d_{SW}^2(\mu, \cdot) \mid X = x\right]$
		local target	$m_{L,h}^{SAW}(x) = \arg\min_{\omega\in\mathcal{F}} M_{L,h}^{SAW}(\omega, x)$ $M_{L,h}^{SAW}(\cdot, x) = E[s_L(X,x,h) d_{SW}^2(\mu, \cdot)]$
		convergence	$d_{SW}(m^{SAW}(x), \hat{m}_{L,h}^{SAW}(x)) = O_p(n^{-2/5})$
	SWW	target function	$m^{SWW}(x) = \widetilde{\psi}^{-1}\left[\arg\min_{\gamma\in\Gamma_\Theta} M^{SWW}(\gamma, x)\right]$ $M^{SWW}(\cdot, x) = E\left[d_{DW}^2(\widetilde{\psi}(\mu), \cdot) \mid X = x\right]$
		local target	$m_{L,h,\tau}^{SWW}(x) = \widetilde{\psi}_\tau^{-1}\left[\arg\min_{\gamma\in\Gamma_\Theta} M_{L,h}^{SWW}(\gamma, x)\right]$ $M_{L,h}^{SWW}(\cdot, x) = E\left[s_L(X,x,h) d_{DW}^2(\widetilde{\psi}(\mu), \cdot)\right]$
		convergence	$d_\infty(m^{SWW}(x), \hat{m}_{L,h,\tau}^{SWW}(x)) = O_p\left(C_1(\tau) + C_2(\tau) n^{-8/35}\right)$

models. The actual distributions may have domains of which D is a subset and these domains may vary from distribution to distribution and could be unbounded.

For the practically important case where the distributions μ_i are not known but must be inferred from observations, we make some robust assumptions that will suffice to side-step the issue of boundary effects in density estimation without delving into complex technical details. A first assumption is that there exists $\epsilon > 0$ so that $D_\epsilon = \bigcup_{z\in D} B(z, \epsilon)$ is a subset of the domain of each underlying distribution, where $B(z, \epsilon)$ is a ball with radius ϵ centered at z. This assumption makes it possible to avoid a detailed discussion of boundary effects when estimating densities from available data they generate. We furthermore assume that the continuous differentiability assumption (F1) holds on D_ϵ. If a random distribution in the sample with density f_{D_f} has the (random) domain D_f, we require that $D_f \supset D_\epsilon$ and that on D_ϵ the density f_{D_f} is uniformly bounded from below and above, continuously differentiable of order k and has uniformly bounded partial derivatives, where the uniformity requirement extends across all f_{D_f}. Formally,

(F1′) There exists a constant $M_0 > 0$ and an integer $k \geqslant 2$ such that for all $f \in \mathcal{F}$, the density function f is the density of a truncated distribution on the domain D, where the original distribution has a domain D_f with $D_f \supset D_\epsilon$ and a density f_{D_f}

that is defined on D_f. It holds that $\max\{\|f_{D_f}\|_\infty, \|1/f_{D_f}\|_\infty\} \leqslant M_0$ on D_ϵ and that f_{D_f} is continuously differentiable of order k on D_ϵ, with uniformly bounded partial derivatives. The target distribution with density f is the truncated version of the f_{D_f} on D, so that for all $z \in D$ one has $f(z) = f_{D_f}(z)/\int_D f_{D_f}(u)du$.

Assume we have a sample of distributions $\mu_{1,D_1}, \ldots, \mu_{n,D_n}$ with densities $f_{1,D_1}, \ldots, f_{n,D_n}$ and domains $D_i \supset D_\epsilon$ and N_i i.i.d. observations generated by f_{i,D_i} by a random mechanism that is independent of the random mechanism that generates the μ_{i,D_i}. Let $\tilde{N}_i$ be the number of observations made for f_{i,D_i} that fall inside the domain D_ϵ. We impose the following assumption (P1) on the N_i to ensure that the L^2 convergence rate of the density estimates is faster than or at least as fast as the parametric rate $n^{-1/2}$.

(P1) $N(n) = \min_{1\leqslant i\leqslant n} N_i \gtrsim n^{(p+4)/4}$, where N_i is the number of random observations for the i-th distribution μ_{i,D_i}.

The following proposition shows that $\tilde{N}_i$ follows the same rate as (P1) with probability approaching one.

Proposition 10 *Assume (P1) and (F1′). Let $\tilde{N} = \min_{1\leqslant i\leqslant n} \tilde{N}_i$. There exists a constant $\tilde{c}$ such that*

$$P\left(\tilde{N} \geqslant \tilde{c} n^{(p+4)/4}\right) \to 1, \quad \text{as } n \to \infty.$$

To obtain the consistency of estimates of SAW regression and SWW regression when estimated densities are used, one needs to quantify the discrepancy between the true densities of the distributions $\{\mu_i\}_{i=1}^n$ and their estimates. This preliminary density estimation step can be implemented with standard density estimation methods, which have been extensively studied in the literature for both univariate and multivariate scenarios (Cowling and Hall, 1996; Hazelton and Marshall, 2009).

Let μ_{D_f} be a random probability distribution with density function f_{D_f} that satisfies (F1′), from which random observations $Z_1, ..., Z_N$ are independently sampled. Then a standard kernel estimator $\check{f}$ for f_{D_f} and its truncated version $\hat{f}$ for the density f truncated to the domain D is

$$\check{f}(z) = \frac{1}{Nb^p} \sum_{j=1}^{N} \kappa\left(\frac{z - Z_j}{b}\right), \quad \hat{f}(z) = \check{f}(z) \Big/ \int_D \check{f}(u)du\,, \quad z \in D \subset \mathbb{R}^p. \tag{27}$$

Here, κ is a kernel function and b a positive bandwidth (tuning parameter) satisfying Assumptions (K1) and (K2) listed below.

(K1) The kernel function κ as per (27) is a probability density function that has compact support and is symmetric, bounded and k times continuously differentiable (without loss of generality, the support is assumed to be contained in the unit cube of $\mathbb{R}^p$).

(K2) For some $A > 0$ and $\omega > 0$, the class of functions $\mathcal{I}_b = \{\kappa\left(\frac{z-\cdot}{b}\right), z \in \mathbb{R}^p, b > 0\}$ satisfies

$$\sup_{\mathcal{P}} M\left(\mathcal{I}_b, L_2(\mathcal{P}), \varepsilon\|F\|_{L_2(\mathcal{P})}\right) \leqslant \left(\frac{A}{\varepsilon}\right)^\omega,$$

where $M(T, d, \varepsilon)$ denotes the ε-covering number of the metric space (T, d), F is the envelope function of $\mathcal{I}_b$ and the supremum is taken over the set of all probability measures on $\mathbb{R}^p$.

Proposition 11 *Assume (D1), (F1′) and (K1). Choosing $b \sim N^{-\frac{1}{p+4}}$, the kernel density estimator $\hat{f}$ in (27) satisfies that*

$$\sup_{f \in \mathcal{F}} E\left[d_2\left(\hat{f}, f\right)^2 \middle| f \right] = O(N^{-4/(4+p)}). \tag{28}$$

We note that by construction, $\hat{f} \geqslant 0, \int_D \hat{f}(z)dz = 1$ and since $b \to 0$ as $N \to \infty$, we only need to consider the scenario where $b < \epsilon$, when due to the compactness of the support of the kernel κ there are no boundary effects when estimating f, as for any $x \in D$ one has $B(x, \epsilon) \subset D_\epsilon$. The following proposition formalizes this in terms of a uniform convergence rate.

Proposition 12 *Assume (D1), (F1′) and (K1)–(K2) and $b \sim N^{-\frac{1}{p+4}}$. The kernel density estimator $\hat{f}$ in (27) satisfies*

$$\sup_{f \in \mathcal{F}} d_\infty\left(\hat{f}, f\right) = O_p(\sqrt{\log N} N^{-\frac{2}{p+4}}). \tag{29}$$

We have assumed symmetric non-negative kernel functions, thus their first non-zero moments are of order two (alternatively one can also use products of kernels that are one-dimensional symmetric densities). If one wants to take full advantage of the higher order smoothness of the density function f in view of Assumptions (F2) or (F3), one can use higher order kernels κ of order k, where the order of the first non-zero moments is k. We refer to Müller and Stadtmüller (1999) for a more detailed account on multivariate kernels and boundary effects in multivariate density estimation. It is well known that such higher order kernel functions cannot be non-negative valued and thus it is not guaranteed anymore that the resulting kernel density estimates on D satisfy $\hat{f} \geqslant 0$, however one can still show that, since boundary effects can be ignored,

$$\sup_{f \in \mathcal{F}} E\left[d_2\left(\hat{f}, f\right)^2 \middle| f \right] = O(N^{-2k/(2k+p)}).$$

For example, under the condition $k \geqslant p + 1$ in (F3) one has

$$\sup_{f \in \mathcal{F}} E\left[d_2\left(\hat{f}, f\right)^2 \middle| f \right] = O(N^{-2/3}).$$

Under (F3), the condition (P1) can then be replaced by a dimension-independent rate, if one adopts the following condition (P1′).

(P1′) $N(n) = \min_{1 \leqslant i \leqslant n} N_i \gtrsim n^{3/2}$, where N_i is the number of random observations for the i-th distribution μ_i when densities need to be estimated.

The fact that $\hat{f} \geqslant 0$ is not guaranteed for these faster converging estimates makes this option less practical for our setting where the density estimates must be bona fide densities. The gains for higher order kernels are a consequence of their bias-reducing property, while the variance typically suffers a substantial increase through larger constants that are associated with these kernels. One can artificially enforce bona fide density estimates by modifying the estimator (Gajek, 1986) but this leads to new problems. If one nevertheless follows through with these adjustments, it may be possible to relax Condition (P1) to (P1′).

Appendix D. Local Sliced Wasserstein Regression

D.1 Regression Model

For local Fréchet regression (Petersen and Müller, 2019) we consider the case of a scalar predictor $X \in \mathbb{R}$ while the extension to $X \in \mathbb{R}^q$ with $q > 1$ is easily possible but tedious and for $q > 3$ usually subject to the curse of dimensionality, just like ordinary nonparametric regression approaches. For a smoothing kernel $K(\cdot)$ corresponding to a probability density and $K_h = h^{-1}K(\cdot/h)$, where h is a bandwidth, local Fréchet regression at x is defined as

$$m_{L,h}(x) = \operatorname*{arg\,min}_{\omega \in \mathscr{F}} M_{L,h}(\omega, x), \quad M_{L,h}(\cdot, x) = E\left[s_L(X, x, h) d^2(\mu, \cdot)\right], \tag{30}$$

where $s_L(X, x, h) = K_h(X - x)[\upsilon_2 - \upsilon_1(X - x)]/\sigma_0^2$ with $\upsilon_j = E[K_h(X - x)(X - x)^j]$ for $j = 0, 1, 2$ and $\sigma_0^2 = \upsilon_0\upsilon_2 - \upsilon_1^2$. The proposed local slice-averaged Wasserstein (LSAW) regression at x is

$$m_{L,h}^{SAW}(x) = \operatorname*{arg\,min}_{\omega \in \mathscr{F}} M_{L,h}^{SAW}(\omega, x), \quad M_{L,h}^{SAW}(\cdot, x) = E\left[s_L(X, x, h) d_{SW}^2(\mu, \cdot)\right]. \tag{31}$$

The proposed local slice-wise Wasserstein (LSWW) regression at x is defined as

$$m_{L,h,\tau}^{SWW}(x) = \widetilde{\psi}_\tau^{-1}\left[\operatorname*{arg\,min}_{\gamma \in \Gamma_\Theta} M_{L,h}^{SWW}(\gamma, x)\right], \quad M_{L,h}^{SWW}(\cdot, x) = E\left[s_L(X, x, h) d_{DW}^2(\widetilde{\psi}(\mu), \cdot)\right]. \tag{32}$$

In analogy to Proposition 5 for GSWW, the following result shows that the LSWW regression is equivalent to applying the Fréchet regression along each slice θ followed by an inverse transform.

Proposition 13 *The minimizing argument* $\gamma_{L,h,x} = \arg\min_{\gamma \in \Gamma_\Theta} M_{L,h}^{SWW}(\gamma, x)$, *see* (32), *is characterized as*

$$\gamma_{L,h,x}(\theta) = \operatorname*{arg\,min}_{\nu \in \mathscr{G}} E\left[s_L(X, x, h) d_W^2\left(G^{-1}\left(\widetilde{\psi}(\mu)(\theta)\right), \nu\right)\right], \quad \textit{for almost all } \theta \in \Theta.$$

D.2 Estimation

Suppose we have a sample of independent random pairs $\{(X_i, \mu_i)\}_{i=1}^n \sim F$, then the sample mean and variance are

$$\bar{X} = n^{-1}\sum_{i=1}^n X_i, \quad \hat{\Sigma} = n^{-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\intercal.$$

If random distributions μ_i are fully observed, sample estimators of LSAW and LSWW are obtained as

$$\check{m}_{L,h}^{SAW}(x) = \underset{\omega \in \mathscr{F}}{\arg\min}\, \check{M}_{L,h}^{SAW}(\omega, x), \quad \check{M}_{L,h}^{SAW}(\cdot, x) = n^{-1} \sum_{i=1}^{n} \left[s_{iL,h}(x) d_{SW}^2 \left(\mu_i, \cdot \right) \right], \tag{33}$$

$$\check{m}_{L,h,\tau}^{SWW}(x) = \widetilde{\psi}_{\tau}^{-1} \left[\underset{\gamma \in \Gamma_\Theta}{\arg\min}\, \check{M}_{L,h}^{SWW}(\gamma, x) \right], \; \check{M}_{L,h}^{SWW}(\cdot, x) = n^{-1} \sum_{i=1}^{n} \left[s_{iL,h}(x) d_{DW}^2 (\widetilde{\psi}(\mu_i), \cdot) \right], \tag{34}$$

where $s_{iL,h}(x) = K_h(X_i - x)[\hat{v}_2 - \hat{v}_1(X_i - x)]/\hat{\sigma}_0^2$ with $\hat{v}_j = n^{-1} \sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^j$, $j = 0, 1, 2$ and $\hat{\sigma}_0^2 = \hat{v}_0 \hat{v}_2 - \hat{v}_1^2$. If the random distributions μ_i are estimated by $\hat{\mu}_i$ as per (27), the corresponding estimators are

$$\hat{m}_{L,h}^{SAW}(x) = \underset{\omega \in \mathscr{F}}{\arg\min}\, \hat{M}_{L,h}^{SAW}(\omega, x), \quad \hat{M}_{L,h}^{SAW}(\cdot, x) = n^{-1} \sum_{i=1}^{n} \left[s_{iL,h}(x) d_{SW}^2 \left(\hat{\mu}_i, \cdot \right) \right]; \tag{35}$$

$$\hat{m}_{L,h,\tau}^{SWW}(x) = \widetilde{\psi}_{\tau}^{-1} \left[\underset{\gamma \in \Gamma_\Theta}{\arg\min}\, \hat{M}_{L,h}^{SWW}(\gamma, x) \right], \; \hat{M}_{L,h}^{SWW}(\cdot, x) = n^{-1} \sum_{i=1}^{n} \left[s_{iL,h}(x) d_{DW}^2 (\widetilde{\psi}(\hat{\mu}_i), \cdot) \right]. \tag{36}$$

A practical data-driven approach to select the tuning parameter τ and bandwidth h when an i.i.d. sample of random pairs $\{(X_i, \mu_i)\}_{i=1}^n$ is available can be obtained through leave-one-out cross-validation. Specifically, we aim to minimize the discrepancy between predicted and observed distributions, given by

$$\hat{\tau}, \hat{h} = \underset{\tau, h}{\arg\min} \sum_{i=1}^{n} d_{SW}^2 \left(\mu_i, \hat{m}_{L,h,\tau,-i}^{SWW}(X_i) \right),$$

where $\hat{m}_{L,h,\tau,-i}^{SWW}(X_i)$ is the prediction at X_i from the LSWW regression of the ith-left-out sample $\{(X_{i'}, \mu_{i'})\}_{i' \neq i}$. When the sample size n exceeds 30, we substitute leave-one-out cross-validation with 5-fold cross-validation to strike a balance between computational efficiency and the accuracy of the tuning parameter selection.

D.3 Asymptotic Convergence

Some additional assumptions listed in Section B are required to derive asymptotic convergence results for the local models. Assumption (A3) ensures the underlying minimizer of the LSWW regression belongs to the image space of the slicing transform. Additional kernel and distributional assumptions (L1)–(L4) are standard for local regression estimation. We provide the following convergence result for LSAW. Here Theorem 14 provides rates of convergence for the case where the densities in the random sample are known and Theorem 15 for the case where they are unknown and must be estimated from the data.

Theorem 14 *(LSAW for known distributions). Assume (D1), (F1)–(F2), (A1), (T0)–(T2) and (L1)–(L2). When adopting the slice-averaged Wasserstein distance, for a fixed*

$x \in \mathbb{R}$ and $m^{SAW}(x), m_{L,h}^{SAW}(x)$, and $\check{m}_{L,h}^{SAW}(x)$ as per (15), (31) *and* (33),

$$d_{SW}(m^{SAW}(x), m_{L,h}^{SAW}(x)) = O\left(h^2\right),$$
$$d_{SW}(m_{L,h}^{SAW}(x), \check{m}_{L,h}^{SAW}(x)) = O_p\left((nh)^{-1/2}\right),$$

and when taking $h \sim n^{-1/5}$,

$$d_{SW}(m^{SAW}(x), \check{m}_{L,h}^{SAW}(x)) = O_p(n^{-2/5}).$$

Furthermore, assuming (L3)–(L4) for a closed interval $\mathcal{T} \subset \mathbb{R}$, if $h \to 0$, $nh^2(-\log h)^{-1} \to \infty$ as $n \to \infty$, for any $\epsilon > 0$,

$$\sup_{x \in \mathcal{T}} d_{SW}(m^{SAW}(x), m_{L,h}^{SAW}(x)) = O\left(h^2\right),$$
$$\sup_{x \in \mathcal{T}} d_{SW}(m_{L,h}^{SAW}(x), \check{m}_{L,h}^{SAW}(x)) = O_p(\max\{\left(nh^2\right)^{-1/(2+\epsilon)}, [nh^2(-\log h)^{-1}]^{-1/2}\}),$$

and when taking $h \sim n^{-1/(6+2\epsilon)}$,

$$\sup_{x \in \mathcal{T}} d_{SW}(m^{SAW}(x), \check{m}_{L,h}^{SAW}(x)) = O_p\left(n^{-1/(3+\epsilon)}\right).$$

Theorem 15 *(LSAW for estimated distributions). Assume (D1), (F1)–(F2), (A1), (T0)–(T2), (L1)–(L2) and assumptions based on kernel density estimation (P1), (F1′), (K1)–(K2). When adopting the slice-averaged Wasserstein distance, for a fixed $x \in \mathbb{R}$ and $m^{SAW}(x)$, $m_{L,h}^{SAW}(x)$, and $\hat{m}_{L,h}^{SAW}(x)$ as per* (15), (31) *and* (35),

$$d_{SW}(m^{SAW}(x), m_{L,h}^{SAW}(x)) = O\left(h^2\right),$$
$$d_{SW}(m_{L,h}^{SAW}(x), \hat{m}_{L,h}^{SAW}(x)) = O_p\left((nh)^{-1/2}\right),$$

and when taking $h \sim n^{-1/5}$,

$$d_{SW}(m^{SAW}(x), \hat{m}_{L,h}^{SAW}(x)) = O_p(n^{-2/5}).$$

Furthermore, assuming (L3)–(L4) for a closed interval $\mathcal{T} \subset \mathbb{R}$, if $h \to 0$, $nh^2(-\log h)^{-1} \to \infty$ as $n \to \infty$, for any $\epsilon > 0$,

$$\sup_{x \in \mathcal{T}} d_{SW}(m^{SAW}(x), m_{L,h}^{SAW}(x)) = O\left(h^2\right),$$
$$\sup_{x \in \mathcal{T}} d_{SW}(m_{L,h}^{SAW}(x), \hat{m}_{L,h}^{SAW}(x)) = O_p(\max\{\left(nh^2\right)^{-1/(2+\epsilon)}, [nh^2(-\log h)^{-1}]^{-1/2}\}),$$

and when taking $h \sim n^{-1/(6+2\epsilon)}$,

$$\sup_{x \in \mathcal{T}} d_{SW}(m^{SAW}(x), \hat{m}_{L,h}^{SAW}(x)) = O_p\left(n^{-1/(3+\epsilon)}\right).$$

The pointwise convergence rate of the LSAW estimator achieves the optimal rates established for local linear estimators for the special case of real-valued responses.

Turning to the convergence of LSWW, under Assumption (A3), we define the population-level targets m^{SWW} as

$$m^{SWW} = \widetilde{\psi}^{-1}\left[\underset{\gamma\in\Gamma_\Theta}{\arg\min}\, M^{SWW}(\gamma, x)\right]. \tag{37}$$

For LSWW we again obtain results for two scenarios, when the densities are known in Theorem 16 and when they are unknown and must be estimated in Theorem 17.

Theorem 16 *(LSWW for known distributions). Assume (D1), (F1), (F3), (A3), (T0)–(T4) and (L1)–(L2). For LSWW regression, for a fixed $x \in \mathbb{R}$ and $m^{SWW}(x)$, $m^{SWW}_{L,h,\tau}(x)$ and $\check{m}^{SWW}_{L,h,\tau}(x)$ as per* (37), (32) *and* (34)*, with $C_1(\tau)$ and $C_2(\tau)$ from (T3),*

$$d_\infty(m^{SWW}(x), m^{SWW}_{L,h,\tau}(x)) = O\left(C_1(\tau) + C_2(\tau)h^{8/7}\right),$$
$$d_\infty(m^{SWW}_{L,h,\tau}(x), \check{m}^{SWW}_{L,h,\tau}(x)) = O_p\left(C_2(\tau)\,(nh)^{-2/7}\right),$$

and when taking $h \sim n^{-1/5}$,

$$d_\infty(m^{SWW}(x), \check{m}^{SWW}_{L,h,\tau}(x)) = O_p\left(C_1(\tau) + C_2(\tau)n^{-8/35}\right).$$

Furthermore, assume (L3)–(L4) for a closed interval $\mathcal{T} \subset \mathbb{R}$, if $h \to 0$, $nh^2(-\log h)^{-1} \to \infty$ as $n \to \infty$, then for any $\epsilon > 0$,

$$\sup_{x\in\mathcal{T}} d_\infty(m^{SWW}(x), m^{SWW}_{L,h,\tau}(x)) = O\left(C_1(\tau) + C_2(\tau)h^{8/7}\right),$$
$$\sup_{x\in\mathcal{T}} d_\infty(m^{SWW}_{L,h,\tau}(x), \check{m}^{SWW}_{L,h,\tau}(x)) = O_p\left(C_2(\tau)\max\left\{\left(nh^2\right)^{-2/(7+\epsilon)}, \left[nh^2(-\log h)^{-1}\right]^{-2/7}\right\}\right),$$

and when taking $h \sim n^{-1/(6+\epsilon)}$,

$$\sup_{x\in\mathcal{T}} d_\infty\left(m^{SWW}(x), \check{m}^{SWW}_{L,h,\tau}(x)\right) = O_p\left(C_1(\tau) + C_2(\tau)n^{-4/(21+\epsilon)}\right).$$

Theorem 17 *(LSWW for estimated distributions). Assume (D1), (F1), (F3), (A3), (T0)–(T4), (L1)–(L2) and assumptions based on kernel density estimation (P1), (F1′), (K1)–(K2). For LSWW regression, for a fixed $x \in \mathbb{R}$ and $m^{SWW}(x)$, $m^{SWW}_{L,h,\tau}(x)$ and $\hat{m}^{SWW}_{L,h,\tau}(x)$ as per* (37), (32) *and* (36)*, with $C_1(\tau)$ and $C_2(\tau)$ from (T3),*

$$d_\infty(m^{SWW}(x), m^{SWW}_{L,h,\tau}(x)) = O\left(C_1(\tau) + C_2(\tau)h^{8/7}\right),$$
$$d_\infty(m^{SWW}_{L,h,\tau}(x), \hat{m}^{SWW}_{L,h,\tau}(x)) = O_p\left(C_2(\tau)\,(nh)^{-2/7}\right),$$

and when taking $h \sim n^{-1/5}$,

$$d_\infty(m^{SWW}(x), \hat{m}^{SWW}_{L,h,\tau}(x)) = O_p\left(C_1(\tau) + C_2(\tau)n^{-8/35}\right).$$

Furthermore, assume (L3)–(L4) for a closed interval $\mathcal{T} \subset \mathbb{R}$*, if* $h \to 0$*,* $nh^2(-\log h)^{-1} \to \infty$ *as* $n \to \infty$*, then for any* $\epsilon > 0$*,*

$$\sup_{x\in\mathcal{T}} d_\infty(m^{SWW}(x), m^{SWW}_{L,h,\tau}(x)) = O\left(C_1(\tau) + C_2(\tau)h^{8/7}\right),$$
$$\sup_{x\in\mathcal{T}} d_\infty(m^{SWW}_{L,h,\tau}(x), \hat{m}^{SWW}_{L,h,\tau}(x)) = O_p\left(C_2(\tau)\max\left\{\left(nh^2\right)^{-2/(7+\epsilon)}, \left[nh^2(-\log h)^{-1}\right]^{-2/7}\right\}\right),$$

and when taking $h \sim n^{-1/(6+\epsilon)}$*,*

$$\sup_{x\in\mathcal{T}} d_\infty\left(m^{SWW}(x), \hat{m}^{SWW}_{L,h,\tau}(x)\right) = O_p\left(C_1(\tau) + C_2(\tau)n^{-4/(21+\epsilon)}\right).$$

This result demonstrates the convergence rate of LSWW regression and provides a decomposition of the reconstruction error into two components, similar to the situation for GSWW. In the special case of a Radon transform, Corollary 18 below shows that the curse of dimensionality is manifested in both $C_1(\tau)$ and $C_2(\tau)$, as a higher order of smoothness is required for a higher dimensional distribution to achieve the same convergence rate. Here we only give an explicit result for the more intricate scenario where densities are not fully observed and must be estimated; however, analogous results are available for scenarios with fully observed densities.

Corollary 18 *When taking the Radon transform* $\mathcal{R}$ *and the corresponding regularized inverse* $\mathcal{R}^{-1}_\tau$ *as per* (3) *and* (6)*, under the assumptions of Theorem 17,*

$$d_\infty\left(m^{SWW}(x), \hat{m}^{SWW}_{L,h,\tau}(x)\right) = O_p\left(\tau^{-(k-p)} + \tau^p n^{-8/35}\right),$$
$$\sup_{x\in\mathcal{T}} d_\infty\left(m^{SWW}(x), \hat{m}^{SWW}_{L,h,\tau}(x)\right) = O_p\left(\tau^{-(k-p)} + \tau^p n^{-4/(21+\epsilon)}\right),$$

and with $\tau \sim n^{8/(35k)}$*,*

$$d_\infty\left(m^{SWW}(x), \hat{m}^{SWW}_{L,h,\tau}(x)\right) = O_p\left(n^{-8(k-p)/35k}\right).$$

Furthermore, for $\tau \sim n^{4/(21k)}$*,*

$$\sup_{x\in\mathcal{T}} d_\infty\left(m^{SWW}(x), \hat{m}^{SWW}_{L,h,\tau}(x)\right) = O_p\left(n^{-4(k-p)/(21k+\epsilon)}\right).$$

We note that the uniform convergence results for LSAW and LSWW can be applied to scenarios where one aims at maximizing/minimizing a functional of the unknown distribution in dependence on the predictor. Then the target is the location x^* and the value of the functional where the extremum occurs; see Müller (1985) for the case of real-valued responses and Chen and Müller (2022) for univariate distribution-valued responses.

D.4 Numerical Algorithm

Following the notations in Section 7, the LSAW regression of (35) given $X = x$ is defined as,

$$\underset{\mathbf{W}\in\mathbb{R}^{p\times N}}{\arg\min}\, \mathcal{M}_{L,h}(\mathbf{W}, x) = n^{-1}\sum_{i=1}^{n}\left[s_{iL,h}(x)d^2_{SW}(\mu_{\mathbf{W}^{(i)}}, \mu_{\mathbf{W}})\right]. \tag{38}$$

The following proposition states that the target function of LSAW is smooth for the case of the Radon transform.

Proposition 19 (Theorem 1 (Bonneel et al., 2015)) *For each fixed x and $N_i \equiv N$, $\mathcal{M}_{L,h}(\mathbf{W}, x) : \mathbb{R}^{p\times N} \mapsto \mathbb{R}$ is an L^1 function with a uniformly ρ_L-Lipschitz gradient for some $\rho_L > 0$ given by*

$$\nabla\mathcal{M}_{L,h}(\mathbf{W}, x) = n^{-1}\sum_{i=1}^{n}\left[s_{iL,h}(x)\int_{\Theta}\theta\left(\mathbf{W}(\theta) - \Pi^{-1}_{\mathbf{W}(\theta)}\circ\Pi_{\mathbf{W}^{(i)}(\theta)}\circ\mathbf{W}^{(i)}(\theta)\right)^{\intercal}d\theta\right].$$

In analogy to Algorithm 1, we use the following gradient descent algorithm to find a stationary point.

Algorithm 3 LSAW Algorithm when using the Radon Transform

1: Initialize a grid $(\theta_1, \theta_2, ..., \theta_L)$ along Θ
2: Set $N = \min_{i=1,...,n} N_i$, convergence threshold ε and learning rate η
3: For each $\mu_{\mathbf{W}^{(i)}}$, downsample $\mathbf{W}^{(i)}$ such that $\mathbf{W}^{(i)} \in \mathbb{R}^{p\times N}$
4: Initialize $\mathbf{W}^{[0]} \in \mathbb{R}^{p\times N}$ arbitrarily and fix the output predictor $X = x$
5: **repeat**
6: Calculate $\nabla\mathcal{M}_{L,h}(\mathbf{W}^{[k]}, x)$ through

$$\nabla\mathcal{M}_{L,h}(\mathbf{W}^{[k]}, x) = (nL)^{-1}\sum_{i=1}^{n}\sum_{l=1}^{L}s_{iL,h}(x)\theta_l\left(\mathbf{W}(\theta_l) - \Pi^{-1}_{\mathbf{W}(\theta_l)}\circ\Pi_{\mathbf{W}^{(i)}(\theta_l)}\circ\mathbf{W}^{(i)}(\theta_l)\right)^{\intercal}$$

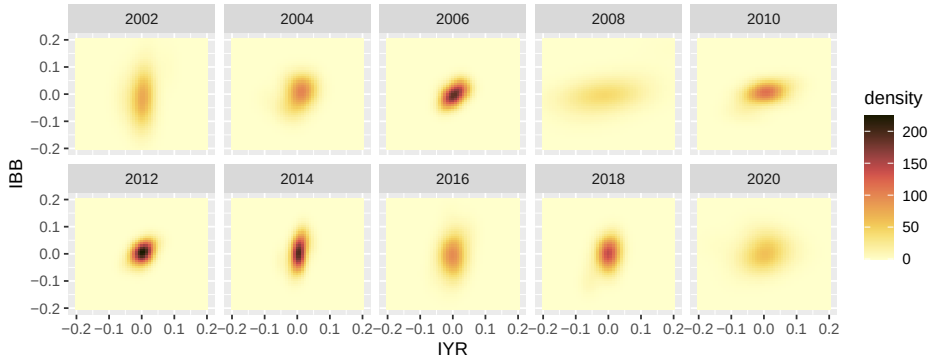
7: $\mathbf{W}^{[k+1]} = \mathbf{W}^{[k]} - \eta\nabla\mathcal{M}_{L,h}(\mathbf{W}^{[k]}, x)$
8: **until** Algorithm converges with $\|\mathbf{W}^{[k+1]} - \mathbf{W}^{[k]}\|_2/\|\mathbf{W}^{[k]}\|_2 < \varepsilon$ to $\mathbf{W}^{[\infty]}$
9: Consider each column of $\mathbf{W}^{[\infty]}$ as a sample from $\hat{m}^{SAW}_{L,h}(x)$ and apply the kernel density estimator (27) to derive the density estimator $\hat{f}$

Appendix E. Additional Data Application

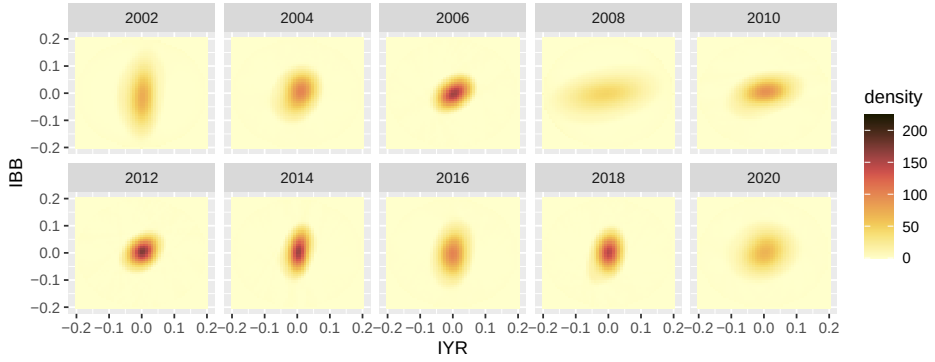
E.1 Exchange Traded Funds Modeling

We provide an additional data application to further illustrate the LSAW and LSWW regression models, aiming to demonstrate that these models can be used for the smoothing (nonparametric regression) of multivariate distributions against a one-dimensional covariate, for which we use calendar year in this application. Exchange Traded Funds (ETFs) are investment vehicles that invest assets to track a benchmark, such as a general index, sector, bonds, fixed income, etc. Sector ETFs that track a particular industry have become popular among investors and are widely used for hedging and statistical arbitrage. Historical data for sector ETFs can be obtained from Yahoo Finance `https://finance.yahoo.com/`.

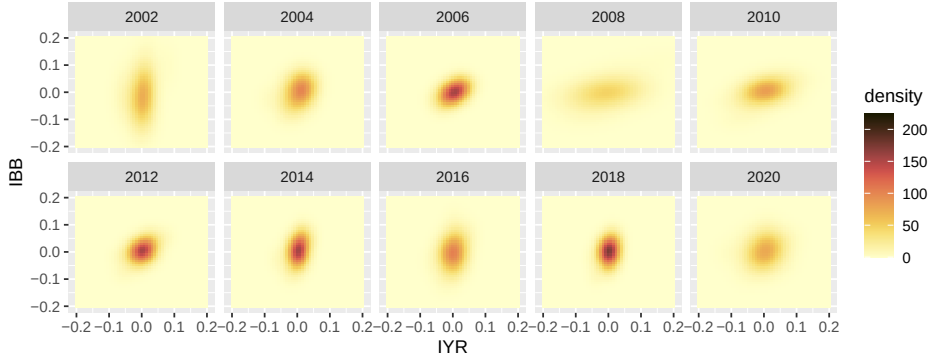
For each year, we model the bivariate distribution between the weekly return of iShares Biotechnology ETF (IBB) and the weekly return of iShares U.S. Real Estate ETF (IYR). Kernel density estimates of the data are in Figure 9(a). Figure 10 displays the local Fréchet regressions for representative angles for LSWW. The variance of weekly returns for the



(a)



(b)



(c)

Figure 9: (a) Observed smoothed densities and fitted densities from (b) local slice-wise Wasserstein regression (LSWW) and (c) local slice-averaged Wasserstein regression (LSAW) of the joint distributions of Biotechnology (IBB) (y-axis) and Real Estate (IYR) ETFs (x-axis). Each panel is labeled by year. The sliced Wasserstein fractions of variance explained are 0.95 for LSWW and 0.74 for LSAW.

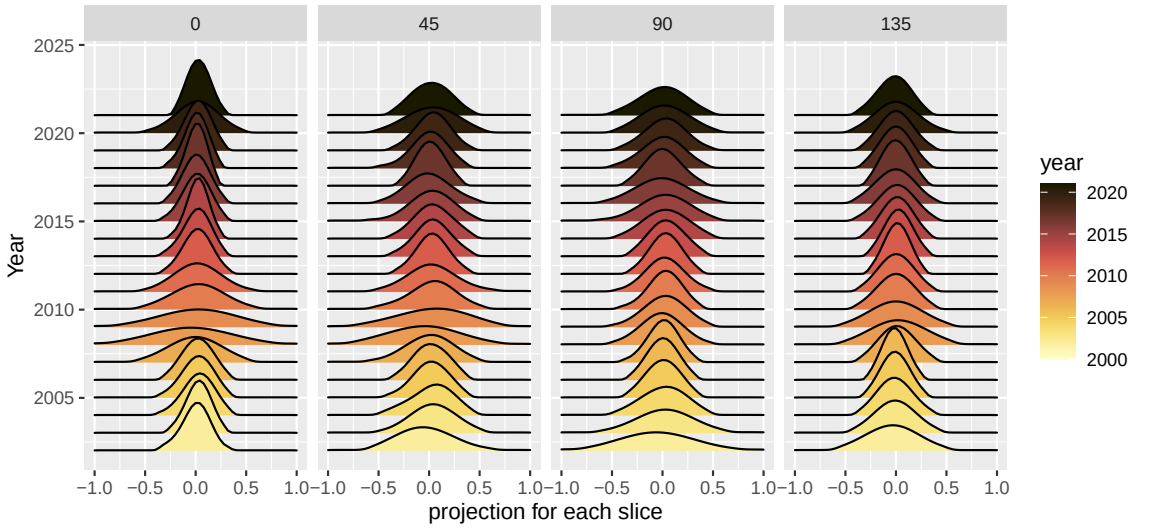


Figure 10: Fréchet regressions for LSWW between year (predictor, on the y-axis) and fitted slicing distributions (response, on the x-axis) for various projections. The number at the top of each panel indicates the angle of the respective projection with the x-axis (IYR) in Figure 9(a).

two included sectors increased during the financial crisis of 2008-2009 and the COVID-19 pandemic in 2020, indicating increased market uncertainty during these periods. The real estate market was impacted more significantly than biotechnology, likely due to its sensitivity to economic and financial conditions.

Reconstructed density surfaces obtained from LSWW are shown in Figure 9(b) while those obtained from LSAW are shown in Figure 9(c). The sliced Wasserstein fraction of variance explained for LSWW and LSAW models, as per (23), is 0.95 and 0.74. During periods of economic expansion and positive investor sentiment, the biotech and real estate sectors both benefit and become positively correlated, as was the case during the housing bubble from 2002 to 2008 and the post-crisis period from 2010 to 2018. Conversely, during times of economic uncertainty or market volatility, correlations between sectors tend to decrease as investors shift toward safe-haven assets. This was evident during the financial crisis of 2008-2009 and the COVID-19 pandemic in 2020. Both LSWW and LSAW models reveal the general trends of the joint distribution and achieve satisfactory values of the sliced Wasserstein fraction of variance explained.

Appendix F. Slicing Transforms

Various slicing transforms besides the Radon transform may be of interest. One of these is the circular Radon transform $\mathcal{CR}$ (Kuchment, 2006). This transform is the integral of a

function f over a sphere centered at $\theta \in \Theta$, where

$$\mathcal{CR}(f)(\theta, u) = \int_{|z-\theta|=u} f(z) d\sigma(z), \quad \theta \in \Theta.$$

The injectivity property of the circular Radon transform has been studied previously, especially for the case when Θ is a unit sphere (Agranovsky and Quinto, 1996; Ehrenpreis, 2003). However, the analytic inverse formulas for even dimensions are still unknown (Finch and Patch, 2004).

Another transform of interest is the generalized Radon transform, which extends the classic Radon transform (Beylkin, 1984; Ehrenpreis, 2003). The generalized Radon transform $\mathcal{GR}$ is defined as an integral over $I_{u,\theta} = \{z \in D | \chi(z, \theta) = u\}$,

$$\mathcal{GR}\left(f\right)(\theta, u) = \int_{\chi(z,\theta)=u} f(z) d\sigma(z),$$

where $d\sigma(z)$ integrates the surface area on $I_{u,\theta}$ and χ is a so-called defining function if it satisfies some regularity conditions (Beylkin, 1984). The sliced Wasserstein distance has been extended to the generalized Radon transform (Kolouri et al., 2019).

Appendix G. Auxiliary Lemmas and Propositions

Lemma 20 *Assume (D1) and (F1). For any vector* $\mathbf{k} = (k_1, ..., k_p)$ *of* p *non-negative integers with* $\sum_{l=1}^{p} k_l = k$ *let* $\mathcal{D}^{\mathbf{k}} = \frac{\partial^k}{\partial z_{k_1} \cdots \partial z_{k_p}}$. *It holds that* $\mathcal{R}(f)(\theta, \cdot), f \in \mathcal{F}$, *is* $k-$*times differentiable, and*

$$\mathcal{R}\left[\mathcal{D}^{\mathbf{k}} f\right](\theta, u) = (-1)^k \left(\prod_{l=1}^{p} \theta_l^{k_l}\right) \frac{\partial^k \left(\mathcal{R}(f)\right)}{\partial u^k}(\theta, u), \tag{39}$$

for each $\theta = (\theta_1, ..., \theta_p) \in \Theta$. *Furthermore,*

$$\left|\frac{\partial^k (\mathcal{R}(f))}{\partial u^k}(\theta, u)\right| \leqslant B_1 p^{k/2} < \infty,$$

for each $\theta \in \Theta$ *where the constant* B_1 *does not depend on* f *or* θ.

Proof Assumptions (D1) and (F1) imply f has bounded support and continuous partial derivatives of order k, and formula (39) follows from Proposition 6.1.3 in Epstein (2007). For each $\theta = (\theta_1, ..., \theta_p) \in \Theta$ with $\sum_{j=1}^{p} \theta_j^2 = 1$ one has that

$$\max_{j=1,..,p} |\theta_j| \geqslant p^{-1/2}.$$

Setting $j(\theta) = \arg\max_{j=1,\cdots,p} |\theta_j|$ and $\mathbf{k}(\theta) = (0, ..., 0, k, 0, ..., 0) = (k_1, \ldots, k_p)$ with $k_l = 0$, $l \neq j(\theta)$, and $k_{j(\theta)} = k$,

$$\left|\left(\prod_{l=1}^{p} \theta_l^{k_l}\right)\right| = \left(\prod_{l=1}^{p} |\theta_l|^{k_l}\right) \geqslant p^{-k/2}.$$

Since $\left\|\mathcal{D}^{\mathbf{k}(\theta)}f\right\|_\infty \leqslant B_1$ for a constant B_1 from (F1), it follows from the definition of the Radon transform that $\left\|\mathcal{R}\left[\mathcal{D}^{\mathbf{k}(\theta)}f\right](\theta,u)\right\|_\infty \leqslant B_1$. Hence,

$$\left|\frac{\partial^k\left(\mathcal{R}(f)\right)}{\partial u^k}(\theta,u)\right| \leqslant B_1 p^{k/2}.$$

■

Lemma 21 *Suppose h_1, h_2 are three times continuously differentiable on [0,1] with $\max\{\|h_1^{(3)}(s)\|_\infty, \|h_2^{(3)}(s)\|_\infty\} \leqslant B_0$ where B_0 is a constant, then*

$$|h_1'(s) - h_2'(s)| \leqslant C(B_0) d_2^{4/7}(h_1, h_2), \quad \forall z \in [0,1].$$

Proof Consider a kernel function $\mathcal{K}(\cdot)$ that is a symmetric probability density function with compact support on [-1,1] and a bounded derivative $\mathcal{K}'$. Furthermore, assume $\mathcal{K}$ satisfies that $\sigma^2(\mathcal{K}) = \int u^2\mathcal{K}(u)du < \infty$ and $\sigma^3(\mathcal{K}) = \int |u|^3\mathcal{K}(u)du < \infty$. Adopting Lemma 1 in Chen et al. (2020), arbitrarily fix $s \in (0,1)$, and assume for the bandwidth a that $a \leqslant \min\{s, 1-s\}$. As $a \to 0$,

$$\int_0^1 a^{-1}h_1(u)\mathcal{K}\left(\frac{s-u}{a}\right)du = h_1(s) + \frac{1}{2}a^2\sigma^2(\mathcal{K})h_1^{(2)}(s) + R_{11}(a), \tag{40}$$

$$\int_0^1 a^{-2}h_1(u)\mathcal{K}'\left(\frac{s-u}{a}\right)du = h_1'(s) + \frac{1}{2}a^2\sigma^2(\mathcal{K})h_1^{(3)}(s) + R_{12}(a). \tag{41}$$

Expanding $h_1(u)$,

$$h_1(u) = h_1(s) + h_1'(s)(u-s) + h_1^{(2)}(s)(u-s)^2/2 + \int_s^u \frac{h_1^{(3)}(t)(u-t)^2}{3!}dt.$$

Combining the expansion of $h_1(u)$ with (40) and (41), the remainder terms $R_{11}(a), R_{12}(a)$ can be bounded as follows

$$|R_{11}(a)| = \left|\int_0^1 a^{-1}\mathcal{K}\left(\frac{s-u}{a}\right)\int_s^u \frac{h_1^{(3)}(t)(u-t)^2}{3!}dtdu\right| \lesssim B_0 a^3,$$

$$\begin{aligned} R_{12}(a) = &\int_0^1 a^{-2}\mathcal{K}'\left(\frac{s-u}{a}\right)\int_s^u \frac{h_1^{(3)}(t)(u-t)^2}{3!}dtdu \\ &- \int_0^1 a^{-1}\mathcal{K}\left(\frac{s-u}{a}\right)\frac{h_1^{(3)}(s)(u-s)^2}{3!}du \lesssim B_0 a^2. \end{aligned}$$

Similarly for $h_2(u)$, we have

$$\int_0^1 a^{-2}h_2(u)\mathcal{K}'\left(\frac{s-u}{a}\right)du = h_2'(s) + \frac{1}{2}a^2\sigma^2(\mathcal{K})h_2^{(3)}(s) + R_{22}(a).$$

Then

$$\begin{aligned}|h_1'(s)-h_2'(s)| &\leqslant a^{-2}\int_0^1 |h_1(u)-h_2(u)|\mathcal{K}'\left(\frac{s-u}{a}\right)du + \frac{1}{2}a^2\sigma^2(\mathcal{K})|h_1^{(3)}(u)-h_2^{(3)}(u)| \\ &\quad + |R_{12}(a)-R_{22}(a)| \\ &\leqslant a^{-3/2}d_2(h_1,h_2)\left(\int\left(\mathcal{K}'(u)\right)^2du\right)^{1/2} + B_0a^2\sigma^2(\mathcal{K}) + |R_{12}(a)+R_{22}(a)| \\ &\lesssim C_1(B_0)\left(a^{-3/2}d_2(h_1,h_2)\right) + C_2(B_0)a^2,\end{aligned}$$

for constants $C_1(B_0), C_2(B_0)$ which only depend on B_0. By choosing $a \sim d_2^{2/7}(h_1,h_2)$,

$$|h_1'(s)-h_2'(s)| \leqslant C(B_0)d_2^{4/7}(h_1,h_2).$$

■

Lemma 22 *If D is compact, the univariate Wasserstein distance is bounded by the L^2 distance of the corresponding densities*

$$d_W(\nu_1,\nu_2) \lesssim d_2\left(\varphi(\nu_1),\varphi(\nu_2)\right), \quad \nu_1,\nu_2 \in \mathscr{G}. \tag{42}$$

Proof From Theorem 4 in Gibbs and Su (2002),

$$d_W(\nu_1,\nu_2) \leqslant \operatorname{diam}(D)d_{TV}(\nu_1,\nu_2), \quad \nu_1,\nu_2 \in \mathscr{G},$$

where $\operatorname{diam}(D) = \sup\{d(z_1,z_2) : z_1,z_2 \in D\}$ and $d_{TV}(\nu_1,\nu_2)$ represents the total variation distance $d_{TV} = \sup_{S\subset D}|\nu_1(S)-\nu_2(S)|$. Note that when ν_1 and ν_2 have densities $\varphi(\nu_1)$ and $\varphi(\nu_2)$, it follows that

$$d_{TV}(\nu_1,\nu_2) = \frac{1}{2}\int_D |\varphi(\nu_1)(u)-\varphi(\nu_2)(u)|du \leqslant \frac{1}{2}\sqrt{\operatorname{diam}(D)}d_2(\varphi(\nu_1),\varphi(\nu_2)).$$

If D is a compact set, $\operatorname{diam}(D)$ is bounded. We conclude that

$$d_W(\nu_1,\nu_2) \lesssim d_2(\varphi(\nu_1),\varphi(\nu_2)). \tag{43}$$

■

Proposition 23 *Assume (T0) and (T2), for m_G^{SAW} as per (16) and M_G^{SWW} as per (18),*

$$m_G^{SAW}(x) = \widetilde{\psi}^{-1}\left[\underset{\gamma\in\Gamma_\Theta\cap\widetilde{\psi}(\mathscr{F})}{\arg\min}\ M_G^{SWW}(\gamma,x)\right].$$

Similarly, for $m_{L,h}^{SAW}$ as per (31) and $M_{L,h}^{SWW}$ as per (32),

$$m_{L,h}^{SAW}(x) = \widetilde{\psi}^{-1}\left[\underset{\gamma\in\Gamma_\Theta\cap\widetilde{\psi}(\mathscr{F})}{\arg\min}\ M_{L,h}^{SWW}(\gamma,x)\right].$$

Proof The change of variable $\gamma = \widetilde{\psi}(\omega) \in \widetilde{\psi}(\mathscr{F})$, $\omega \in \mathscr{F}$ is a bijection because of the injectivity of the transform $\widetilde{\psi}$. Hence,

$$\begin{aligned} m_G^{SAW}(x) &= \underset{\omega\in\mathscr{F}}{\arg\min}\, E\left[s_G(X,x)d_{SW}^2(\mu,\omega)\right] \\ &= \underset{\omega\in\mathscr{F}}{\arg\min}\, E\left[s_G(X,x)d_{DW}^2(\widetilde{\psi}(\mu),\widetilde{\psi}(\omega))\right] \\ &= \widetilde{\psi}^{-1}\left(\underset{\gamma\in\Gamma_\Theta\bigcap\widetilde{\psi}(\mathscr{F})}{\arg\min}\, E\left[s_G(X,x)d_{DW}^2(\widetilde{\psi}(\mu),\gamma)\right]\right) \\ &= \widetilde{\psi}^{-1}\left[\underset{\gamma\in\Gamma_\Theta\bigcap\widetilde{\psi}(\mathscr{F})}{\arg\min}\, M_G^{SWW}(\gamma,x)\right]. \end{aligned}$$

Similarly,

$$m_{L,h}^{SAW}(x) = \widetilde{\psi}^{-1}\left[\underset{\gamma\in\Gamma_\Theta\cap\widetilde{\psi}(\mathscr{F})}{\arg\min}\, M_{L,h}^{SWW}(\gamma,x)\right].$$

■

Appendix H. Proofs

H.1 Proof of Proposition 10

Proof Denoting by $\tilde{N}_i$ the number of observations made for $f_{i,D_i|D_\epsilon}$ that fall within the domain D_ϵ, it is clear that each $\tilde{N}_i$ follows a binomial distribution $\mathscr{B}(p_i, N_i)$, where $p_i = \int_{D_\epsilon} f_{i,D_i|D_\epsilon}(z)dz \geqslant M_0^{-1}|D_\epsilon|$ according to (F1′). From (P1), there exists a constant $c > 0$, such that $\min_{1\leqslant i\leqslant n} N_i \geqslant cn^{1+p/4}$. Let $\tilde{c} = c\cdot|D_\epsilon|/(2M_0)$. Hoeffding's inequality (Hoeffding, 1994) implies that for any $\rho \leqslant N_i p_i$,

$$P\left(\tilde{N}_i \leqslant \rho\right) \leqslant \exp\left(-2N_i\left(p_i - \frac{\rho}{N_i}\right)^2\right),$$

whence

$$\begin{aligned} P\left(\min_{1\leqslant i\leqslant n}\tilde{N}_i > \tilde{c}n^{1+p/4}\right) &= \prod_{i=1}^n P\left(\tilde{N}_i > \tilde{c}n^{1+p/4}\right) \\ &= \prod_{i=1}^n\left(1 - P\left(\tilde{N}_i \leqslant \tilde{c}n^{1+p/4}\right)\right) \\ &\geqslant \prod_{i=1}^n\left(1 - e^{-2N_i(p_i - |D_\epsilon|/(2M_0))^2}\right). \end{aligned}$$

Setting $\tilde{m}_i = p_i - |D_\epsilon|/(2M_0) \geqslant |D_\epsilon|/(2M_0) > 0$,

$$\begin{aligned} P\left(\min_{1\leqslant i\leqslant n} \tilde{N}_i > \tilde{c}n^{1+p/4}\right) &\geqslant \left(1 - e^{-2N_i\tilde{m}_i^2}\right)^n \\ &\geqslant \left(1 - e^{-2c\tilde{m}_i^2 n^{1+p/4}}\right)^n \\ &= e^{n\log\left(1-e^{-2c\tilde{m}_i^2 n^{1+p/4}}\right)} \\ &\to 1 \text{ as } n \to \infty. \end{aligned}$$

■

H.2 Proof of Proposition 3

Proof First, we show that if Assumptions (D1), (F1) are satisfied, $\mathcal{R}(\mathcal{F})$ satisfies Assumptions (D2) and (G1). It is clear that $\mathcal{R}(f) \geqslant 0$ and $\int_u \mathcal{R}(f)(\theta, u)du = 1$, where we denote $\mathcal{R}(f)(\theta, u)$ as $\mathcal{R}(f)(\theta)(u)$ for simplicity. The validity of (D2) follows from (D1) and the definition of the Radon transform. Furthermore, (G1) holds based on (D1), (F1) and the implications of Lemma 20. Next, we prove that (T0) and (T1) are satisfied. The injectivity of the Radon transform is evident from the definition. Hence, (T0) is satisfied. Since D is compact, there exists a constant C such that $\|D\|_\infty \leqslant C$. Computing the L^2-distance between $\mathcal{R}(f_1)$ and $\mathcal{R}(f_2)$ for fixed $\theta \in \Theta$,

$$\begin{aligned} &\int_{\mathbb{R}} |\mathcal{R}(f_1)(\theta, u) - \mathcal{R}(f_2)(\theta, u)|^2 du \\ &= \int_{-C}^{C} \left(\int_{-C}^{C} \cdots \int_{-C}^{C} \left((f_1 - f_2)\left(u\theta + \sum_{j=1}^{p-1} s_j e_j \right) \right) ds_1 \cdots ds_p \right)^2 du \\ &\leqslant (2C)^{p-1} \int_{-C}^{C} \cdots \int_{-C}^{C} \left((f_1 - f_2)\left(u\theta + \sum_{j=1}^{p-1} s_j e_j \right) \right)^2 ds_1 \cdots ds_p du \\ &\leqslant (2C)^{p-1} d_2^2(f_1, f_2). \end{aligned}$$

The first inequality follows from the Cauchy-Schwarz inequality. Thus, (T1) holds as well. Assumptions (T2) and (T3) naturally follow as a corollary from Theorem 2. ■

H.3 Proof of Theorem 2

Proof Set $\overline{\mathcal{R}(f)}(\theta, r) = \mathcal{J}_1(\mathcal{R}(f)(\theta))(r)$, $r \in \mathbb{R}$ and $\check{\Delta}_{\tau,1} = f - \check{\mathcal{R}}_\tau(\mathcal{R}(f))$. Then we have

$$\check{\Delta}_{\tau,1} = \frac{1}{2(2\pi)^p} \int_\Theta \int_{|r|>\tau} \overline{\mathcal{R}(f)}(\theta, r) e^{ir\langle\theta,z\rangle} |r|^{p-1} dr d\theta.$$

Since $\mathcal{R}(f)(\theta)(\cdot)$ has uniform bounded k-derivative for each $\theta \in \Theta$ from (G1), it follows from Proposition 4.2.1 in Epstein (2007) that

$$|\overline{\mathcal{R}(f)}(\theta, r)| \leqslant M_3 |r|^{-k},$$

where M_3 is a constant such that $\|\mathcal{R}(f)(\theta)(\cdot)\|_\infty \leqslant M_3$. Therefore, we get

$$\begin{aligned}\left\|\check{\Delta}_{\tau,1}\right\|_\infty &\leqslant \frac{1}{2(2\pi)^p}\int_\Theta\int_{|r|>\tau}\left|\overline{\mathcal{R}(f)}(\theta,r)\right||r|^{p-1}drd\theta \\ &\leqslant \frac{M_3}{2(2\pi)^p}\int_\Theta d\theta\int_{|r|>\tau}|r|^{p-k-1}dr \\ &= O\left(\tau^{-(k-p)}\right),\end{aligned}$$

where we note that $k \geqslant p+1$ from (F3). Since $\int_D f(z)dz = 1 > 0$ and $\|f\|_\infty \leqslant M_0$ from (F1), we obtain $\|\Delta_{\tau,1}\|_\infty = O\left(\tau^{-(k-p)}\right)$.

Next, set $\lambda_f = \mathcal{R}(f)$, $\lambda_f^* = \widetilde{\mathcal{R}(f)} \in \Lambda_\Theta$ and $\check{\Delta}_{\tau,2} = \check{\mathcal{R}}_\tau^{-1}(\lambda_f) - \check{\mathcal{R}}_\tau^{-1}(\lambda_f^*)$,. This yields

$$\begin{aligned}\left\|\check{\Delta}_{\tau,2}\right\|_\infty &= \frac{1}{2(2\pi)^p}\int_\Theta\int_{|r|\leqslant\tau}(\overline{\lambda_f}(\theta,r) - \overline{\lambda_f^*}(\theta,r))|r|^{p-1}e^{ir\langle\theta,z\rangle}drd\theta \\ &\leqslant \frac{\tau^{p-1}}{2(2\pi)^p}\int_\Theta\int_{|r|\leqslant\tau}\left|\overline{\lambda_f}(\theta,r) - \overline{\lambda_f^*}(\theta,r)\right|drd\theta \\ &\leqslant \frac{\tau^{p-1}}{2(2\pi)^p}\int_\Theta\int_{|r|\leqslant\tau}\int_{\mathbb{R}}|\lambda_f(\theta,u) - \lambda_f^*(\theta,u)|dudrd\theta \\ &\lesssim \frac{\tau^p}{2(2\pi)^p}d_2(\lambda_f,\lambda_f^*).\end{aligned}$$

The last inequality follows from the bounded support of Λ_Θ and $\mathcal{F}$. Note that $\|\check{\mathcal{R}}_\tau(\mathcal{R}(f)) - f\|_\infty = O\left(\tau^{-(k-p)}\right)$ as $\tau \to \infty$. The boundedness of $\Delta_{\tau,2}$ follows from the fact $\int_D f(z)dz = 1$ and the condition that $d_2(\lambda_f^*, \lambda_f) \to 0$, whence

$$\|\Delta_{\tau,2}\|_\infty = O\left(\tau^p d_2(\lambda_f,\lambda_f^*)\right) = O\left(\tau^p d_2\left(\mathcal{R}(f),\widetilde{\mathcal{R}(f)}\right)\right).$$

■

H.4 Proof of Proposition 4

This proof follows a similar approach as Proposition 1 of Kolouri et al. (2019), but we present it for a more generalized version of the slicing transform as follows. If $\mu_1 = \mu_2$, $\mu_1, \mu_2 \in \mathscr{F}$, the slice-averaged Wasserstein distance satisfies $d_{SW}(\mu_1,\mu_2) = 0$. The non-negativity and symmetry properties of the slice-averaged Wasserstein distance are direct consequences of the Wasserstein distance being a metric. We next consider the triangle inequality. For any

$\mu_1, \mu_2, \mu_3 \in \mathscr{F}$,

$$\begin{aligned} d_{SW}(\mu_1, \mu_3) &= \left(\int_\Theta d_W^2(G^{-1}\left(\widetilde{\psi}(\mu_1)(\theta)\right), G^{-1}\left(\widetilde{\psi}(\mu_3)(\theta)\right)) \right)^{1/2} \\ &\leqslant \left(\int_\Theta d_W^2(G^{-1}\left(\widetilde{\psi}(\mu_1)(\theta)\right), G^{-1}\left(\widetilde{\psi}(\mu_2)(\theta)\right)) \right)^{1/2} \\ &\quad + \left(\int_\Theta d_W^2(G^{-1}\left(\widetilde{\psi}(\mu_2)(\theta)\right), G^{-1}\left(\widetilde{\psi}(\mu_3)(\theta)\right)) \right)^{1/2} \\ &= d_{SW}(\mu_1, \mu_2) + d_{SW}(\mu_2, \mu_3) \end{aligned}$$

The last inequality is obtained using the Minkowski inequality. We have thus established that the slice-averaged Wasserstein distance satisfies non-negativity, symmetry and the triangle inequality, hence it is a pseudo-metric. If $d_{SW}(\mu_1, \mu_2) = 0$,

$$d_W\left(G^{-1}\left(\widetilde{\psi}(\mu_1)(\theta)\right), G^{-1}\left(\widetilde{\psi}(\mu_2)(\theta)\right)\right) = 0, \quad \text{for almost all } \theta \in \Theta.$$

Equivalently, considering that the Wasserstein distance is a metric, we have $\widetilde{\psi}(\mu_1) = \widetilde{\psi}(\mu_2)$. Therefore, the slice-averaged Wasserstein distance is a distance if and only if $\widetilde{\psi}(\mu_1) = \widetilde{\psi}(\mu_2)$ implies $\mu_1 = \mu_2$, which is equivalent to the injectivity of $\widetilde{\psi}$. Note that $\widetilde{\psi}$ is induced from ψ, and since φ and ϱ are bijective, the injectivity of $\widetilde{\psi}$ is equivalent to (T0).

H.5 Proof of Proposition 5 and Proposition 13

Proof For any $\gamma_1, \gamma_2 \in \Gamma_\Theta$,

$$\begin{aligned} d_{DW}^2(\gamma_1, \gamma_2) &= \int_\Theta d_W^2(G^{-1}(\gamma_1(\theta)), G^{-1}(\gamma_2(\theta)))d\theta \\ &\lesssim \int_\Theta d_2^2(\varrho(\gamma_1)(\theta), \varrho(\gamma_2)(\theta))d\theta, \\ &\lesssim \int_\Theta d_\infty^2(\varrho(\gamma_1)(\theta), \varrho(\gamma_2)(\theta))d\theta. \end{aligned}$$

The first inequality comes from the distance relationship discussed in Lemma 22. Following the boundedness condition in (D1) and (G1), $d_{DW}(\gamma_1, \gamma_2)$ is bounded from above. Therefore,

$$\begin{aligned} & E\left[s_G(X, x) d_{DW}^2(\widetilde{\psi}(\mu), \gamma)\right] \\ &= \int_{X,\mu} s_G(X, x) \int_\Theta d_W^2\left(G^{-1}\left(\widetilde{\psi}(\mu)(\theta)\right), G^{-1}(\gamma(\theta))\right) d\theta d_F(X, \mu) \\ &= \int_\Theta \int_{X,\mu} s_G(X, x) d_W^2\left(G^{-1}\left(\widetilde{\psi}(\mu)(\theta)\right), G^{-1}(\gamma(\theta))\right) d_F(X, \mu) d\theta \\ &= \int_\Theta E\left[s_G(X, x) d_W^2\left(G^{-1}\left(\widetilde{\psi}(\mu)(\theta)\right), G^{-1}(\gamma(\theta))\right)\right] d\theta. \end{aligned}$$

The second equation follows from the Fubini theorem. The last equation indicates that finding the minimum in $E\left[s_G(X, x) d_{DW}^2(\widetilde{\psi}(\mu), \gamma)\right]$ is equivalent to finding the minimum in

$E\left[s_G(X,x)d_W^2\left(G^{-1}\left(\widetilde{\psi}(\mu)(\theta)\right),G^{-1}\left(\gamma(\theta)\right)\right)\right]$ for almost all $\theta\in\Theta$. The result for the local slice-wise Wasserstein regression can be derived analogously. ■

H.6 Proof of Proposition 11 and Proposition 12

Proof Set $z=(z_1,...,z_p)^\intercal\in D\subset D_\epsilon$, $t=(t_1,...,t_p)^\intercal\in D_\epsilon$ and the density f_{D_f} on domain D_ϵ as $f_{D_\epsilon}=f_{D_f}\mathbf{1}_{D_\epsilon}$. We can express the expected value as follows

$$E\left[\check{f}(z)\middle|f\right]=\frac{1}{b^p}\int_{D_\epsilon}\kappa\left(\frac{z-t}{b}\right)f_{D_\epsilon}(t)dt,\quad z\in D. \tag{44}$$

Expanding the multivariate density function $f_{D_\epsilon}(t)$ yields

$$f_{D_\epsilon}(t)=f_{D_\epsilon}(z)+\sum_{s=1}^{p}\frac{\partial f_{D_\epsilon}(z)}{\partial z_s}(t_s-z_s)+\sum_{s_1=1}^{p}\sum_{s_2=1}^{p}\frac{\partial f^2_{D_\epsilon}(z^*)}{\partial z_{s_1}z_{s_2}}(t_{s_1}-z_{s_1})(t_{s_2}-z_{s_2})/2,\quad t\in D_\epsilon,$$

where $z^*=z+\alpha(t-z)\in D_\epsilon$ for a $\alpha\in(0,1)$. With (44),

$$E\left[\check{f}(z)\middle|f\right]=f_{D_\epsilon}(z)+\frac{1}{2b^p}\sum_{s=1}^{p}\int_{D_\epsilon}\frac{\partial^2 f_{D_\epsilon}(z^*)}{\partial z_s^2}\kappa\left(\frac{z-t}{b}\right)(z_s-t_s)^2dt,\quad z\in D.$$

Here we use the symmetry property (K1) of the kernel function. From (F1′), there exists a constant M_3 that does not depend on the function f_{D_ϵ} such that the partial derivative $|\partial^2 f_{D_\epsilon}(z^*)/\partial z_s^2|$ is uniformly bounded by M_3 for $s=1,...,p$. Then

$$\sup_{z\in D}\sup_{f\in\mathcal{F}}\left|E\left[\check{f}(z)\middle|f\right]-f_{D_\epsilon}(z)\mathbf{1}_D\right|=O(b^2). \tag{45}$$

Next, we establish the boundedness of the variance of $\check{f}(z)$,

$$\begin{aligned}\mathrm{Var}\left(\check{f}(z)\middle|f\right)&=\frac{1}{N}\mathrm{Var}\left(\kappa\left(\frac{z-Z_j}{b}\right)b^{-p}\middle|f\right)\\&\leqslant\frac{1}{N}\int_{D_\epsilon}\kappa\left(\frac{z-t}{b}\right)^2b^{-2p}f_{D_\epsilon}(t)dt\\&\leqslant\frac{M_0}{N}\int_{D_\epsilon}\kappa\left(\frac{z-t}{b}\right)^2b^{-2p}dt,\end{aligned}$$

where M_0 from Assumption (F1) is a constant that does not depend on the density function f_{D_ϵ}. From (K2),

$$\sup_{z\in D}\sup_{f\in\mathcal{F}}\mathrm{Var}\left(\check{f}(z)\middle|f\right)=O\left(\frac{1}{Nb^p}\right).$$

Combining the above results, we obtain

$$\sup_{f\in\mathcal{F}}E\left(d_2\left(\check{f},f_{D_\epsilon}\mathbf{1}_D\right)^2\middle|f\right)=O\left(\frac{1}{Nb^p}+b^4\right).$$

Choosing $b \sim N^{-\frac{1}{p+4}}$ leads to

$$\sup_{f\in\mathcal{F}} E\left(d_2\left(\check{f}, f_{D_\epsilon}\mathbf{1}_D\right)^2 \Big| f\right) = O\left(N^{-4/(4+p)}\right).$$

For a fixed $f \in \mathcal{F}$ and the corresponding f_{D_ϵ}, by (45) and Proposition 9 in Rinaldo and Wasserman (2010),

$$\sup_{f\in\mathcal{F}} d_\infty\left(\check{f}, f_{D_\epsilon}\mathbf{1}_D\right) = O(b^2) + O_p\left(\sqrt{\frac{\log N}{Nb^p}}\right) = O_p(\sqrt{\log N}N^{-\frac{2}{p+4}}).$$

The error of uniform bound can be decomposed into a deterministic term $E[\check{f}(z)] - f_{D_\epsilon}(z)$ and a probabilistic term $\check{f} - E[\check{f}(z)]$. The first part depends on smoothness properties of f_{D_ϵ} only while the second part is characterized via empirical process techniques (Rinaldo and Wasserman, 2010; Jiang, 2017; Chen, 2017).

Note that the truncated density on domain D is presented as $f(z) = f_{D_\epsilon}(z)/\int_D f_{D_\epsilon}(u)du, z \in D$. We can then provide the convergence result for the truncated density $\hat{f}$. Note that when N is large enough we have,

$$\int_D f_{D_\epsilon}(u)du \geqslant \frac{|D|}{M_0}, \quad \int_D \check{f}(u)du \geqslant \frac{|D|}{2M_0}.$$

Whence,

$$\begin{aligned} d_2\left(\hat{f}, f\right) &\leqslant d_2\left(\frac{\check{f}(z)}{\int_D \check{f}(u)du}, \frac{f_{D_\epsilon}(z)\mathbf{1}_D}{\int_D \check{f}(u)du}\right) + d_2\left(\frac{f_{D_\epsilon}(z)\mathbf{1}_D}{\int_D \check{f}(u)du}, \frac{f_{D_\epsilon}(z)\mathbf{1}_D}{\int_D f_{D_\epsilon}(u)du}\right) \\ &\leqslant \frac{2M_0}{|D|} d_2(\check{f}, f_{D_\epsilon}\mathbf{1}_D) + \frac{2M_0^3}{|D|^{3/2}} d_2(\check{f}, f_{D_\epsilon}\mathbf{1}_D). \end{aligned}$$

Similarly, we have

$$d_\infty\left(\hat{f}, f\right) \leqslant \frac{2M_0}{|D|} d_\infty(\check{f}, f_{D_\epsilon}\mathbf{1}_D) + \frac{2M_0^3}{|D|} d_\infty(\check{f}, f_{D_\epsilon}\mathbf{1}_D).$$

It follows that when choosing $b \sim N^{-\frac{1}{p+4}}$

$$\begin{aligned} &\sup_{f\in\mathcal{F}} E\left[d_2\left(\hat{f}, f\right)^2\right] = O\left(N^{-4/(4+p)}\right), \\ &\sup_{f\in\mathcal{F}} d_\infty\left(\hat{f}, f\right) = O_p(\sqrt{\log N}N^{-\frac{2}{p+4}}). \end{aligned}$$

■

H.7 Proof of Theorem 6

Proof The proof proceeds in two steps.

Step 1. We first prove $d_{SW}\left(m_G^{SAW}(x), \check{m}_G^{SAW}(x)\right) = O_p\left(n^{-1/2}\right)$ and establish the following three properties for the SAW estimator $\check{m}_G^{SAW}(x)$.

(R1) The objects $m_G^{SAW}(x)$ and $\check{m}_G^{SAW}(x)$ exist and are unique, the latter almost surely, and, for any $\epsilon > 0$,

$$\inf_{d_{SW}\left(m_G^{SAW}(x),\omega\right)>\epsilon} M_G^{SAW}(\omega, x) > M_G^{SAW}\left(m_G^{SAW}(x), x\right).$$

(R2) Let $B_\delta\left[m_G^{SAW}(x)\right]$ be the δ-ball in $\mathscr{F}$ centered at $m_G^{SAW}(x)$ and $N(\epsilon, \mathscr{F}, d_{SW})$ its covering number using balls of size ϵ. Then

$$\int_0^1 \left(1 + \log N\left\{\delta\epsilon, B_\delta\left[m_G^{SAW}(x)\right], d_{SW}\right\}\right)^{1/2} d\epsilon = O(1) \quad \text{as } \delta \to 0.$$

(R3) There exists $\eta_0 > 0, \beta_0 > 0$, possibly depending on x, such that

$$\inf_{d_{SW}\left(m_G^{SAW}(x),\omega\right)<\eta_0} \left\{M_G^{SAW}(\omega, x) - M_G^{SAW}\left(m_G^{SAW}(x), x\right) - \beta_0 d_{SW}\left(m_G^{SAW}(x), \omega\right)^2\right\} \geqslant 0.$$

Proof of (R2). We note that $d_W^2(\nu_1, \nu_2) = O(d_2(\varphi(\nu_1), \varphi(\nu_2)))$ from Lemma 22, whence

$$\begin{aligned} d_{SW}(\mu_1, \mu_2) &= \left(\int_\Theta d_W^2(G^{-1}\left(\widetilde{\psi}(\mu_1)(\theta)\right), G^{-1}\left(\widetilde{\psi}(\mu_2)(\theta)\right))d\theta\right)^{1/2} \\ &\lesssim \left(\int_\Theta d_2^2(\psi \circ \varphi(\mu_1)(\theta), \psi \circ \varphi(\mu_2)(\theta))d\theta\right)^{1/2} \\ &\lesssim d_2(\varphi(\mu_1), \varphi(\mu_2)), \end{aligned} \tag{46}$$

where the last inequality follows from (T1) and the compactness of Θ. Using Theorem 2.7.1 of van der Vaart and Wellner (2023), there exists a constant A_1 depending only on k and p such that

$$\log N(\epsilon, \mathcal{F}, \|\|_\infty) \leqslant A_1 \epsilon^{-p/k},$$

for every $\epsilon > 0$ and $k > p/2$ from (F2). Note that $d_{SW}(\mu_1, \mu_2) = O(d_2(\varphi(\mu_1), \varphi(\mu_2))) = O(d_\infty(\varphi(\mu_1), \varphi(\mu_2))$ and from the boundedness of the support D, we have $B_{A_2\epsilon}(\mu, \|\|_\infty) \subset B_\epsilon(\mu, d_{SW})$ for some constant A_2. Thus, we have $N(\epsilon, \mathscr{F}, d_{SW}) \leqslant A_3 N(\epsilon, \mathcal{F}, \|\|_\infty)$. For $\tilde{k} = p/k < 2$ it follows that

$$\begin{aligned} \int_0^1 \sqrt{1 + \log N\{\delta\epsilon, B_\delta[m_G^{SAW}(x)], d_{SW}\}}d\epsilon &< \int_0^1 \sqrt{1 + \log N\{\delta\epsilon, \mathscr{F}, d_{SW}\}}d\epsilon \\ &\leqslant \int_0^1 \sqrt{1 + A_3 A_1 \epsilon^{-p/k}}d\epsilon \qquad (47) \\ &\lesssim \int_0^1 \epsilon^{-\tilde{k}/2} < \infty. \qquad (48) \end{aligned}$$

Proof of (R1) and (R3). We define the Hilbert space $\mathcal{H}$ as the set of all functions

$$\mathcal{H} := \{\varsigma \in \mathcal{H} : \Theta \times [0,1] \mapsto \mathbb{R}, \int_\Theta \int_{[0,1]} \varsigma^2(\theta, s) ds d\theta < \infty\}$$

with the inner product

$$\langle \varsigma_1, \varsigma_2 \rangle = \int_\Theta \int_{[0,1]} \varsigma_1(\theta, s)\varsigma_2(\theta, s) ds d\theta.$$

Here the integral is well defined due to the Cauchy-Schwarz inequality. It is easy to verify that the space $\mathcal{H}$ is a vector space over the field $\mathbb{R}$. The inner product satisfies the conditions of conjugate symmetry, linearity, and positive definiteness. The compactness of the Hilbert space follows from the fact that $\mathcal{H}$ consists of measurable functions that are square integrable. We define the L^2 distance between two functions $\varsigma_1, \varsigma_2 \in \mathcal{H}$ as $d_2(\varsigma_1, \varsigma_2) = \langle \varsigma_1 - \varsigma_2, \varsigma_1 - \varsigma_2 \rangle$. Note that the quantile slicing space Γ_Θ is a subspace of $\mathcal{H}$ and the distribution slicing Wasserstein metric coincides with the L^2 distance, i.e., $d_{DW}(\gamma_1, \gamma_2) = d_2(\gamma_1, \gamma_2)$ for $\gamma_1, \gamma_2 \in \Gamma_\Theta$. Let $B_G(x) = E[s_G(X, x)\widetilde{\psi}(\mu)]$, then for any fixed $\omega \in \mathscr{F}$,

$$\begin{aligned}
M_G^{SAW}(\omega, x) &= E[s_G(X, x) d_{SW}^2(\mu, \omega)] \\
&= E[s_G(X, x) d_2^2(\widetilde{\psi}(\mu), \widetilde{\psi}(\omega))] \\
&= E[s_G(X, x) \langle \widetilde{\psi}(\mu) - \widetilde{\psi}(\omega), \widetilde{\psi}(\mu) - \widetilde{\psi}(\omega) \rangle] \\
&= E[s_G(X, x) \langle \widetilde{\psi}(\mu) - B_G(x), \widetilde{\psi}(\mu) - B_G(x) \rangle] \\
&\quad + E[s_G(X, x) \langle B_G(x) - \widetilde{\psi}(\omega), B_G(x) - \widetilde{\psi}(\omega) \rangle] \\
&\quad + 2E[s_G(X, x) \langle \widetilde{\psi}(\mu) - B_G(x), B_G(x) - \widetilde{\psi}(\omega) \rangle] \\
&= E[s_G(X, x) d_2^2(\widetilde{\psi}(\mu), B_G(x))] + E[s_G(X, x) d_2^2(B_G(x), \widetilde{\psi}(\omega))] \\
&= E[s_G(X, x) d_2^2(\widetilde{\psi}(\mu), B_G(x))] + d_2^2(B_G(x), \widetilde{\psi}(\omega)),
\end{aligned}$$

where the last equation follows from the fact that $E[s_G(X, x)] = 1$. Thus,

$$m_G^{SAW}(x) = \arg\min_{\omega \in \mathscr{F}} d_2^2(B_G(x), \widetilde{\psi}(\omega)).$$

Using $n^{-1} \sum_{i=1}^n s_{iG}(x) = 1$, one can similarly show that

$$\check{m}_G^{SAW}(x) = \arg\min_{\omega \in \mathscr{F}} d_2^2(\check{B}_G(x), \widetilde{\psi}(\omega)), \tag{49}$$

where $\check{B}_G(x) = n^{-1} \sum_{i=1}^n s_{iG}(x)\widetilde{\psi}(\mu_i)$. From the convexity and closedness of space $\widetilde{\psi}(\mathscr{F})$, the minimizers $m_G^{SAW}(x)$ and $\check{m}_G^{SAW}(x)$ exist and are unique for any $x \in \mathbb{R}^q$, so that (R1) is satisfied. The best approximation $m_G^{SAW}(x) \in \mathscr{F}$ can be characterized by (Deutsch, 2012, chap.4)

$$\langle B_G(x) - \widetilde{\psi}(m_G^{SAW}(x)), \widetilde{\psi}(\omega) - \widetilde{\psi}(m_G^{SAW}(x)) \rangle \leqslant 0, \text{ for all } \omega \in \mathscr{F}. \tag{50}$$

Consequently, $d_2^2(B_G(x), \widetilde{\psi}(\omega)) \geqslant d_2^2(B_G(x), \widetilde{\psi}(m_G^{SAW}(x))) + d_2^2(\widetilde{\psi}(m_G^{SAW}(x)), \widetilde{\psi}(\omega))$. Then,

$$\begin{aligned} M_G^{SAW}(\omega, x) &\geqslant M_G^{SAW}(m_G^{SAW}(x), x) + d_2^2(\widetilde{\psi}(m_G^{SAW}(x)), \widetilde{\psi}(\omega)) \\ &= M_G^{SAW}(m_G^{SAW}(x), x) + d_{SW}^2(m_G^{SAW}(x), \omega), \end{aligned}$$

for all $\omega \in \mathscr{F}$. Hence, we may choose $\beta_0 = 1$.

Under Properties (R1)–(R3), it follows from Petersen and Müller (2019) that

$$d_{SW}(m_G^{SAW}(x), \check{m}_G^{SAW}(x)) = O_p(n^{-1/2}). \tag{51}$$

We mention here that one can circumvent (R2) by directly appealing to the representation of one-dimensional distributions as quantile functions that form a convex subset of L^2. The basic assumptions on the slicewise distributions imply that one can then directly utilize the results of Petersen et al. (2021) on the weighted averages of quantile functions.

Step 2. Here we show that $d_{SW}(m_G^{SAW}(x), \hat{m}_G^{SAW}(x)) = O_p(n^{-1/2})$. As before, $\varphi(\mu_i)$ and $\varphi(\hat{\mu}_i)$ denote the density functions of the i-th sample distribution μ_i and the estimator $\hat{\mu}_i$ (see (27)). From Proposition 11,

$$\max_{i=1,\dots,n} d_2(\varphi(\mu_i), \varphi(\hat{\mu}_i)) = O_p\left(N^{-2/(4+p)}\right),$$

and the derivation of (46) implies

$$d_2(\widetilde{\psi}(\mu_i), \widetilde{\psi}(\hat{\mu}_i)) = d_{SW}(\mu_i, \hat{\mu}_i) \lesssim d_2\left(\varphi(\mu_i), \varphi(\hat{\mu}_i)\right).$$

Consequently,

$$\max_{i=1,\dots,n} d_2(\widetilde{\psi}(\mu_i), \widetilde{\psi}(\hat{\mu}_i)) = O_p\left(N^{-2/(4+p)}\right).$$

Setting $\hat{B}_G(x) = n^{-1}\sum_{i=1}^n s_{iG}(x)\widetilde{\psi}(\hat{\mu}_i)$ and noting that $n^{-1}\sum_{i=1}^n s_{iG}(x) = 1$,

$$d_2(\check{B}_G(x), \hat{B}_G(x)) = O_p\left(N^{-2/(4+p)}\right).$$

Similarly from the derivation of (49), we have

$$\hat{m}_G^{SAW}(x) = \operatorname*{arg\,min}_{\omega \in \mathscr{F}} d_2(\hat{B}_G(x), \widetilde{\psi}(\omega)).$$

By Theorem 5.3 of Deutsch (2012), considering the closedness and convexity of $\widetilde{\psi}(\mathcal{F})$, we obtain

$$d_2(\widetilde{\psi}(\check{m}_G^{SAW}(x)), \widetilde{\psi}(\hat{m}_G^{SAW}(x))) = O_p\left(N^{-2/(4+p)}\right),$$

whence

$$d_{SW}\left(\check{m}_G^{SAW}(x), \hat{m}_G^{SAW}(x)\right) = O_p\left(N^{-2/(4+p)}\right). \tag{52}$$

From (51) and Assumption (P1), we conclude that

$$d_{SW}(m_G^{SAW}(x), \hat{m}_G^{SAW}(x)) = O_p(n^{-1/2}).$$

Uniform convergence results require stronger versions of these properties provided by (R4)–(R6) as stated below. Let $\|\cdot\|_E$ be the Euclidean norm on $\mathbb{R}^q$ and $B > 0$ a constant.

(R4) Almost surely, for all $\|x\|_E \leqslant B$, the objects $m_G^{SAW}(x)$ and $\check{m}_G^{SAW}(x)$ exist and are unique. Additionally, for any $\epsilon > 0$,

$$\inf_{\|x\|_E \leqslant B} \inf_{d_{SW}\left(m_G^{SAW}(x),\omega\right)>\epsilon} \left\{M_G^{SAW}(\omega, x) - M_G^{SAW}\left[m_G^{SAW}(x), x\right]\right\} > 0$$

and there exists $\zeta = \zeta(\epsilon) > 0$ such that

$$\mathrm{pr}\left(\inf_{\|x\|_E \leqslant B} \inf_{d_{SW}\left[\check{m}_G^{SAW}(x),\omega\right]>\epsilon} \left\{\check{M}_G^{SAW}(\omega, x) - \check{M}_G^{SAW}\left[\check{m}_G^{SAW}(x), x\right]\right\} \geqslant \zeta\right) \to 1.$$

(R5) With $B_\delta[m_G^{SAW}(x)]$ and $N\left\{\epsilon, B_\delta\left[m_G^{SAW}(x)\right], d_{SW}\right\}$ as in Condition (R2),

$$\int_0^1 \sup_{\|x\|_E \leqslant B} \left(1 + \log N\left\{\delta\epsilon, B_\delta\left[m_G^{SAW}(x)\right], d_{SW}\right\}\right)^{1/2} d\epsilon = O(1) \quad \text{as } \delta \to 0.$$

(R6) There exist $\tau_0 > 0$, $C_0 > 0$, possibly depending on B, such that

$$\inf_{\substack{\|x\|_E \leqslant B, \\ d_{SW}\left(m_G^{SAW}(x),\omega\right)<\tau_0}} \left\{M_G^{SAW}(\omega, x) - M_G\left[m_G^{SAW}(x), x\right] - C_0 d_{SW}\left[m_G^{SAW}(x), \omega\right]^2\right\} \geqslant 0.$$

The derivation of (R1) and (R3) lead to (R4) and (R6) with $C_0 = 1$. From equation (47), we find that (R5) is satisfied. From Theorem 2 of Petersen and Müller (2019), we conclude that

$$\sup_{\|x\|_E \leqslant B} d_{SW}\left(m_G^{SAW}(x), \check{m}_G^{SAW}(x)\right) = O_p\left(n^{-1/(2+\epsilon)}\right).$$

Note that formula (52) is uniform for $\|x\| \leqslant B$, leading to

$$\sup_{\|x\|_E \leqslant B} d_{SW}\left(m_G^{SAW}(x), \hat{m}_G^{SAW}(x)\right) = O_p\left(n^{-1/(2+\epsilon)}\right).$$

■

H.8 Proof of Theorem 15

Proof We first show that

$$d_{SW}\left(m^{SAW}(x), m_{L,h}^{SAW}(x)\right) = O\left(h^2\right), \tag{53}$$

$$d_{SW}\left(m_{L,h}^{SAW}(x), \check{m}_{L,h}^{SAW}(x)\right) = O_p\left((nh)^{-1/2}\right). \tag{54}$$

We establish the following three properties for the SAW estimator $\check{m}_{L,h}^{SAW}(x)$.

(U1) The minimizers $m^{SAW}(x), m^{SAW}_{L,h}(x)$ and $\check{m}^{SAW}_{L,h}(x)$ exist and are unique, the last almost surely. Additionally, for any $\epsilon > 0$,

$$\inf_{d_{SW}(m^{SAW}(x),\omega)>\epsilon} \{M^{SAW}(\omega, x) - M^{SAW}[m^{SAW}(x), x]\} > 0,$$
$$\liminf_{h\to 0} \inf_{d_{SW}\left(m^{SAW}_{L,h}(x),\omega\right)>\epsilon} \left\{M^{SAW}_{L,h}(\omega, x) - M^{SAW}_{L,h}\left[m^{SAW}_{L,h}(x), x\right]\right\} > 0.$$

(U2) Let $B_\delta[m^{SAW}(x)]$ be the ball of radius δ centered at $m^{SAW}(x)$ with covering number $N\left\{\delta\epsilon, B_\delta[m^{SAW}(x)], d_{SW}\right\}$. Then

$$\int_0^1 \left(1 + \log N\left\{\delta\epsilon, B_\delta[m^{SAW}(x)], d_{SW}\right\}\right)^{1/2} d\epsilon = O(1) \quad \text{as } \delta \to 0.$$

(U3) There exists $\eta_1, \eta_2 > 0$, $\beta_1, \beta_2 > 0$ such that

$$\inf_{d_{SW}(m^{SAW}(x),\omega)<\eta_1} \left\{M^{SAW}(\omega, x) - M^{SAW}[m^{SAW}(x), x] - \beta_1 d_{SW}\left(m^{SAW}(x), \omega\right)^2\right\} \geqslant 0,$$
$$\liminf_{h\to 0} \inf_{d_{SW}\left(m^{SAW}_{L,h}(x),\omega\right)<\eta_2} \left\{M^{SAW}_{L,h}(\omega, x) - M^{SAW}_{L,h}\left[m^{SAW}_{L,h}(x), x\right] - \beta_2 d_{SW}\left(m^{SAW}_{L,h}(x), \omega\right)^2\right\} \geqslant 0.$$

Proof of (U2). Similar to the derivation of (47), we have

$$\int_0^1 \sqrt{1 + \log N\{\delta\epsilon, B_\delta[m^{SAW}(x)], d_{SW}\}}d\epsilon < \int_0^1 \sqrt{1 + \log N\{\delta\epsilon, \mathscr{F}, d_{SW}\}}d\epsilon \leqslant \int_0^1 \sqrt{1 + A_3 A_1 \epsilon^{-p/k}}d\epsilon < \infty.$$

We note that condition (U2) (analogously to (R2)) can be bypassed by using the weak convergence results of Petersen et al. (2021).

Proof of (U1) and (U3). For any distribution $\mu \in \mathscr{F}$, let $\widetilde{\psi}(\mu) \in \widetilde{\psi}(\mathscr{F})$ and $B(x) = E[\widetilde{\psi}(\mu)|X = x]$. Then

$$\begin{aligned} M^{SAW}(\omega, x) &= E[d^2_{SW}(\mu, \omega)|X = x] \\ &= E[d_2^2(\widetilde{\psi}(\mu), \widetilde{\psi}(\omega))|X = x] \\ &= E[d_2^2(\widetilde{\psi}(\mu), B(x))|X = x] + E[d_2^2(B(x), \widetilde{\psi}(\omega))|X = x] \\ &= M^{SAW}\left[B(x), x\right] + d_2^2\left(B(x), \widetilde{\psi}(\omega)\right) \end{aligned}$$

for all $\omega \in \mathscr{F}$, whence

$$m^{SAW}(x) = \arg\min_{\omega\in\mathscr{F}} d_2^2\left(B(x), \widetilde{\psi}(\omega)\right).$$

Set $B_{L,h}(x) = E\left[s_L(X,x,h)\widetilde{\psi}(\mu)\right], \check{B}_{L,h}(x) = n^{-1}\sum_{i=1}^n s_{iL,h}(x)\widetilde{\psi}(\mu_i)$. Considering $E[s_L(x,h)] = 1$ and $n^{-1}\sum_{i=1}^n s_{iL,h}(x) = 1$, one finds

$$m_{L,h}^{SAW}(x) = \arg\min_{\omega\in\mathscr{F}} d_2^2\left(B_{L,h}(x), \widetilde{\psi}(\omega)\right),$$
$$\check{m}_{L,h}^{SAW}(x) = \arg\min_{\omega\in\mathscr{F}} d_2^2\left(\check{B}_{L,h}(x), \widetilde{\psi}(\omega)\right).$$

From the convexity and closedness of space $\widetilde{\psi}(\mathscr{F})$, the minimizers $m^{SAW}(x)$, $m_{L,h}^{SAW}(x)$ and $\check{m}_{L,h}^{SAW}(x)$ exist uniquely for any $x\in\mathbb{R}$, so that (U1) is satisfied. Following the characterization of the best approximation as per (50), we have

$$M^{SAW}(\omega,x) \geqslant M^{SAW}\left(m^{SAW}(x),x\right) + d_{SW}^2\left(m^{SAW}(x),\omega\right).$$

Similarly,

$$M_{L,h}^{SAW}(\omega,x) \geqslant M_{L,h}^{SAW}\left(m_{L,h}^{SAW}(x),x\right) + d_{SW}^2\left(m_{L,h}^{SAW}(x),\omega\right).$$

Thus, (U3) is satisfied with $\beta_1 = \beta_2 = 1$ and η_1, η_2 chosen arbitrarily. Under Conditions (L1)–(L2) and (U1)–(U3), or replacing (U2) by directly applying the weak convergence results of Petersen et al. (2021), it follows from Theorem 3 and Theorem 4 in Petersen and Müller (2019) that

$$d_{SW}\left(m^{SAW}(x), m_{L,h}^{SAW}(x)\right) = O\left(h^2\right),$$
$$d_{SW}\left(m_{L,h}^{SAW}(x), \check{m}_{L,h}^{SAW}(x)\right) = O_p\left((nh)^{-1/2}\right).$$

Hence, we conclude that (53) and (54) are satisfied. Similarly to the derivation of (52), it follows that

$$d_{SW}\left(m_{L,h}^{SAW}(x), \hat{m}_{L,h}^{SAW}(x)\right) = O_p\left((nh)^{-1/2} + N^{-4/(4+p)}\right).$$

From Assumption (P1), we conclude that

$$d_{SW}\left(m_{L,h}^{SAW}(x), \hat{m}_{L,h}^{SAW}(x)\right) = O_p\left((nh)^{-1/2}\right).$$

Next, we provide stronger versions of the assumptions and then show that the previous results can be extended to hold uniformly over a closed interval $\mathcal{T}\subset\mathbb{R}$.

(U4) For all $x\in\mathcal{T}$, the minimizers $m^{SAW}(x), m_{L,h}^{SAW}(x)$ and $\check{m}_{L,h}^{SAW}(x)$ exist and are unique, the latter almost surely. Additionally, for any $\epsilon > 0$,

$$\inf_{x\in\mathcal{T}} \inf_{d_{SW}(m(x),\omega)>\epsilon} \left\{M^{SAW}(\omega,x) - M^{SAW}[m^{SAW}(x),x]\right\} > 0,$$
$$\liminf_{h\to 0} \inf_{x\in\mathcal{T}} \inf_{d_{SW}\left(m_{L,h}^{SAW}(x),\omega\right)>\epsilon} \left\{M_{L,h}^{SAW}(\omega,x) - M_{L,h}^{SAW}\left[m_{L,h}^{SAW}(x),x\right]\right\} > 0,$$

and there exists $\zeta = \zeta(\epsilon) > 0$ such that

$$\mathrm{pr}\left(\inf_{x\in\mathcal{T}} \inf_{d_{SW}\left(\check{m}_{L,n}^{SAW}(x),\omega\right)>\epsilon} \left\{\check{M}_{L,n}^{SAW}(\omega,x) - \check{M}_{L,n}^{SAW}\left[\check{m}_{L,n}^{SAW}(x),x\right]\right\} \geqslant \zeta\right) \to 1.$$

(U5) With $B_\delta\left[m^{SAW}(x)\right]$ and $N\left\{\delta\epsilon, B_\delta[m^{SAW}(x)], d_{SW}\right\}$ as in Condition (R5),

$$\int_0^1 \sup_{x\in\mathcal{T}} \left(1+\log N\left\{\delta\epsilon, B_\delta[m^{SAW}(x)], d_{SW}\right\}\right)^{1/2} d\epsilon = O(1) \quad \text{as } \delta\to 0.$$

(U6) There exists $\tau_1,\tau_2>0$ and $C_1,C_2>0$, such that

$$\inf_{x\in\mathcal{T}} \inf_{d_{SW}(m^{SAW}(x),\omega)<\tau_1} \left\{M^{SAW}(\omega,x) - M^{SAW}[m^{SAW}(x),x] - C_1 d_{SW}(m^{SAW}(x),\omega)^2\right\} \geqslant 0,$$

$$\liminf_{h\to 0} \inf_{x\in\mathcal{T}} \inf_{d_{SW}\left(m_{L,h}^{SAW}(x),\omega\right)<\tau_2} \left\{M_{L,h}^{SAW}(\omega,x) - M_{L,h}^{SAW}\left[m_{L,h}^{SAW}(x),x\right] - C_2 d_{SW}\left(m_{L,h}^{SAW}(x),\omega\right)^2\right\} \geqslant 0.$$

The arguments provided for (U1) and (U3) lead to (U4) and (U6) with $C_1=C_2=1$. Using (47), one finds that (U5) is satisfied. With Assumptions (L1)–(L2) from Theorem 1 of Chen and Müller (2022), we may conclude that

$$\sup_{x\in\mathcal{T}} d_{SW}\left(m^{SAW}(x), m_{L,h}^{SAW}(x)\right) = O\left(h^2\right),$$

$$\sup_{x\in\mathcal{T}} d_{SW}\left(m_{L,h}^{SAW}(x), \check{m}_{L,h}^{SAW}(x)\right) = O_p\left(\max\left\{\left(nh^2\right)^{-1/(2+\epsilon)}, [nh^2(-\log h)^{-1}]^{-1/2}\right\}\right).$$

From (52) which holds uniformly for $x\in\mathcal{T}$ and Assumption (P1),

$$\sup_{x\in\mathcal{T}} d_{SW}\left(m_{L,h}^{SAW}(x), \hat{m}_{L,h}^{SAW}(x)\right) = O_p\left(\max\left\{\left(nh^2\right)^{-1/(2+\epsilon)}, \left[nh^2(-\log h)^{-1}\right]^{-1/2}\right\}\right).$$

■

H.9 Proof of Theorem 7

Proof Let $\gamma_G = \arg\min_{\gamma\in\Gamma_\Theta} M_G^{SWW}(\gamma,x)$ and $\hat{\gamma}_G = \arg\min_{\gamma\in\Gamma_\Theta} \hat{M}_G^{SWW}(\gamma,x)$. Note that the space Γ_Θ is convex and bounded. Then applying the arguments used for Theorem 6 to the metric space (Γ_Θ, d_{DW}), it follows that

$$d_{DW}^2\left(\gamma_G(x), \hat{\gamma}_G(x)\right) = \int_\Theta \int_{[0,1]} \left(\gamma_G(\theta)(s) - \hat{\gamma}_G(\theta)(s)\right)^2 ds d\theta = O_p\left(n^{-1}\right). \tag{55}$$

Recall that G maps a univariate distribution to the corresponding quantile function while G^{-1} maps a quantile function to the corresponding distribution. Let Ψ map the quantile function to the cumulative distribution function. The arguments given in the proof of Proposition 2 in Petersen and Müller (2016) lead to

$$\begin{aligned}&\int_\Theta \int_{\mathbb{R}} \left(\Psi\left(\gamma_G(\theta)\right)(u) - \Psi\left(\hat{\gamma}_G(\theta)\right)(u)\right)^2 du d\theta \\ \lesssim &\int_\Theta \int_{[0,1]} \left(\gamma_G(\theta)(s) - \hat{\gamma}_G(\theta)(s)\right)^2 ds d\theta,\end{aligned}$$

where we observe that by Assumption (G1) the derivatives of $\Psi\left(\gamma_G(\theta)\right)$ and $\Psi\left(\hat{\gamma}_G(\theta)\right)$ are bounded above and below.

From Lemma 21 and the fact that $\varphi(G^{-1}\left(\gamma_G(\theta)\right)) = \partial\Psi\left(\gamma_G(\theta)\right)(u)/\partial u$ we obtain

$$\begin{aligned}
&\int_\Theta \|\varphi(G^{-1}(\gamma_G(\theta)))(u) - \varphi(G^{-1}(\hat{\gamma}_G(\theta)))(u)\|_\infty^2 d\theta \\
\leqslant &\int_\Theta \left(\int_{\mathbb{R}} \left(\Psi(\gamma_G(\theta))(u) - \Psi(\hat{\gamma}_G(\theta)(u))\right)^2 du\right)^{4/7} d\theta \\
\lesssim &\left(\int_\Theta \int_{\mathbb{R}} \left(\Psi(\gamma_G(\theta))(u) - \Psi(\hat{\gamma}_G(\theta)(u))\right)^2 du d\theta\right)^{4/7} = O_p(n^{-4/7}).
\end{aligned}$$

Here the first inequality follows from Lemma 21 and the second inequality from Hölder's inequality and the compactness of Θ. By Assumption (D2) the support of elements in $\mathcal{G}$ is uniformly bounded, whence

$$\int_\Theta \int_{\mathbb{R}} \left(\varphi(G^{-1}(\gamma_G(\theta)))(u) - \varphi(G^{-1}(\hat{\gamma}_G(\theta)))(u)\right)^2 du d\theta = O_p(n^{-4/7}).$$

Recalling the notation $\varrho(\gamma_G)(\theta) = \varphi(G^{-1}(\gamma_G(\theta)))$ and the definition of L^2 distance in (2), we conclude

$$d_2\left(\varrho(\gamma_G), \varrho(\hat{\gamma}_G)\right) = O_p\left(n^{-2/7}\right). \tag{56}$$

By Theorem 2, we have

$$d_\infty\left(m_G^{SWW}(x), \hat{m}_{G,\tau}^{SWW}\right) = O_p\left(C_1(\tau) + C_2(\tau)n^{-2/7}\right).$$

Uniform convergence then follows similarly to the uniform convergence in Theorem 6,

$$\sup_{\|x\| \leqslant B} d_{DW}\left(\gamma_G(x), \hat{\gamma}_G(x)\right) = O_p\left(n^{-1/(2+\epsilon)}\right),$$

for a given $B > 0$ and any $\epsilon > 0$. Since this holds uniformly across x, we have

$$\sup_{\|x\| \leqslant B} d_\infty\left(m_G^{SWW}(x), \hat{m}_{G,\tau}^{SWW}\right) = O_p\left(C_1(\tau) + C_2(\tau)n^{-2/(7+\epsilon)}\right).$$

■

H.10 Proof of Theorem 17

Proof Consider the following notations

$$\begin{aligned}
\gamma_x &= \arg\min_{\gamma \in \Gamma_\Theta} M^{SWW}(\gamma, x), \\
\gamma_{x,L,h} &= \arg\min_{\gamma \in \Gamma_\Theta} M_{L,h}^{SWW}(\gamma, x), \\
\hat{\gamma}_{x,L,h} &= \arg\min_{\gamma \in \Gamma_\Theta} \hat{M}_{L,h}^{SWW}(\gamma, x).
\end{aligned}$$

Note that the space Γ_Θ is convex and closed. Applying the arguments in the proof of Theorem 15 to the metric space (Γ_Θ, d_{DW}), it follows that

$$d_{DW}(\gamma_{x,L,h}, \gamma_x) = O_p\left(h^2\right),$$
$$d_{DW}(\gamma_{x,L,h}, \hat{\gamma}_{x,L,h}) = O_p\left((nh)^{-1/2}\right).$$

In analogy to the derivation of (56), we have

$$d_2\left(\varrho(\gamma_{x,L,h}), \varrho(\gamma_x)\right) = O_p\left(h^{8/7}\right),$$
$$d_2\left(\varrho(\gamma_{x,L,h}), \varrho(\hat{\gamma}_{x,L,h})\right) = O_p\left((nh)^{-2/7}\right).$$

From Assumption (T3), we obtain

$$d_\infty\left(m^{SWW}(x), m^{SWW}_{L,h,\tau}(x)\right) = O_p\left(C_1(\tau) + C_2(\tau)h^{8/7}\right),$$
$$d_\infty\left(m^{SWW}_{L,h,\tau}(x), \hat{m}^{SWW}_{L,h,\tau}(x)\right) = O_p\left(C_2(\tau)\,(nh)^{-2/7}\right).$$

When choosing $h \sim n^{-1/5}$, the resulting convergence rate is

$$d_\infty\left(m^{SWW}(x), \hat{m}^{SWW}_{L,h,\tau}(x)\right) = O_p\left(C_1(\tau) + C_2(\tau)n^{-8/35}\right).$$

Under the additional assumptions (L3)–(L4), the uniform convergence rate similarly is obtained as

$$\sup_{x\in\mathcal{T}} d_\infty(m^{SWW}(x), m^{SWW}_{L,h,\tau}(x)) = O\left(C_1(\tau) + C_2(\tau)h^{8/7}\right),$$
$$\sup_{x\in\mathcal{T}} d_\infty(m^{SWW}_{L,h,\tau}(x), \hat{m}^{SWW}_{L,h,\tau}(x)) = O_p\left(C_2(\tau)\max\left\{\left(nh^2\right)^{-2/(7+\epsilon)}, \left[nh^2(-\log h)^{-1}\right]^{-2/7}\right\}\right),$$

and when taking $h \sim n^{-1/(6+\epsilon)}$,

$$\sup_{x\in\mathcal{T}} d_\infty\left(m^{SWW}(x), \hat{m}^{SWW}_{L,h,\tau}(x)\right) = O_p\left(C_1(\tau) + C_2(\tau)n^{-4/(21+\epsilon)}\right).$$

■

Appendix I. Gaussian Compatibility in Slice-Wise Wasserstein Regression

We verify the compatibility of the slice-wise Wasserstein (SWW) regression framework with multivariate Gaussian distributions in a tractable setting. This result illustrates the role of Assumption (A2) in Section 6, which ensures that the population slice-wise minimizer lies in the image space $\widetilde{\psi}(\mathscr{F})$ and therefore corresponds to a valid multivariate distribution under the inverse slicing transform. The following calculation is for the Radon, or equivalently projection, slicing transform; it should be interpreted as a Gaussian compatibility result for the corresponding untruncated Gaussian family.

Proposition 24 (Gaussian Compatibility) *Let $\psi = \mathcal{R}$ be the Radon transform, and let $\widetilde{\psi}$ be the induced quantile slicing transform. Consider a regression model with scalar predictor $X \in \mathbb{R}$, $E(X^2) < \infty$ and $\mathrm{Var}(X) > 0$, and conditional response*

$$\mu|(X = x) = \mathcal{N}(x\mathbf{1}_p, I_p),$$

where $\mathbf{1}_p \in \mathbb{R}^p$ is the all-ones vector and I_p is the $p \times p$ identity matrix. Let

$$\gamma_{G,x} = \operatorname*{arg\,min}_{\gamma \in \Gamma_\Theta} M_G^{\mathrm{SWW}}(\gamma, x)$$

be the population global SWW minimizer. Then, for each $\theta \in \Theta$,

$$G^{-1}\{\gamma_{G,x}(\theta)\} = \mathcal{N}\left(x \sum_{i=1}^{p} \theta_i, 1\right).$$

Equivalently,

$$\gamma_{G,x}(\theta)(s) = x \sum_{i=1}^{p} \theta_i + \Phi^{-1}(s), \qquad s \in (0,1),$$

where Φ is the standard normal distribution function. Thus, $\gamma_{G,x}$ is precisely the collection of one-dimensional projections of the multivariate Gaussian distribution $\mathcal{N}(x\mathbf{1}_p, I_p)$. Consequently,

$$\gamma_{G,x} \in \widetilde{\psi}(\mathscr{F}),$$

within this Gaussian projection family, verifying the compatibility condition in Assumption (A2).

Proof For the conditional response

$$\mu|(X = x) = \mathcal{N}(x\mathbf{1}_p, I_p),$$

the Radon, or projection, slice in direction $\theta \in \Theta$ is the distribution of $\langle Z, \theta \rangle$ for $Z \sim \mathcal{N}(x\mathbf{1}_p, I_p)$. Since Gaussian distributions are closed under linear projections,

$$\langle Z, \theta \rangle \sim \mathcal{N}\left(\langle x\mathbf{1}_p, \theta \rangle, \theta^\top I_p \theta\right) = \mathcal{N}\left(x \sum_{i=1}^{p} \theta_i, 1\right),$$

where we used $\|\theta\|_2 = 1$.

By Proposition 5, the SWW criterion decouples across slices. Hence, for almost every $\theta \in \Theta$, the slice-wise minimizer satisfies

$$\gamma_{G,x}(\theta) = \operatorname*{arg\,min}_{q \in Q} E\left[s_G(X, x) d_W^2\left(G^{-1}\{\widetilde{\psi}(\mu)(\theta)\}, G^{-1}(q)\right)\right].$$

For the present Gaussian model,

$$G^{-1}\{\widetilde{\psi}(\mu)(\theta)\} = \mathcal{N}\left(X \sum_{i=1}^{p} \theta_i, 1\right).$$

The quantile function of this univariate Gaussian distribution is

$$Q_X^\theta(s) = X\sum_{i=1}^p \theta_i + \Phi^{-1}(s), \qquad s \in (0,1).$$

Since the squared Wasserstein distance between univariate distributions is the squared $L^2(0,1)$ distance between their quantile functions, the weighted Fréchet minimizer in each slice is obtained by the weighted average of the quantile functions:

$$\gamma_{G,x}(\theta)(s) = E\left[s_G(X,x)Q_X^\theta(s)\right].$$

Using

$$s_G(X,x) = 1 + \frac{(X - E[X])(x - E[X])}{\mathrm{Var}(X)},$$

we have

$$E\{s_G(X,x)\} = 1 \quad \text{and} \quad E\{s_G(X,x)X\} = x.$$

Therefore,

$$\begin{aligned}\gamma_{G,x}(\theta)(s) &= E\left[s_G(X,x)\left\{X\sum_{i=1}^p \theta_i + \Phi^{-1}(s)\right\}\right] \\ &= \sum_{i=1}^p \theta_i\, E\{s_G(X,x)X\} + \Phi^{-1}(s)E\{s_G(X,x)\} \\ &= x\sum_{i=1}^p \theta_i + \Phi^{-1}(s).\end{aligned}$$

This is the quantile function of $\mathcal{N}\left(x\sum_{i=1}^p \theta_i, 1\right)$ and thus

$$G^{-1}\{\gamma_{G,x}(\theta)\} = \mathcal{N}\left(x\sum_{i=1}^p \theta_i, 1\right),$$

which is exactly the collection of one-dimensional projections of $\mathcal{N}(x\mathbf{1}_p, I_p)$.

By injectivity of the Radon transform, this collection of projected distributions uniquely determines the multivariate Gaussian distribution $\mathcal{N}(x\mathbf{1}_p, I_p)$. Hence, the slice-wise population minimizer belongs to the image of the induced slicing transform, namely $\gamma_{G,x} \in \widetilde{\psi}(\mathscr{F})$, within this Gaussian projection family. This verifies Assumption (A2) in this setting. ■

Appendix J. Additional Figures

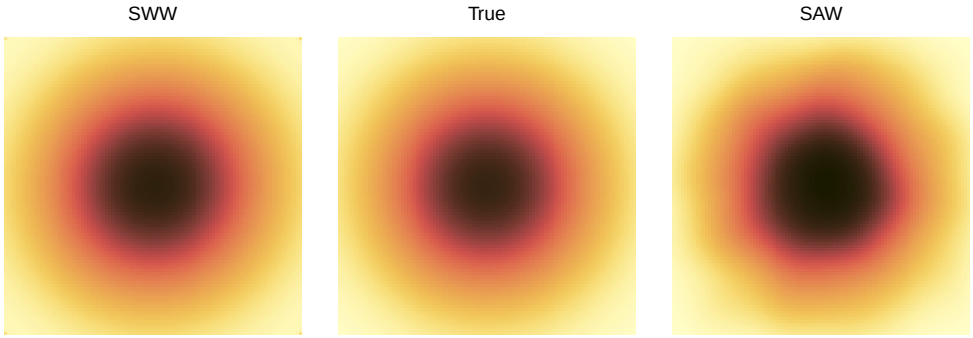


Figure 11: Comparison of the fitted bivariate Gaussian densities obtained by SWW (left) and SAW (right) with the true density (center).

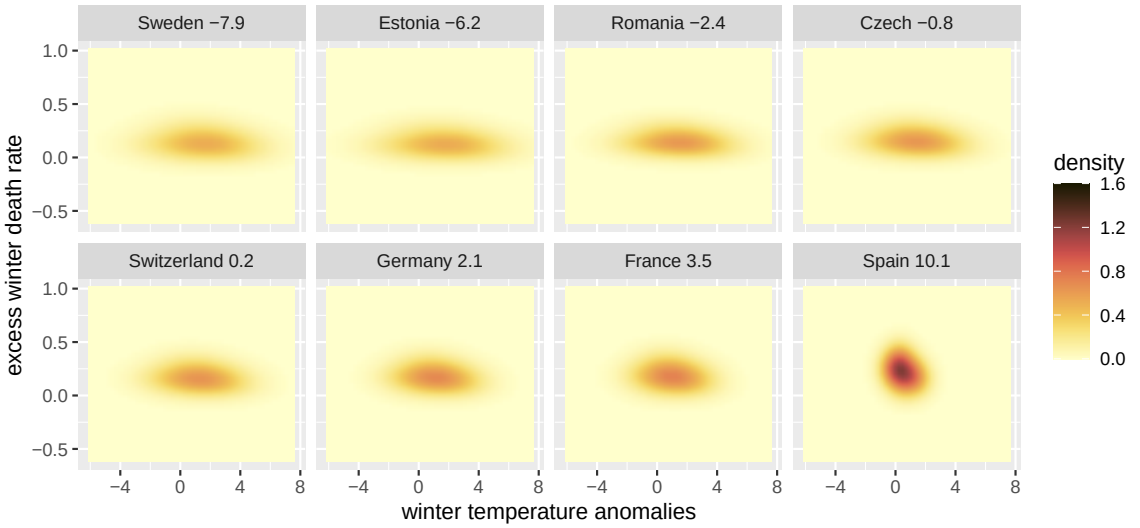


Figure 12: Excess Winter Death Rates: Fitted density surfaces for randomly selected countries obtained by the GSAW version of sliced Wasserstein regression with sliced Wasserstein fraction of variance explained at level 0.28.

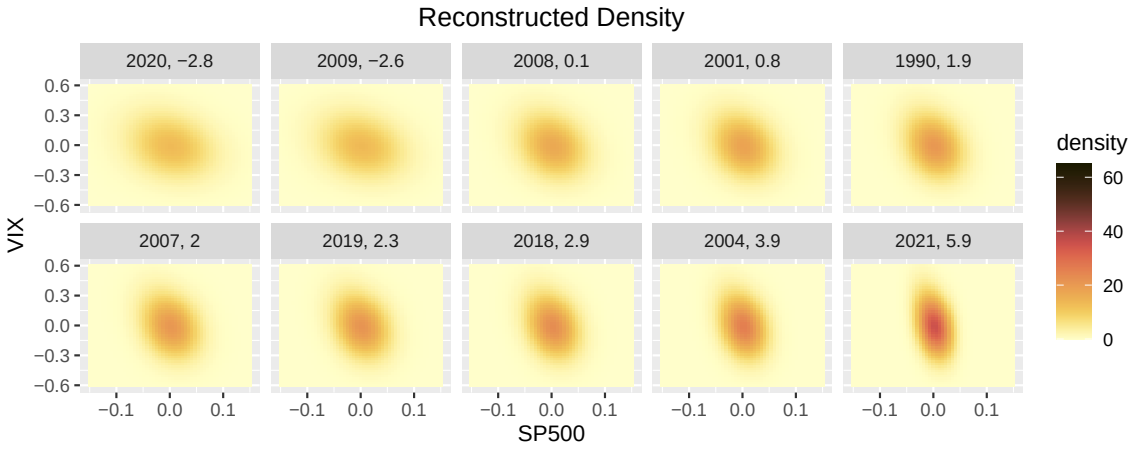


Figure 13: S&P 500 and VIX Index Data: Fitted density surfaces for randomly selected years obtained by the GSAW version of sliced Wasserstein regression with sliced Wasserstein fraction of variance explained at level 0.13.

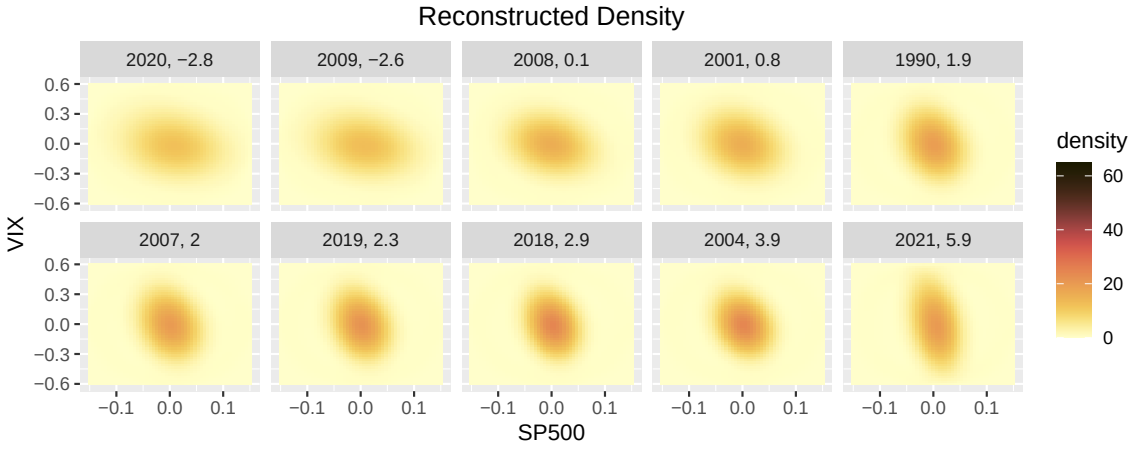


Figure 14: S&P 500 and VIX Index data: Fitted density surfaces for randomly selected years obtained by the LSWW version of sliced Wasserstein regression with sliced Wasserstein fraction of variance explained at level 0.31.

References

H. Abeida, Q. Zhang, J. Li, and N. Merabtine. Iterative sparse asymptotic minimum variance based approaches for array processing. *IEEE Transactions on Signal Processing*, 61(4):933–944, 2012.

M. L. Agranovsky and E. T. Quinto. Injectivity sets for the Radon transform over circles and complete systems of radial functions. *Journal of Functional Analysis*, 139(2):383–414, 1996.

G. Beylkin. The inversion problem and applications of the generalized Radon transform. *Communications on Pure and Applied Mathematics*, 37(5):579–599, 1984.

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

E. Boissard, T. L. Gouic, and J.-M. Loubes. Distribution's template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

R. N. Bracewell. Strip integration in radio astronomy. *Australian Journal of Physics*, 9(2): 198–217, 1956.

H. Chen and H.-G. Müller. Functional data analysis for multivariate distributions through Wasserstein slicing. In *39th Conference on Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

Y. Chen and H.-G. Müller. Uniform convergence of local Fréchet regression with applications to locating extrema and time warping for metric space valued trajectories. *Annals of Statistics*, 50(3):1573–1592, 2022.

Y. Chen, M. Dawson, and H.-G. Müller. Rank dynamics for functional data. *Computational Statistics & Data Analysis*, 149:106963, 2020.

Y. Chen, Z. Lin, and H.-G. Müller. Wasserstein regression. *Journal of the American Statistical Association*, 118(542):869–882, 2023a.

Y. Chen, Y. Zhou, H. Chen, Á. Gajardo, J. Fan, Q. Zhong, P. Dubey, K. Han, S. Bhattacharjee, C. Zhu, S. I. Iao, P. Kundu, A. Petersen, and H.-G. Müller. *frechet: Statistical Analysis for Random Objects and Non-Euclidean Data*, 2023b. URL `https://CRAN.R-project.org/package=frechet`. R package version 0.3.0.

Y.-C. Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.

U. Cherubini, E. Luciano, and W. Vecchiato. *Copula Methods in Finance*. John Wiley & Sons, 2004.

N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *31st Conference on Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

A. Cowling and P. Hall. On pseudodata methods for removing boundary effects in kernel density estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(3):551–563, 1996.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

X. Dai. Statistical inference on the Hilbert sphere with application to random densities. *Electronic Journal of Statistics*, 16(1):700–736, 2022.

F. Deutsch. *Best Approximation in Inner Product Spaces*. CMS Books in Mathematics. Springer New York, 2012. ISBN 9781468492989.

L. Ehrenpreis. *The Universality of the Radon Transform*. Oxford University Press, 10 2003. ISBN 9780198509783. doi: 10.1093/acprof:oso/9780198509783.001.0001.

C. L. Epstein. *Introduction to the Mathematics of Medical Imaging*. SIAM, 2007.

J. Fan and H.-G. Müller. Conditional distribution regression for functional responses. *Scandinavian Journal of Statistics*, 49(2):502–524, 2022.

J. Fan and H.-G. Müller. Conditional Wasserstein barycenters and interpolation/extrapolation of distributions. *IEEE Transactions on Information Theory*, 71(5): 3835–3853, 2024.

D. Finch and S. K. Patch. Determining a function from its mean values over a family of spheres. *SIAM Journal on Mathematical Analysis*, 35(5):1213–1240, 2004.

T. Fowler, R. J. Southgate, T. Waite, R. Harrell, S. Kovats, A. Bone, Y. Doyle, and V. Murray. Excess Winter deaths in Europe: A multi-country descriptive analysis. *European Journal of Public Health*, 25(2):339–345, 2015.

L. Gajek. On improving density estimators which are not bona fide functions. *Annals of Statistics*, 14(4):1612–1618, 1986.

W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177:113–161, 1996.

L. Ghodrati and V. M. Panaretos. Distribution-on-distribution regression via optimal transport maps. *Biometrika*, 109(4):957–974, 2022.

L. Ghodrati and V. M. Panaretos. Transportation of measure regression in higher dimensions. *arXiv preprint arXiv:2305.17503*, 2023.

R. Ghosal, V. R. Varma, D. Volfson, I. Hillel, J. Urbanek, J. M. Hausdorff, A. Watts, and V. Zipunnikov. Distributional data analysis via quantile functions and its application to modeling digital biomarkers of gait in Alzheimer's Disease. *Biostatistics*, 24(3):539–561, 2023.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

D. Guégan and M. Iacopini. Nonparametric forecasting of multivariate probability density functions. *arXiv preprint arXiv:1803.06823*, 2018.

M. L. Hazelton and J. C. Marshall. Linear boundary kernels for bivariate density estimation. *Statistics & Probability Letters*, 79(8):999–1003, 2009.

J. D. Healy. Excess winter mortality in Europe: A cross country analysis identifying key risk factors. *Journal of Epidemiology & Community Health*, 57(10):784–789, 2003.

S. Helgason. *Integral Geometry and Radon Transforms*. Springer New York, 2010. ISBN 9781441960542.

G. T. Herman. *Fundamentals of Computerized Tomography: Image Reconstruction from Projections*. Springer Science & Business Media, 2009.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

S. Horbelt, M. Liebling, and M. Unser. Discretization of the Radon transform and of its inverse by spline convolutions. *IEEE Transactions on Medical Imaging*, 21(4):363–376, 2002.

K. Hron, A. Menafoglio, M. Templ, K. Hruzova, and P. Filzmoser. Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics and Data Analysis*, 94:330–350, 2016.

K. Hron, J. Machalová, and A. Menafoglio. Bivariate densities in Bayes spaces: Orthogonal decomposition and spline representation. *Statistical Papers*, 64:1629–1667, 2023.

H. Jiang. Uniform convergence rates for kernel density estimation. In *International Conference on Machine Learning*, pages 1694–1703. PMLR, 2017.

A. C. Kak and M. Slaney. *Principles of Computerized Tomographic Imaging*. SIAM, 2001.

L. V. Kantorovich. On the translocation of masses. *Journal of Mathematical Sciences*, 133 (4):1381–1382, 2006.

J. Kitagawa and A. Takatsu. Sliced optimal transport: is it a suitable replacement? *Indiana University Mathematics Journal*, to appear, 2026.

S. Kolouri, Y. Zou, and G. K. Rohde. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde. Generalized sliced Wasserstein distances. In *33rd Conference on Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

U. Küchler and S. Tappe. Bilateral gamma distributions and processes in financial mathematics. *Stochastic Processes and their Applications*, 118(2):261–283, 2008.

P. Kuchment. Generalized transforms of Radon type and their applications. In *Proceedings of Symposia in Applied Mathematics*, volume 63, page 67, 2006.

D. B. Madan. Multivariate distributions for financial returns. *International Journal of Theoretical and Applied Finance*, 23(06):2050041, 2020.

D. B. Madan and K. Wang. Asymmetries in financial returns. *International Journal of Financial Engineering*, 4(04):1750045, 2017.

M. Matabuena, A. Petersen, J. C. Vidal, and F. Gude. Glucodensities: A new representation of glucose profiles using distributional data analysis. *Statistical Methods in Medical Research*, 30(6):1445–1464, 2021.

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

R. M. Mersereau and A. V. Oppenheim. Digital reconstruction of multidimensional signals from their projections. *Proceedings of the IEEE*, 62(10):1319–1338, 1974.

D. Meunier, M. Pontil, and C. Ciliberto. Distribution regression with sliced Wasserstein kernels. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 15501–15523. PMLR, 17–23 Jul 2022.

H.-G. Müller. Kernel estimators of zeros and location and size of extrema of regression functions. *Scandinavian Journal of Statistics*, 12:221–232, 1985.

H.-G. Müller and U. Stadtmüller. Multivariate boundary kernels and a continuous least squares principle. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(2):439–458, 1999.

K. Nadjahi, A. Durmus, U. Simsekli, and R. Badeau. Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. In *33rd Conference on Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

K. Nadjahi, A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Simsekli. Statistical and topological properties of sliced probability divergences. In *34th Conference on Neural Information Processing Systems*, volume 33, pages 20802–20812. Curran Associates, Inc., 2020.

F. Natterer. *The Mathematics of Computerized Tomography*. SIAM, 2001.

R. R. Officer. The distribution of stock returns. *Journal of the American Statistical Association*, 67(340):807–812, 1972.

R. Okano and M. Imaizumi. Distribution-on-distribution regression with Wasserstein metric: Multivariate Gaussian case. *Journal of Multivariate Analysis*, 203:105334, 2024.

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

S. Park and D. Slepčev. Geometry and analytic properties of the sliced Wasserstein space. *Journal of Functional Analysis*, 289(7):110975, 2025.

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

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

A. Petersen and H.-G. Müller. Fréchet regression for random objects with Euclidean predictors. *Annals of Statistics*, 47:691–719, 2019.

A. Petersen, X. Liu, and A. A. Divani. Wasserstein F-tests and confidence bands for the Fréchet regression of density response curves. *The Annals of Statistics*, 49(1):590–611, 2021.

A. Petersen, C. Zhang, and P. Kokoszka. Modeling probability density functions as data objects. *Econometrics and Statistics*, 21:159–178, 2022.

G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

M. Quellmalz, R. Beinert, and G. Steidl. Sliced optimal transport on the sphere. *Inverse Problems*, 39(10):105005, 2023.

S. A. Qureshi, S. M. Mirza, and M. Arif. Inverse Radon transform-based image reconstruction using various frequency domain filters in parallel beam transmission tomography. In *2005 Student Conference on Engineering Sciences and Technology*, pages 1–8. IEEE, 2005.

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

J. Radon. Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Akad. Wiss.*, 69:262–277, 1917.

A. Rinaldo and L. Wasserman. Generalized density clustering. *Annals of Statistics*, 38(5): 2678–2722, 2010.

R. M. Rustamov and S. Majumdar. Intrinsic sliced Wasserstein distances for comparing collections of probability distributions on manifolds and graphs. In *International Conference on Machine Learning*, pages 29388–29415. PMLR, 2023.

L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging*, 1(2):113–122, 1982.

E. Tanguy, R. Flamary, and J. Delon. Reconstructing discrete measures from projections. consequences on the empirical sliced Wasserstein distance. *Comptes Rendus. Mathématique*, 362(G10):1121–1129, 2024.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer New York, 2nd edition, 2023.

C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

Q. Zhang, B. Li, and L. Xue. Nonlinear sufficient dimension reduction for distribution-on-distribution regression. *Journal of Multivariate Analysis*, 202:105302, 2024.

Y. Zhou and H.-G. Müller. Network regression with graph Laplacians. *Journal of Machine Learning Research*, 23(1):14383–14423, 2022.

Y. Zhou and H.-G. Müller. Wasserstein regression with empirical measures and density estimation for sparse data. *Biometrics*, 80(4):ujae127, 2024.

C. Zhu and H.-G. Müller. Autoregressive optimal transport models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):1012–1033, 2023.

C. Zhu and H.-G. Müller. Spherical autoregressive models, with application to distributional and compositional time series. *Journal of Econometrics*, 239(2):105389, 2024.

C. Zhu and H.-G. Müller. Geodesic optimal transport regression. *Biometrika*, 113(1): asaf086, 2026.