Biometrics XX, 1–25 DOI: xx
May 2024

Wasserstein regression with empirical measures and density estimation for sparse data

Yidong Zhou

Department of Statistics, University of California, Davis, Davis, CA 95616, U.S.A.

email: ydzhou@ucdavis.edu

and

Hans-Georg Müller

Department of Statistics, University of California, Davis, Davis, CA 95616, U.S.A.

email: hgmueller@ucdavis.edu

Summary: The problem of modeling the relationship between univariate distributions and one or more explanatory variables lately has found increasing interest. Existing approaches proceed by substituting proxy estimated distributions for the typically unknown response distributions. These estimates are obtained from available data but are problematic when for some of the distributions only few data are available. Such situations are common in practice and cannot be addressed with currently available approaches, especially when one aims at density estimates. We show how this and other problems associated with density estimation such as tuning parameter selection and bias issues can be side-stepped when covariates are available. We also introduce a novel version of distribution-response regression that is based on empirical measures. By avoiding the preprocessing step of recovering complete individual response distributions, the proposed approach is applicable when the sample size available for each distribution varies and especially when it is small for some of the distributions but large for others. In this case, one can still obtain consistent distribution estimates even for distributions with only few data by gaining strength across the entire sample of distributions, while traditional approaches where distributions or densities are estimated individually fail, since sparsely sampled densities cannot be consistently estimated. The proposed model is demonstrated to outperform existing approaches through simulations and Environmental Influences on Child Health Outcomes (ECHO) data.

Key words: distributional data analysis, Fréchet mean, multi-cohort study, optimal transport, sample of distributions, Wasserstein distance.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Data consisting of samples of univariate probability distributions are increasingly prevalent across various research areas. Data that contain distributions as a basic unit of observation are encountered in the analysis of mortality (Chen et al., 2023; Ghodrati and Panaretos, 2022), population pyramids (Hron et al., 2016; Bigot et al., 2017), brain connectivity (Petersen and Müller, 2016), and financial returns (Zhang et al., 2022; Zhu and Müller, 2023), among many other applications. This has led to the emerging field of distributional data analysis (Petersen et al., 2022). Distributions, represented as probability density functions, cumulative distribution functions or quantile functions, come with inherent constraints that are not present in traditional functional data. The space where distributions are situated is not a vector space, and linear methods developed for functional data cannot be directly applied. Various approaches have been proposed to address this challenge, including global transformations of univariate distributions to a Hilbert space (Hron et al., 2016; Petersen and Müller, 2016), which however do not take the geometry of the space of distributions into account and are not isometric.

More recently, attention has focused on directly modeling distributions as elements of the Wasserstein space, a geodesic metric space related to optimal transport (Villani, 2003; Panaretos and Zemel, 2020). The pseudo-Riemannian structure of this space can be used to construct isometric exponential maps from the space to tangent bundles, where linear operations can be deployed (Bigot et al., 2017; Chen et al., 2023); however the inverse log map is not defined on the entire tangent space, which causes substantial difficulties and requires an ad-hoc constraint or projection step (Fletcher, 2013; Pegoraro and Beraha, 2022).

A commonly encountered problem is to model the relationship between distributions and one or more explanatory variables. Distribution-response regression problems arise in many modern data analysis settings. Examples include neuroimaging, where varying patterns

of brain connectivity distributions across age and other clinical covariates are of interest (Petersen and Müller, 2016), metabolomics, where the aim is to model the dependence of a metabolite distribution on birth weight (Talská et al., 2018), and also mortality where one is interested in the dependence of age-at-death distributions on economic indicators (Petersen et al., 2022). Using the Wasserstein metric, distribution-response regression can be implemented as a special case of Fréchet regression (Petersen and Müller, 2019), which models the relationship between random objects that lie in a generic metric space as responses and scalar or vector covariates as predictors.

In line with all local and global transformation methods, Fréchet regression requires that each distribution is observed as a distributional object. This is however usually unrealistic in practice as one rarely observes data samples where the atoms of the samples are entire distributions. Instead, one usually has samples of independent data $\{Y_{ij}\}_{j=1}^{N_i}$ that are generated by the distribution ν_i for each i; see Figure 1 in Section 5.1 for a demonstration. To address this, current approaches for virtually all distributional data analysis methods include a prior distribution estimation step where one substitutes a smooth estimate for the unobservable distribution. Commonly this is done through a density estimate such as a kernel estimate or smoothed histogram that is obtained from data $\{Y_{ij}\}_{j=1}^{N_i}$ (Panaretos and Zemel, 2016; Bigot et al., 2018; Petersen and Müller, 2019; Niles-Weed and Berthet, 2022) or a smooth quantile function obtained by smoothing the empirical quantile function (Petersen et al., 2021; Gajardo and Müller, 2021).

In the current literature, the preliminary smoothing step relies on the assumption that the random distribution is absolutely continuous with respect to the Lebesgue measure and thus possesses a density (Petersen et al., 2021; Niles-Weed and Berthet, 2022). To achieve a reasonable rate of convergence, the corresponding random density is assumed to follow certain smoothness or regularity conditions (Niles-Weed and Berthet, 2022). The minimum

number of observations $\min_{1\leqslant i\leqslant n} N_i$, where n is the number of response distributions, is required to increase to infinity at a fast rate, typically faster than n (Chen et al., 2023; Chen and Müller, 2023). This entails the convergence of the density estimation step at a rate that is faster than that of the subsequent Fréchet regression or other operation so that the preliminary estimation step can be ignored in the overall asymptotic analysis. However, the assumption of a fast increase in the number of observations made for each of the distributions across the board is often unrealistic and the need to choose a tuning parameter for each distribution is another downside.

This state of affairs motivates the approach proposed in this paper, namely to use empirical measures $\widehat{\nu}_i = (1/N_i)\sum_{j=1}^{N_i} \delta_{Y_{ij}}$, where $N_i \geqslant 1$ and $\delta_{Y_{ij}}$ denotes the Dirac measure at Y_{ij}, i.e., using the observations made for each random distribution directly, rather than first forming density estimates or other distributional estimates when conducting Fréchet regression. Instead of requiring all N_i to diverge to infinity, we consider N_i as a random variable where $E(N_i^{-1}) \to 0$ as $n \to \infty$. Such a distributional condition on N_i enables us to obtain a consistent density estimate for distributions for which one has only very few observations, where the number of observations may not even increase with n. This becomes possible by harnessing the data across all distributions and exploiting the assumed smooth dependency of the response distributions on the predictors.

The proposed regression model, implemented through least common multiples and supported with asymptotic theory, avoids smoothing bias and tuning parameter choice in the pre-smoothing step and hence is more broadly applicable in practice, especially when the available sample size for each distribution greatly varies (Qiu et al., 2024). A typical example is provided by multi-cohort studies (O'Connor et al., 2022), where the availability and cost of observations highly varies across cohorts (Bonevski et al., 2014). We illustrate such a scenario with Environmental influences on Child Health Outcomes (ECHO) data (Knapp

et al., 2023), where the dependence of body mass index distributions for children across cohorts on demographic covariates is of interest.

The proposed approach offers several key innovations and strengths. First, it demonstrates that the universally employed pre-smoothing step is superfluous for distribution-response Fréchet regression; omitting this step avoids initial smoothing bias and tuning parameter selection. This makes the proposed approach computationally more feasible, especially considering the time-consuming nature of automatically selecting the bandwidth for density estimation in a data-driven manner for each individual measure. Second, the proposed method does not require the random measure to be absolutely continuous, nor does it necessitate any smoothness or regularity conditions on the corresponding random density, if it exists. Third, the method achieves consistent density estimates even for distributions with sparse numbers of observations by leveraging information across the sample. This includes achieving corresponding pointwise and uniform rates of convergence. Fourth, empirical results demonstrate that the proposed regression model outperforms existing smoothing approaches in finite sample situations. In the following, we refer to the proposed approach as Regression with Empirical Measures (REM) and discuss both global and local versions.

The rest of the paper is organized as follows. In Section 2, we introduce some basic background and notation for the Wasserstein space of distributions. The proposed REM approach provides a regression model for empirical measure responses and vector covariates and is introduced in Section 3. Pointwise and uniform rates of convergence for the estimators are established in Section 4. Computational details and simulation results are presented in Section 5. The proposed framework is illustrated in Section 6 using publicly available data on child development from the ECHO study, followed by a brief discussion. Detailed theoretical proofs are in the Supplementary Materials.

2. Preliminaries

For a closed interval Ω with Borel σ-algebra $\mathcal{B}(\Omega)$, we denote the set of probability measures, also referred to as measures or distributions, μ with domain Ω as $\mathcal{P}(\Omega)$ and define the space of measures with finite second moments as

$$\mathcal{W} = \left\{ \mu \in \mathcal{P}(\Omega) : \int_\Omega x^2 \mu(dx) < \infty \right\}.$$

For any measure $\mu \in \mathcal{W}$ with cumulative distribution function F_μ, we consider the quantile function F_μ^{-1} to be the left continuous inverse of F_μ, i.e., $F_\mu^{-1}(\alpha) = \inf\{x \in \Omega : F_\mu(x) \geqslant \alpha\}$, for $\alpha \in (0, 1)$. The space $\mathcal{W}$ is a metric space with the 2-Wasserstein, or simply Wasserstein, distance between two measures $\mu_1, \mu_2 \in \mathcal{W}$ defined as

$$d^2_{\mathcal{W}}(\mu_1, \mu_2) = \inf_{\pi \in \Pi(\mu_1, \mu_2)} \int_{\Omega \times \Omega} |x - y|^2 d\pi(x, y),$$

where $\Pi(\mu_1, \mu_2)$ is the set of joint measures on $\Omega \times \Omega$ with marginals μ_1 and μ_2 (Kantorovich, 1942). It is well known (Villani, 2003) that the Wasserstein distance can be expressed as the L^2 distance between the corresponding quantile functions,

$$d^2_{\mathcal{W}}(\mu_1, \mu_2) = \int_0^1 \{F^{-1}_{\mu_1}(\alpha) - F^{-1}_{\mu_2}(\alpha)\}^2 d\alpha. \tag{1}$$

It can be shown that $\mathcal{W}$ endowed with $d_{\mathcal{W}}$ is a complete and separable metric space, the Wasserstein space (Panaretos and Zemel, 2020; Villani, 2003).

Consider a random element ν taking values in the Wasserstein space $\mathcal{W}$, assumed to be square integrable in the sense that $E\{d^2_{\mathcal{W}}(\nu, \mu)\} < \infty$ for all $\mu \in \mathcal{W}$. The Fréchet mean of ν (Fréchet, 1948), extending the usual notion of mean, is

$$\nu_\oplus = \operatorname*{arg\,min}_{\mu \in \mathcal{W}} E\{d^2_{\mathcal{W}}(\nu, \mu)\},$$

which is well-defined and unique as the Wasserstein space $\mathcal{W}$ is a Hadamard space (Kloeckner, 2010). It follows from (1) that the quantile function of the Fréchet mean $\nu_\oplus$ is $F^{-1}_{\nu_\oplus}(\cdot) = E\{F^{-1}_\nu(\cdot)\}$. In the following, the notation μ refers to fixed measures in $\mathcal{W}$, ν to random

measures, $a \lesssim b$ means that there exists a positive constant C such that $a \leqslant Cb$ and $a \asymp b$ that $a \lesssim b$ and $b \lesssim a$. The Euclidean norm in $\mathbb{R}^p$ is denoted by $\|\cdot\|_E$.

3. Wasserstein Regression with Empirical Measures

Let (Z, ν) be a random pair with joint distribution $\mathcal{F}$ on the product space $\mathbb{R}^p \times \mathcal{W}$. Denote the mean and variance of Z by $\theta = E(Z)$ and $\Sigma = \mathrm{Var}(Z)$, with Σ positive definite. To model the regression relation between random measure ν and vector covariates Z, we adopt the framework of Fréchet regression, a version of conditional Fréchet means designed for the regression of metric-space-valued responses on Euclidean predictors (Petersen and Müller, 2019; Chen and Müller, 2022). Fréchet regression targets the conditional Fréchet mean of ν given $Z = z$,

$$m(z) = \underset{\mu \in \mathcal{W}}{\arg\min}\, E\{d^2_{\mathcal{W}}(\nu, \mu) | Z = z\}.$$

A brief description of Fréchet regression is as follows; for more details see Petersen and Müller (2019). Suppose that $\{(Z_i, \nu_i)\}_{i=1}^n$ are n independent realizations of (Z, ν) with

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^n Z_i, \quad \widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (Z_i - \overline{Z})(Z_i - \overline{Z})^{\mathrm{T}}.$$

Global Fréchet regression can be considered as a generalization of classical multiple linear regression for real-valued responses and targets

$$m_G(z) = \underset{\mu \in \mathcal{W}}{\arg\min}\, E\{s_G(z) d^2_{\mathcal{W}}(\nu, \mu)\}, \tag{2}$$

with weight function $s_G(z) = 1 + (Z - \theta)^{\mathrm{T}} \Sigma^{-1} (z - \theta)$ and empirical version

$$\widetilde{m}_G(z) = \underset{\mu \in \mathcal{W}}{\arg\min}\, \frac{1}{n}\sum_{i=1}^n s_{iG}(z) d^2_{\mathcal{W}}(\nu_i, \mu), \tag{3}$$

for a sample $\{(Z_i, \nu_i)\}_{i=1}^n$, where $s_{iG}(z) = 1 + (Z_i - \overline{Z})^{\mathrm{T}} \widehat{\Sigma}^{-1} (z - \overline{Z})$.

Analogously, local Fréchet regression extends local linear regression to metric-space-valued responses. For the special case of a scalar predictor $Z \in \mathbb{R}$ it targets

$$m_{L,h}(z) = \underset{\mu \in \mathcal{W}}{\arg\min}\, E\{s_L(z, h) d^2_{\mathcal{W}}(\nu, \mu)\}, \tag{4}$$

where $s_L(z,h) = K_h(Z-z)\{u_2 - u_1(Z-z)\}/\sigma_0^2$, $u_j = E\{K_h(Z-z)(Z-z)^j\}$ for $j = 0,1,2$, $\sigma_0^2 = u_0u_2 - u_1^2$, and $K_h(\cdot) = h^{-1}K(\cdot/h)$ with $K(\cdot)$ a continuous symmetric probability density function on $[-1,1]$ and h a bandwidth. The empirical version is

$$\widetilde{m}_{L,h}(z) = \operatorname*{arg\,min}_{\mu\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} s_{iL}(z,h)d^2_{\mathcal{W}}(\nu_i,\mu), \tag{5}$$

where $s_{iL}(z,h) = K_h(Z_i-z)\{\widehat{u}_2 - \widehat{u}_1(Z_i-z)\}/\widehat{\sigma}_0^2$, $\widehat{u}_j = n^{-1}\sum_{i=1}^n K_h(Z_i-z)(Z_i-z)^j$ for $j = 0,1,2$ and $\widehat{\sigma}_0^2 = \widehat{u}_0\widehat{u}_2 - \widehat{u}_1^2$.

In previous work on Fréchet regression (Petersen and Müller, 2019), response distributions in the Wasserstein space $\mathcal{W}$ served as one of the key examples but it has been typically assumed that the measures are fully observed while at the same time they may be randomly perturbed in analogy to usual additive noise models (Chen and Müller, 2022). However, none of this applies in the more realistic situation where one has data $\{Y_{ij}\}_{j=1}^{N_i}$ that are generated from each distribution ν_i. The stopgap solution applied previously is a preprocessing density estimation step (Panaretos and Zemel, 2016; Petersen and Müller, 2016), where it is assumed that the number of data available for all measures increases at a faster rate than the available number of observation points (Z_i,ν_i) in order to preserve asymptotic convergence rates.

However, the intermediate kernel density estimates will be biased and inconsistent if some of the measures generate only very few observations, which then renders this approach infeasible. This provides the motivation to replace the unobservable measures ν_i with the empirical measure $\widehat{\nu}_i = (1/N_i)\sum_{j=1}^{N_i}\delta_{Y_{ij}}$, circumventing the pre-smoothing step and hence eliminating the corresponding tuning parameter selection and smoothing bias. Using empirical measures $\widehat{\nu}_i$ in lieu of the unobservable measures ν_i as responses, the approaches proposed here are global and local Regression with Empirical Measures (REM), replacing (3) and (5) by

$$\widehat{m}_G(z) = \operatorname*{arg\,min}_{\mu\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} s_{iG}(z)d^2_{\mathcal{W}}(\widehat{\nu}_i,\mu), \tag{6}$$

$$\widehat{m}_{L,h}(z) = \operatorname*{arg\,min}_{\mu\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} s_{iL}(z,h)d^2_{\mathcal{W}}(\widehat{\nu}_i,\mu). \tag{7}$$

The computation of the minimizers in (6) and (7) is not straightforward, as some of the weights in these minimization problems are inherently negative. We show in Section 5.1 that adopting least common multiples leads to a simple and efficient algorithm.

4. Asymptotic Properties

We establish pointwise and uniform rates of convergence under the framework of M-estimation. To establish rates of convergence for estimates (6) and (7), we require the following condition.

(C1) The numbers of observations $N_1, N_2, \ldots, N_n$ are i.i.d. random variables with distribution N supported on $\{1, 2, \ldots\}$, where $E(N^{-1}) \to 0$ as $n \to \infty$.

Since measures ν_i are discretely observed, rates of convergence will be affected by the rate at which the sample size N_i increases to infinity. Instead of strictly requiring all sample sizes N_i to diverge to infinity, Condition (C1) imposes a distributional condition. Commonly used discrete distributions, such as the negative binomial distribution and the Poisson distribution, can be employed to satisfy (C1). In cases where sample sizes are highly heterogeneous, it may be appropriate to consider $N = U + 1$ where U follows a negative binomial distribution with a low success probability. Otherwise, a Poisson distribution may be adequate (Fournier and Guillin, 2015). For a more detailed discussion, see Web Appendix A. This setup reflects heterogeneity in sample sizes by allowing some N_i to be very small for certain distributions.

The following result formalizes the consistency of the proposed global REM estimates and provides rates of convergence.

THEOREM 1: *Under Condition (C1), for a fixed $z \in \mathbb{R}^p$, the global REM estimate defined in* (6) *satisfies*

$$d_{\mathcal{W}}\{\widehat{m}_G(z), m_G(z)\} = O_p(n^{-1/2} + \{E(N^{-1/2})\}^{1/2}).$$

Furthermore, for a given constant B it holds that for any $\varepsilon > 0$,

$$\sup_{\|z\|_E \leqslant B} d_{\mathcal{W}}\{\widehat{m}_G(z), m_G(z)\} = O_p(n^{-1/\{2(1+\varepsilon)\}} + \{E(N^{-1/2})\}^{1/2}).$$

All proofs are provided in Web Appendix B. The first term originates from regression with fully observed measures, while the second term, $\{E(N^{-1/2})\}^{1/2}$, represents a stability bound of the difference between $\widetilde{m}_G(z)$ and $\widehat{m}_G(z)$ resulting from the replacement of ν_i by $\widehat{\nu}_i$. A comparable stability bound has been discussed in the context of barycenters in the Wasserstein space (Bigot et al., 2018; Carlier et al., 2024). The pointwise rate of convergence is $O_p(n^{-1/2})$ as long as $E(N^{-1}) \asymp n^{-2}$, which is the same as the well-known optimal rate for multiple linear regression. Similarly, we obtain the following result for local REM, where the kernel and distributional conditions (A1)–(A4) listed in the Appendix are standard for local linear regression.

THEOREM 2: *Under Condition (C1), for a fixed $z \in \mathbb{R}$, the local REM estimate defined in* (7) *satisfies*

$$d_{\mathcal{W}}\{\widehat{m}_{L,h}(z), m(z)\} = O_p(n^{-2/5} + \{E(N^{-1/2})\}^{1/2})$$

for $h \sim n^{-1/5}$ if Conditions (A1) and (A2) hold. Furthermore, for a closed interval $\mathcal{T} \subset \mathbb{R}$,

$$\sup_{z \in \mathcal{T}} d_{\mathcal{W}}\{\widehat{m}_{L,h}(z), m(z)\} = O_p(n^{-1/(3+\varepsilon)} + \{E(N^{-1/2})\}^{1/2})$$

for $h \sim n^{-1/(6+2\varepsilon)}$ and any $\varepsilon > 0$ if Conditions (A3) and (A4) hold.

As long as $E(N^{-1}) \asymp n^{-8/5}$, the local REM estimate achieves the pointwise rate $O_p(n^{-2/5})$, corresponding to the well-known optimal rate for standard local linear regression.

REMARK 1: The second term in the rates of convergence provided in Theorems 1 and 2, $\{E(N^{-1/2})\}^{1/2}$, corresponds to the rate of convergence of the empirical measure in Wasserstein distance. The exponent $-1/2$ at N in the above results can be shown to be optimal without further assumptions on the random measure ν (Bobkov and Ledoux, 2019, Theorem 7.9). However, it can be improved from $-1/2$ to to -1 if ν is assumed to be absolutely

continuous with respect to the Lebesgue measure and the corresponding random density is uniformly bounded below by a positive constant (Bigot et al., 2018; Bobkov and Ledoux, 2019). The pointwise rates of convergence for both global and local REM are thus optimal.

5. Implementation and Simulations

5.1 *Implementation details*

To implement the proposed methods, one needs to solve minimization problems (6) and (7). By standard properties of the $L^2(0,1)$ inner product, (6) and (7) can be simplified to

$$\widehat{m}_G(z) = \operatorname*{arg\,min}_{\mu\in\mathcal{W}} d^2_{L^2}\{\widehat{B}_G(z), F_\mu^{-1}\}, \quad \widehat{B}_G(z) = n^{-1}\sum_{i=1}^n s_{iG}(z) F^{-1}_{\widehat{\nu}_i}, \tag{8}$$

$$\widehat{m}_{L,h}(z) = \operatorname*{arg\,min}_{\mu\in\mathcal{W}} d^2_{L^2}\{\widehat{B}_{L,h}(z), F_\mu^{-1}\}, \quad \widehat{B}_{L,h}(z) = n^{-1}\sum_{i=1}^n s_{iL}(z,h) F^{-1}_{\widehat{\nu}_i},$$

where $d_{L^2}(\cdot,\cdot)$ denotes the L^2 distance; see the proof of Theorem 1 for details. The minimizers $\widehat{m}_G(z)$ and $\widehat{m}_{L,h}(z)$, viewed as projections onto $\mathcal{W}$, exist and are unique for any z by convexity and closedness of $\mathcal{W}$. Similar projections for distributional regression have also been discussed by Chen et al. (2023) and Pegoraro and Beraha (2022). To tackle the challenge of substantial variation in the number of jumps for empirical quantile functions $F^{-1}_{\widehat{\nu}_i}$ due to varying sample sizes N_i, we propose an algorithm based on the least common multiple, which allows us to compute $\widehat{B}_G(z)$ and $\widehat{B}_{L,h}(z)$ as straightforward weighted averages. The algorithm for global REM is outlined in Algorithm 1. The only difference for local REM is the substitution of the weight function $s_{iL}(z,h)$ for $s_{iG}(z)$.

In the first step of Algorithm 1, one needs to calculate $M = \mathrm{lcm}(\{N_i\}_{i=1}^n)$, the least common multiple of the ensemble $\{N_i\}_{i=1}^n$. To control the size of $\mathrm{lcm}(\{N_i\}_{i=1}^n)$ especially when n is large and N_i greatly varies, we incorporate a pre-specified constant M_0 and set $M = \min\{\mathrm{lcm}(\{N_i\}_{i=1}^n), M_0\}$. In our applications, $M_0 = 5,000$ has proven to be adequate. In the second step, if M is not divisible by N_i, we replicate each observation $\lfloor \frac{M}{N_i} \rfloor$ times and randomly select $M - N_i\lfloor \frac{M}{N_i} \rfloor$ observations from the N_i available observations to construct

Algorithm 1: Global Regression with Empirical Measures

Input: data $\{(Z_i, \{Y_{ij}\}_{j=1}^{N_i})\}_{i=1}^n$, and a new predictor level z.

Output: prediction $\widehat{m}_G(z)$.

1 $M \longleftarrow$ the least common multiple of $\{N_i\}_{i=1}^n$;

2 $\{\boldsymbol{V}_i\}_{i=1}^n \longleftarrow$ for each $i = 1, \ldots, n$, stretch $(Y_{i1}, \ldots, Y_{iN_i})^{\mathrm{T}}$ to a vector of length M, $\boldsymbol{V}_i = (V_{i1}, \ldots, V_{iM})^{\mathrm{T}}$, by repeating each element $\frac{M}{N_i}$ times and then arranging in ascending order;

3 $s_{iG}(z) \longleftarrow 1 + (Z_i - \overline{Z})^{\mathrm{T}} \widehat{\Sigma}^{-1}(z - \overline{Z})$ where $\overline{Z} = n^{-1} \sum_{i=1}^n Z_i$ and $\widehat{\Sigma} = n^{-1} \sum_{i=1}^n (Z_i - \overline{Z})(Z_i - \overline{Z})^{\mathrm{T}}$ are the sample mean and variance of $\{Z_i\}_{i=1}^n$, respectively;

4 $\overline{\boldsymbol{V}} = (\overline{V}_1, \ldots, \overline{V}_M)^{\mathrm{T}} \longleftarrow n^{-1} \sum_{i=1}^n s_{iG}(z) \boldsymbol{V}_i$, the element-wise weighted average of $\{\boldsymbol{V}_i\}_{i=1}^n$;

5 $\widehat{m}_G(z) \longleftarrow \arg\min_{\mu \in \mathcal{W}} d_{L^2}^2(\widehat{B}_G^*(z), F_\mu^{-1})$, where $\widehat{B}_G^*(z) = \sum_{m=1}^M \overline{V}_m \mathbf{1}_{q \in (\frac{m-1}{M}, \frac{m}{M}]}$ is a step function defined on $q \in [0, 1]$.

the vector $\boldsymbol{V}_i$, where $\lfloor \frac{M}{N_i} \rfloor$ denotes the greatest integer less than or equal to $\frac{M}{N_i}$. The term $\widehat{B}_G^*(z)$ in the fifth step is a numerical approximation of $\widehat{B}_G(z)$ as defined in (8). This approximation becomes exact when M is divisible by N_i for all i. The fifth step of Algorithm 1 involves solving a minimization problem to project $\widehat{B}_G^*(z)$ onto $\mathcal{W}$. Employing a Riemann sum approximation of the L^2 distance leads to the following convex quadratic optimization problem,

$$\begin{aligned} \text{minimize} \quad & \sum_{m=1}^M (q_m - \overline{V}_m)^2 \\ \text{subject to} \quad & q_1 \leqslant q_2 \leqslant \cdots \leqslant q_{M-1} \leqslant q_M, \\ & q_j \in \Omega, \quad j = 1, \ldots, M. \end{aligned} \tag{9}$$

The solution q^* represents a discretized version of the predicted quantile function. We use the `osqp` package (Stellato et al., 2020) in R to solve the optimization problem (9).

An illustrative example is in Figure 1 to demonstrate the proposed approach and algorithm for the case of global REM. In the top left panel of Figure 1, we depict large variation in sample sizes N_i of the observations available for distributions ν_i, where the smallest N_i is $N_i = 1$. The corresponding empirical quantile functions are illustrated in the top right panel of Figure 1. A simple calculation shows that the global weight function is $s_{iG}(z) = 1+(2i-5)(z-5)/5$ for $i = 1, 2, 3, 4$. The predicted distribution at predictor level z is obtained by calculating $\widehat{B}_G(z)$ in (8), a weighted average of empirical quantile functions, utilizing weights $s_{iG}(z)$, followed by a projection onto $\mathcal{W}$. The weighted average is implemented using Algorithm 1. The predicted distribution $\widehat{m}_G(z)$ obtained in Algorithm 1 is represented as an empirical quantile function, which can subsequently be converted into a density function for improved visualization. We adopted the `qf2pdf` function in the `frechet` package (Chen et al., 2023) to construct densities from quantile functions throughout; there are also alternative options (Niles-Weed and Berthet, 2022).

[Figure 1 about here.]

The bottom two panels of Figure 1 illustrate the predicted distributions using global REM for predictor levels $z = 3$ and $z = 5$. For $z = \overline{Z} = 5$, one obtains the Wasserstein barycenter of the four empirical measures as for this case the global weight function is constant, i.e., $s_{iG}(5) = 1$ for all i. Since $z = 3$ is close to the left endpoint, the estimation requires negative weights where $s_{4G}(3) = -1/5$. In the presence of negative weights, $\widehat{B}_G(3)$ no longer resides in $\mathcal{W}$, necessitating the additional projection step as per (8).

5.2 *Simulations*

To assess the finite sample performance of the proposed methods, we construct a generative model that produces random responses ν along with a Euclidean predictor $Z \in \mathbb{R}$. Consider the true regression function $m(Z)$, represented as a quantile function, $m(Z) = E(\eta|Z) + E(\sigma|Z)\Phi^{-1}(\cdot)$, which corresponds to a Gaussian distribution with mean and stan-

dard deviation depending on Z. The distribution parameters of the true regression function $m(Z)$ are generated conditionally on Z, where the mean and the standard deviation are assumed to follow a normal distribution and a Gamma distribution, respectively. Four different simulation settings are examined as summarized in Table 1. In Settings I and II, the response is generated, on average, as a Gaussian distribution with parameters depending on Z. Setting I corresponds to a global scenario where the response ν depends linearly on the predictor Z as $E(\eta|Z) = \eta_0 + \alpha Z$ and $E(\sigma|Z) = \sigma_0 + \beta Z$. In Setting II, the nonlinear relationships $E(\eta|Z) = \eta_0 + \alpha \sin(\pi Z)$ and $E(\sigma|Z) = \sigma_0 + \beta \sin(\pi Z)$ is considered for the true underlying regression model.

To take non-Gaussian distributions into consideration, we apply an additional transport map to the random response in Settings III and IV. Specifically, after sampling the distribution parameters as in the previous settings, the resulting distribution is transported in Wasserstein space via a random transport map T, uniformly sampled from the collection of maps $T_k(x) = x - \sin(kx)/|k|$ for $k \in \{-2, -1, 1, 2\}$ (Panaretos and Zemel, 2016). Such a transport map significantly complicates Settings III and IV and makes the response distribution non-Gaussian. One can show that the true regression function remains the same after this random transportation.

[Table 1 about here.]

We consider $n = 50, 100, 200, 500, 1000$, with 100 Monte Carlo runs for each of the four simulation settings. For each n, the sample size N_i is independently sampled from the negative binomial distribution with parameters $r = 3$ and $p_n = 10n^{-1}$. In each Monte Carlo run, predictors $\{Z_i\}_{i=1}^n$ are independently sampled from $U(-1, 1)$. For each i, mean and standard deviation of ν_i are generated conditionally on Z_i as described in Table 1, with $\eta_0 = 0, \sigma_0 = 3, \alpha = 3, \beta = 0.5, \tau = 0.5$, and $\kappa = 1$. N_i random observations $\{Y_{ij}\}_{j=1}^{N_i}$ are then independently sampled from ν_i. For the qth Monte Carlo run for a given setting, the quality of

the estimation is quantified by the integrated squared error $\text{ISE}_q = \int_{-1}^{1} d_{\mathcal{W}}^2\{\widehat{m}_q(z), m(z)\}dz$, where $m(\cdot)$ is the true regression function and $\widehat{m}_q(z)$ is the fitted regression function. The bandwidths for the local REM in Settings II and IV are chosen as $n^{-1/5}$.

We also include comparisons with previous two-step procedures (Petersen and Müller, 2019) which involved a prior kernel smoothing step to estimate densities of ν_i from the available discrete observations. These comparison methods are implemented in the `frechet` package (Chen et al., 2023) as `GloDenReg` and `LocDenReg` functions, where the bandwidth for density estimation is automatically selected using the data-driven bandwidth selector proposed in Sheather and Jones (1991).

For additional context, we include regression results for the case of fully observed measures, where it is assumed that measures ν_i are directly available without the need for estimation. The integrated squared errors (ISE) for all Monte Carlo runs and different sample sizes n under the four simulation settings using fully observed measures, the comparison methods, and the proposed methods are summarized in the boxplots in Figure 2. With increasing sample size, ISE is seen to decrease for all simulation settings, demonstrating the convergence of the proposed methods to the target. The analysis of distributional data is particularly challenging when only a limited number of observations are available for some measures, leading to higher ISE for both the proposed and comparison methods. However, this challenge diminishes with larger sample sizes, especially for the proposed methods. Notably, the proposed REM consistently outperforms the comparison methods under all simulation settings, with its advantage becoming more pronounced as the sample size increases. The preliminary smoothing step required for the previous approaches in the literature requires tuning parameter selection which can be difficult and often fails when the data for some of the measures are sparse.

[Figure 2 about here.]

To further evaluate the performance of the proposed methods for discrete distributions, we conducted the same procedure as for the Gaussian distribution, but with a discrete base distribution. Specifically, we considered the true regression function $m(Z) = E(\eta|Z) + E(\sigma|Z)Q(\cdot)$ representing quantile functions in dependence on the covariate Z. Here $Q(\cdot)$ represents the quantile function of a binomial distribution with five trials and a success probability 0.5. Four simulation settings, paralleling those in Table 1 were examined, where we replaced the Gaussian quantile function $\Phi^{-1}(\cdot)$ with the binomial quantile function $Q(\cdot)$ while keeping all other parameters the same. For each n, the sample sizes N_i were independently sampled from a Poisson distribution with parameter $\lambda = 0.3n$. The ISE for all Monte Carlo runs and different n under the four simulation settings using fully observed measures, the comparison methods, and the proposed REM method are summarized in the boxplots in Figure 3. Similar patterns were observed as for the Gaussian distribution, including the convergence of the REM to the target and their superiority over previous two-step methods. Additionally, in the case of discrete distributions where a theoretical density does not exist, the previous two-step methods exhibited worse convergence and more pronounced estimation errors. This finding reaffirms the superiority of the proposed REM over previous two-step methods, particularly for discrete data settings.

[Figure 3 about here.]

6. Cohort-Specific BMI Distribution for US Preschool Children

The Environmental influences on Child Health Outcomes (ECHO) program is an NIH-funded nationwide consortium of multiple cohort studies across the United States designed to investigate the effects of early life exposures on child health and development (Gillman and Blaisdell, 2018). The ECHO program combines existing prenatal and pediatric data collected via cohort-specific protocols with a standardized ECHO-wide protocol that was

established in 2019 (Knapp et al., 2023). The de-identified data on participants contributing extant and new data can be accessed through the National Institute of Child Health and Human Development (NICHD) Data and Specimen Hub (DASH); the version we use here has been made available on August 31, 2021 (Gillman, 2022). As a multi-cohort study, ECHO brings separate cohorts together so that researchers can access information from a large and diverse population of children followed from the prenatal period through adolescence.

It is of interest to study the role of demographic factors in child development, measured in terms of body mass index (BMI), calculated as weight in kilograms divided by height in meters squared. We extracted weight and height measurements of preschool children aged approximately 4 years for 17 cohorts from ECHO, along with demographic information for each cohort, aiming to shed light on how the distribution of BMI of preschool children varies across different cohorts in relation to cohort-specific demographic characteristics. The responses specifically are the cohort-specific distributions of BMI for 4-year-old children in the respective cohort. Cohort-specific covariates that reflect important demographic characteristics of each cohort include average BMI of mothers, average parental education, and proportion of Asians.

Due to differences in the cost and accessibility of visits for different cohorts, there is a significant variation in sample size for each cohort. The amount of accessible weight and height measurements for each cohort varies substantially as a result; see Table 2 for further information. Specifically, there is only one weight and height measurement for girls in the AGA01 cohort, while 160 measurements are available for boys in the AAV01 cohort. We applied global REM to boys and girls separately. Figure 4(A) illustrates the fitted BMI densities for two cohorts with the most (AAV01) and least (AGA01) weight and height measurements, alongside the corresponding BMI measurements shown as ticks. Despite having only one measurement for girls in the AGA01 cohort, the proposed regression model

effectively borrows information from cohorts with more measurements to construct a reasonable density estimate.

For comparison, kernel density estimates are also included in Figure 4(A); these do not account for covariate information. With only a single observation, the kernel density estimator for girls in the AGA01 cohort reduces to the kernel itself, centered at the single observation. This estimator is entirely dependent on the chosen kernel and thus not reliable. Additionally, we implemented the previously proposed two-step procedure where in the first step measures are smoothed (Petersen and Müller, 2019), with the resulting fitted densities shown as dashed curves. These densities are overly influenced by the AGA01 cohort and other cohorts with few observations, leading to an undesirable second peak in the resulting BMI distribution estimates.

[Table 2 about here.]

[Figure 4 about here.]

To further investigate the effects of different demographic factors, predicted BMI densities at different predictor levels are shown in Figure 4(B). Separately for each predictor, the predictor level was varied from the first to the third quantile of the sample, while the other two predictors were held fixed at their median level. We observe that the average BMI of mothers is the most influential predictor, where with increasing average BMI of mothers, the BMI distribution for both 4-year-old boys and girls flattens out and becomes less concentrated. Conversely, higher average parental education is associated with a sharper peak and greater concentration in the BMI distribution for both sexes, with the distribution for boys also shifting slightly to the right. The proportion of Asians in the population seems to have little effect on the BMI distribution of preschool children.

Childhood obesity is continuing to rise in the US, and currently about 13.7 million children are considered to be overweight/obese. For the ECHO data, we can compute the probability

of obesity from the BMI distribution to investigate demographic disparities in early childhood obesity. Obesity is defined as a BMI at or above the 95th percentile for children of the same age and sex. According to the sex-specific BMI-for-age 2000 CDC Growth Charts, the BMI threshold for a 4-year-old boy to be obese is 17.8, while for a girl the value is 18. Based on the predicted BMI distributions (see Figure 4), we calculated the corresponding probabilities of obesity at different predictor levels, with results illustrated in Figure 5. The average BMI of mothers is found to be positively correlated with the prevalence of obesity among four-year-old children, while average parental education and proportion of Asians are negatively associated with obesity. None of these associations establishes a causal effect, but these findings are in agreement with the current literature on obesity, where parental overweight and low socioeconomic status were found to be strong risk factors of obesity in children (Danielzik et al., 2004; Vazquez and Cubbin, 2020), while obesity is less common in Asian children (Anderson and Whitaker, 2009).

[Figure 5 about here.]

7. Discussion

Regression with Empirical Measures (REM) as proposed in this article emerges as a flexible and robust regression model for analyzing discretely observed distributions with heterogeneous numbers of observations. The use of empirical measures circumvents the preliminary smoothing step adopted in the existing literature, thereby avoiding unnecessary smoothing bias. The proposed REM regression model is supported by theoretical justifications, including both pointwise and uniform convergence rates. The pointwise rates are optimal for both global and local versions, corresponding to known optimal rates for Euclidean responses. Additionally, the distributional condition (C1) can be utilized in other distributional data analysis contexts, broadening the applicability of the corresponding theory for real data ex-

amples. Both numerical simulations and real data applications demonstrate the effectiveness and utility of the proposed regression model.

Quantile regression (Koenker and Bassett, 1978), a longstanding research area in statistics, aims to model conditional quantiles of response variables. Traditional quantile regression typically deals with pairs (Z_i, Y_i), where Y_i is a scalar or vector response. In contrast, the proposed distributional regression model considers the response to be the whole distribution, which is often not fully observed. This leads to data in the form $(Z_i, \{Y_{ij}\}_{j=1}^{N_i})$, where $\{Y_{ij}\}_{j=1}^{N_i}$ are discrete observations of the target unobservable measure ν_i. The proposed REM approach is thus not related to quantile regression.

For the univariate distributions considered here, the 2-Wasserstein distance between two probability measures corresponds to the L^2 distance between their corresponding quantile functions. The resulting shape-preserving property makes the Wasserstein distance an attractive choice for modeling univariate distributions. Fréchet Regression for univariate distributions using the Wasserstein distance inherits these merits if the distributions are fully observed. In practice, however, one typically only has samples of independent data from each distribution. While previously proposed two-step methods first obtain density estimates from discrete observations, we demonstrate that using empirical measures provides a more direct and preferable approach, especially when the number of observations available per measure fluctuates widely. The proposed REM method eliminates the need for density estimation, reduces computational complexity and avoids tuning parameter selection and other undesirable features.

Supplementary Materials

Web Appendices A and B referenced in Section 4 are available with this paper at the Biometrics website on Oxford Academic. R code implementing the proposed regression

model is also available at the Biometrics website on Oxford Academic and at `https://github.com/yidongzhou/Wasserstein-regression-with-empirical-measures`.

Funding

We acknowledge support from NSF grants DMS-20146260, DMS-2310450 and IOS-2102953.

Conflict of Interest

None declared.

Data Availability

The data that support the findings in this paper can be accessed through the National Institute of Child Health and Human Development Data and Specimen Hub (`https://dash.nichd.nih.gov/study/417122`).

References

Anderson, S. E. and Whitaker, R. C. (2009). Prevalence of obesity among US preschool children in different racial and ethnic groups. *Archives of Pediatrics & Adolescent Medicine* **163,** 344–348.

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

Bigot, J., Gouet, R., Klein, T., and Lopez, A. (2018). Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line. *Electronic Journal of Statistics* **12,** 2253–2289.

Bobkov, S. and Ledoux, M. (2019). *One-dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances*, volume 261. American Mathematical Society.

Bonevski, B., Randell, M., Paul, C., Chapman, K., Twyman, L., Bryant, J. et al. (2014). Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Medical Research Methodology* **14,** 1–29.

Carlier, G., Delalande, A., and Merigot, Q. (2024). Quantitative stability of barycenters in the Wasserstein space. *Probability Theory and Related Fields* **188,** 1257–1286.

Chen, H. and Müller, H.-G. (2023). Sliced Wasserstein regression. *arXiv preprint arXiv:2306.10601* .

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

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

Chen, Y., Zhou, Y., Chen, H., Gajardo, A., Fan, J., Zhong, Q. et al. (2023). *frechet: Statistical Analysis for Random Objects and Non-Euclidean Data.* R package version 0.3.0.

Danielzik, S., Czerwinski-Mast, M., Langnäse, K., Dilba, B., and Müller, M. (2004). Parental overweight, socioeconomic status and high birth weight are the major determinants of overweight and obesity in 5–7 y-old children: baseline data of the Kiel Obesity Prevention Study (KOPS). *International Journal of Obesity* **28,** 1494–1502.

Fletcher, P. T. (2013). Geodesic regression and the theory of least squares on Riemannian manifolds. *International Journal of Computer Vision* **105,** 171–185.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields* **162,** 707–738.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincaré* **10,** 215–310.

Gajardo, Á. and Müller, H.-G. (2021). Cox point process regression. *IEEE Transactions on Information Theory* **68,** 1133–1156.

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

Gillman, M. (2022). Environmental influences on Child Health Outcomes (ECHO)-wide cohort. https://dash.nichd.nih.gov/study/417122. NICHD Data and Specimen Hub.

Gillman, M. W. and Blaisdell, C. J. (2018). Environmental influences on Child Health Outcomes, a research program of the NIH. *Current Opinion in Pediatrics* **30,** 260–262.

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

Kantorovich, L. V. (1942). On the translocation of masses. *Dokl. Akad. Nauk SSSR (translated version in Journal of Mathematical Sciences, 133, 1381-1382, 2006)* **37,** 227–229.

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

Knapp, E. A., Kress, A. M., Parker, C. B., Page, G. P., McArthur, K., Gachigi, K. K. et al. (2023). The Environmental influences on Child Health Outcomes (ECHO)-wide cohort. *American Journal of Epidemiology* **192,** 1249–1263.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46,** 33–50.

Niles-Weed, J. and Berthet, Q. (2022). Minimax estimation of smooth densities in Wasserstein distance. *Annals of Applied Statistics* **50,** 1519–1540.

O'Connor, M., Spry, E., Patton, G., Moreno-Betancur, M., Arnup, S., Downes, M. et al. (2022). Better together: Advancing life course research through multi-cohort analytic approaches. *Advances in Life Course Research* **53,** 100499.

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

Panaretos, V. M. and Zemel, Y. (2020). *An Invitation to Statistics in Wasserstein Space*. Springer New York.

Pegoraro, M. and Beraha, M. (2022). Projected statistical methods for distributional data on the real line with the Wasserstein metric. *Journal of Machine Learning Research* **23,** 1686–1744.

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

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

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

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

Qiu, J., Dai, X., and Zhu, Z. (2024). Nonparametric estimation of repeated densities with heterogeneous sample sizes. *Journal of the American Statistical Association* **119,** 176–188.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B* **53,** 683–690.

Stellato, B., Banjac, G., Goulart, P., Bemporad, A., and Boyd, S. (2020). OSQP: An operator splitting solver for quadratic programs. *Mathematical Programming Computation* **12,** 637–672.

Talská, R., Menafoglio, A., Machalová, J., Hron, K., and Fišerová, E. (2018). Compositional

regression with functional response. *Computational Statistics & Data Analysis* **123,** 66–85.

Vazquez, C. E. and Cubbin, C. (2020). Socioeconomic status and childhood obesity: a review of literature from the past decade to inform intervention research. *Current Obesity Reports* **9,** 562–570.

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

Zhang, C., Kokoszka, P., and Petersen, A. (2022). Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis* **43,** 30–52.

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

Appendix

Conditions for Theorem 1 and Theorem 2

In the following, $f_Z(\cdot)$ and $f_{Z|\nu}(\cdot, \mu)$ stand for the marginal density of Z and the conditional density of Z given $\nu = \mu$, respectively, and $\mathcal{T}$ is a closed interval in $\mathbb{R}$ with interior $\mathcal{T}^o$.

(A1) The kernel $K(\cdot)$ is a probability density function, symmetric around zero. Furthermore, defining $K_{kj} = \int_{\mathbb{R}} K^k(u)u^j du$, $|K_{14}|$ and $|K_{26}|$ are both finite.

(A2) $f_Z(\cdot)$ and $f_{Z|\nu}(\cdot, \mu)$ both exist and are twice continuously differentiable, the latter for all $\mu \in \mathcal{W}$, and $\sup_{z,\mu} |(\partial^2 f_{Z|\nu}/\partial z^2)(z, \mu)| < \infty$. Additionally, for any open set $U \subset \mathcal{W}$, $\int_U dF_{\nu|Z}(z, \mu)$ is continuous as a function of z.

(A3) The kernel $K(\cdot)$ is a probability density function, symmetric around zero, and uniformly continuous on $\mathbb{R}$. Furthermore, defining $K_{jk} = \int_{\mathbb{R}} K(u)^j u^k du$ for $j, k \in \mathbb{N}$, $|K_{14}|$ and $|K_{26}|$ are both finite. The derivative K' exists and is bounded on the support of K, i.e., $\sup_{K(x)>0} |K'(x)| < \infty$; additionally, $\int_{\mathbb{R}} x^2 |K'(x)|(|x \log |x||)^{1/2} dx < \infty$.

(A4) $f_Z(\cdot)$ and $f_{Z|\nu}(\cdot, \mu)$ both exist and are continuous on $\mathcal{T}$ and twice continuously dif-

ferentiable on $\mathcal{T}^o$, the latter for all $\mu \in \mathcal{W}$. The marginal density $f_Z(\cdot)$ is bounded away from zero on $\mathcal{T}$, $\inf_{z\in\mathcal{T}} f_Z(z) > 0$. The second-order derivative f_Z'' is bounded, $\sup_{z\in\mathcal{T}^o} |f_Z''(z)| < \infty$. The second-order partial derivatives $(\partial^2 f_{Z|\nu}/\partial z^2)(\cdot, \mu)$ are uniformly bounded, $\sup_{z\in\mathcal{T}^o,\mu\in\mathcal{W}} |(\partial^2 f_{Z|\nu}/\partial z^2)(z, \mu)| < \infty$. Additionally, for any open set $U \subset \mathcal{W}$, $\int_U dF_{\nu|Z}(z, \mu)$ is continuous as a function of z; for any $z \in \mathcal{T}$, $M(\mu, z) = E\{d^2_{\mathcal{W}}(\nu, \mu)|Z = z\}$ is equicontinuous, i.e.,

$$\limsup_{x\to z} \sup_{\mu\in\mathcal{W}} |M(\mu, x) - M(\mu, z)| = 0.$$

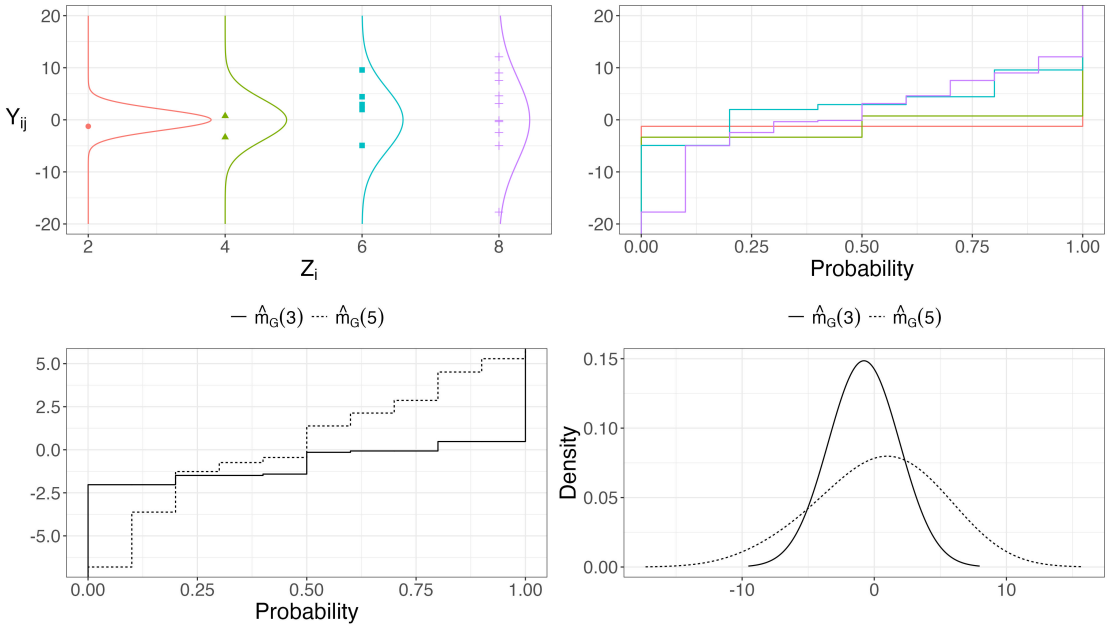

Figure 1. Illustrating global REM, where N_i observations $\{Y_{ij}\}_{j=1}^{N_i}$ are sampled from distributions $\nu_i = N(0, Z_i^2)$ for $i = 1, 2, 3, 4$, with sample sizes $N_1 = 1, N_2 = 2, N_3 = 5, N_4 = 10$ and scalar covariates $Z_1 = 2, Z_2 = 4, Z_3 = 6, Z_4 = 8$. Top left: Visualization of Y_{ij} versus covariate levels Z_i, along with the underlying true densities. Top right: Empirical quantile functions corresponding to the empirical measures $\widehat{\nu}_i = (1/N_i) \sum_{j=1}^{N_i} \delta_{Y_{ij}}$. Bottom left: Predicted quantile functions obtained with global REM for predictor levels $z = 3$ and $z = 5$. Bottom right: Densities corresponding to smoothed versions of the predicted quantile functions.

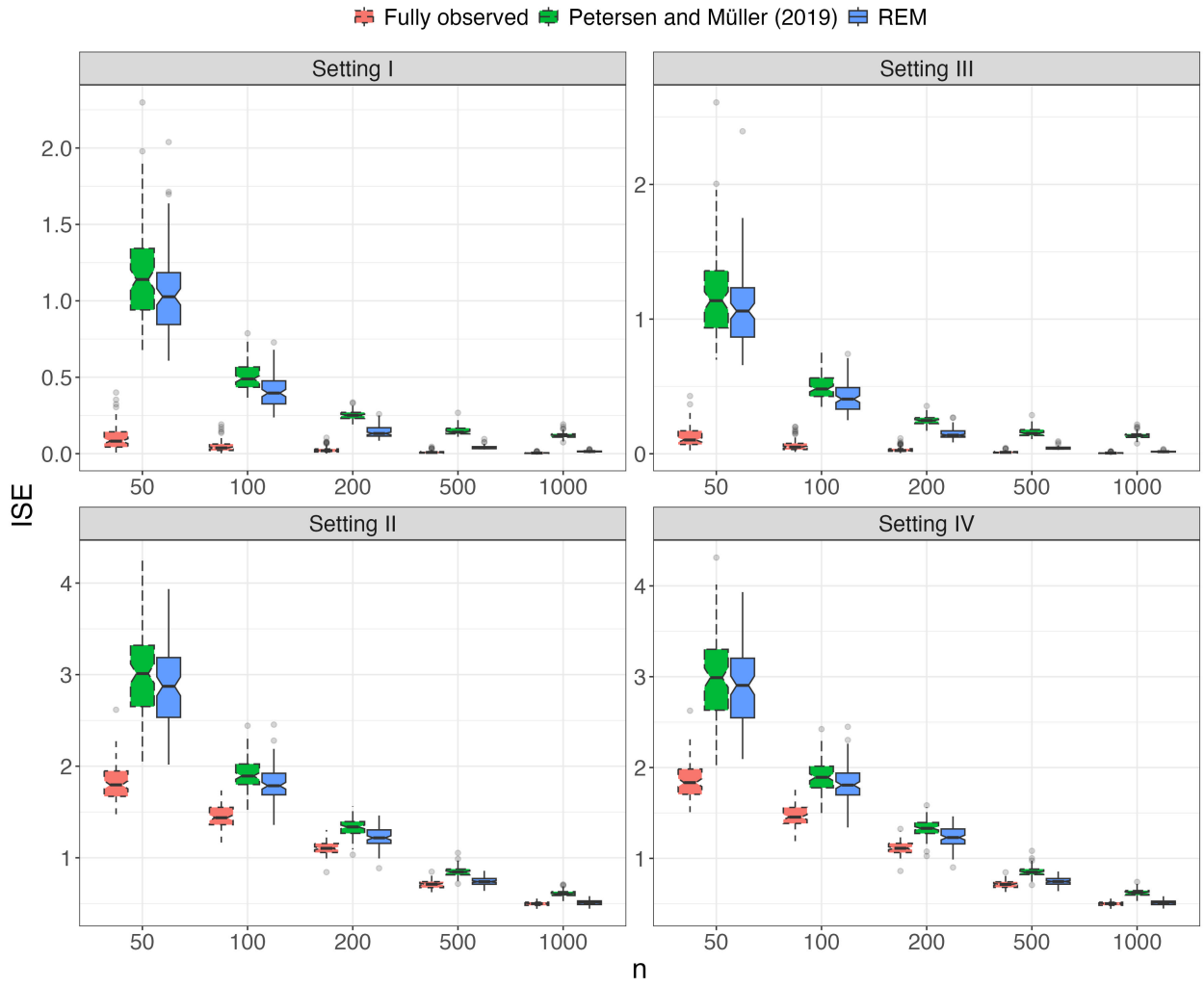

Figure 2. Boxplots of integrated square errors (ISE) using fully observed measures (red, left), presmoothed measures (Petersen and Müller, 2019) (green, middle) and the proposed REM (blue, right), for Gaussian distributions.

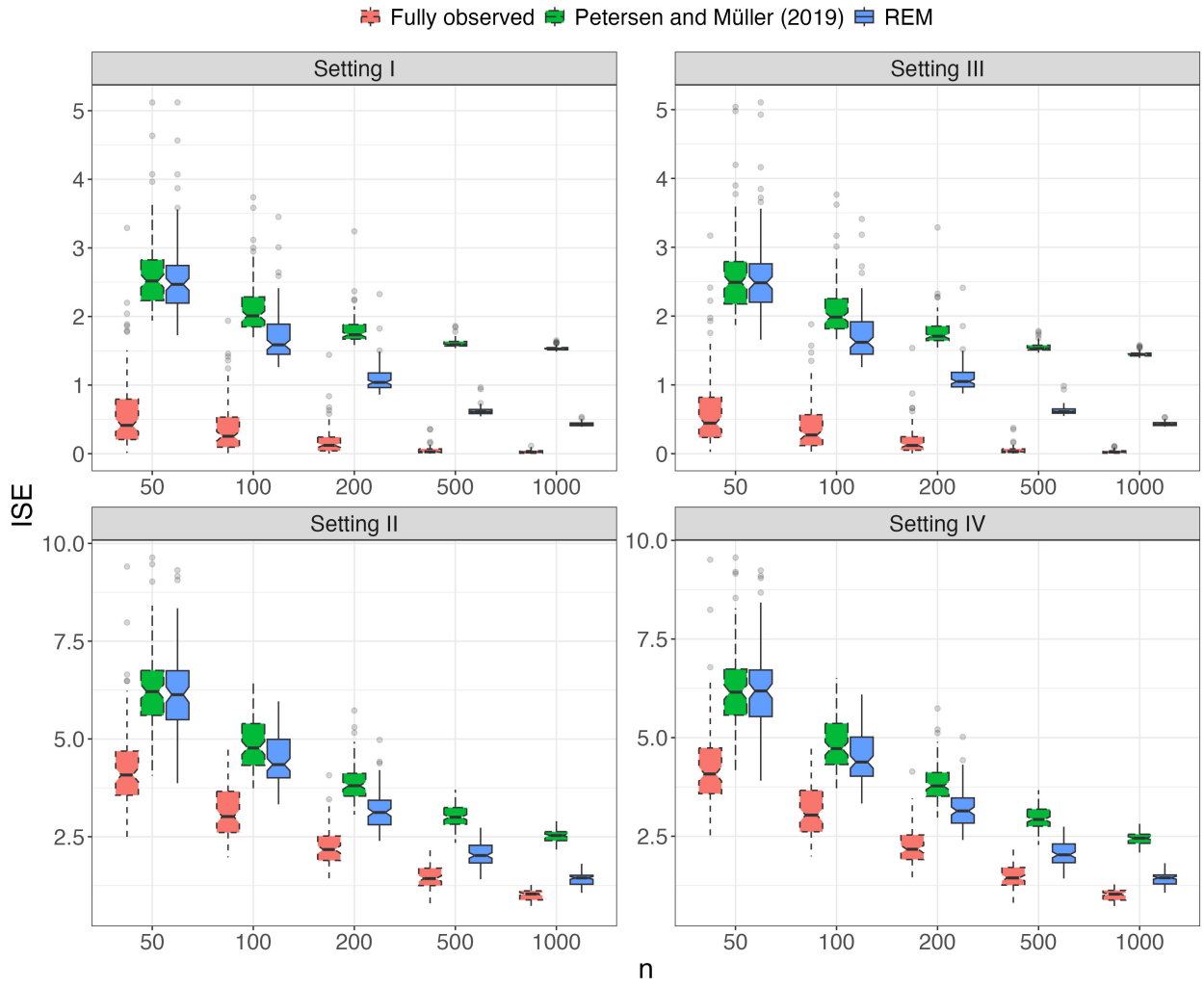

Figure 3. Boxplots of integrated square errors (ISE) using fully observed measures (red, left), presmoothed measures (Petersen and Müller, 2019) (green, middle) and the proposed REM (blue, right), for binomial distributions.

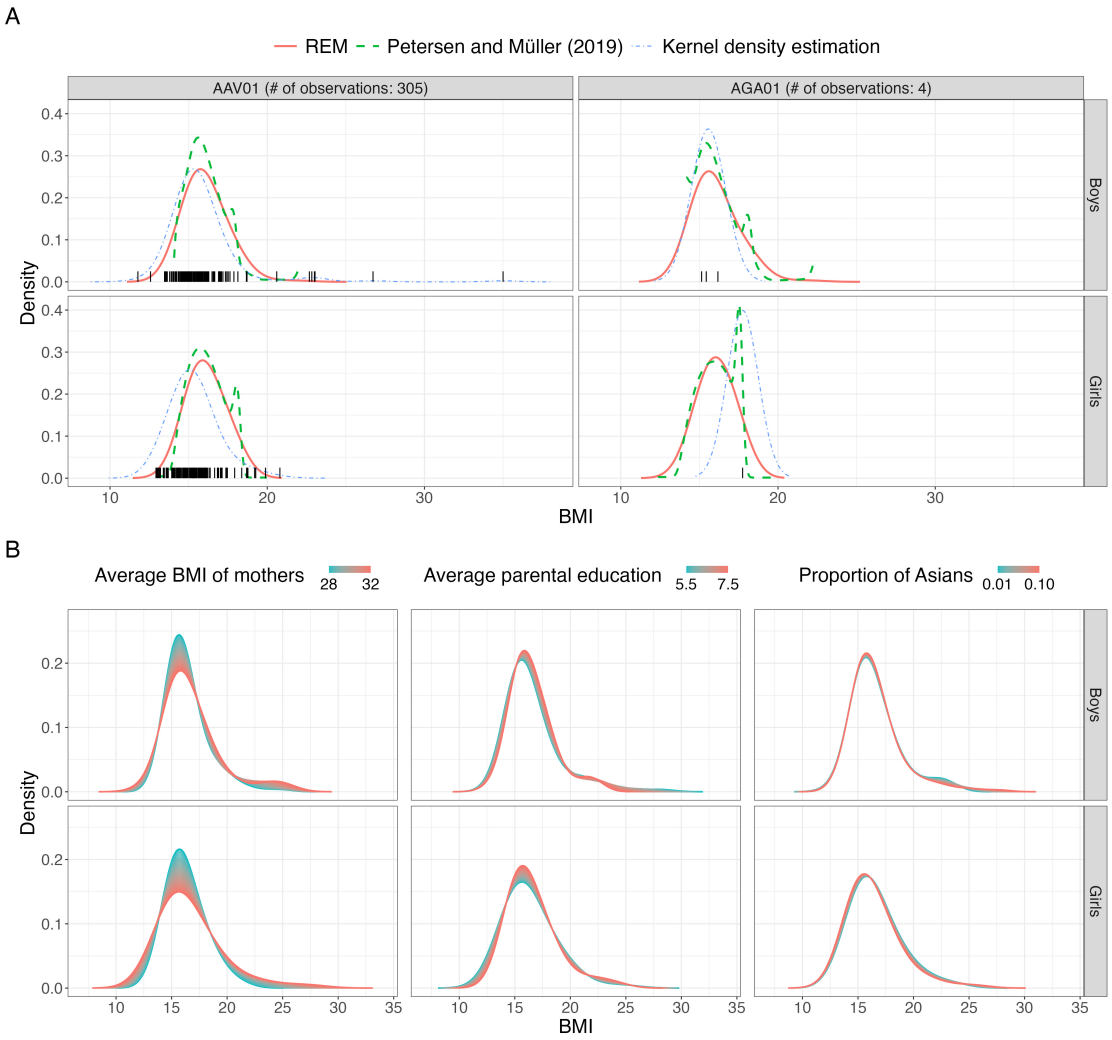

Figure 4. (A) Fitted densities of BMI distributions of US preschool boys and girls for AAV01 and AGA01 cohorts using global REM (solid) and presmoothed measures (Petersen and Müller, 2019) (dashed), along with direct kernel density estimates (dotdash). The corresponding BMI measurements are shown as ticks. (B) Predicted BMI densities of US preschool boys and girls at different predictor levels, obtained with global REM.

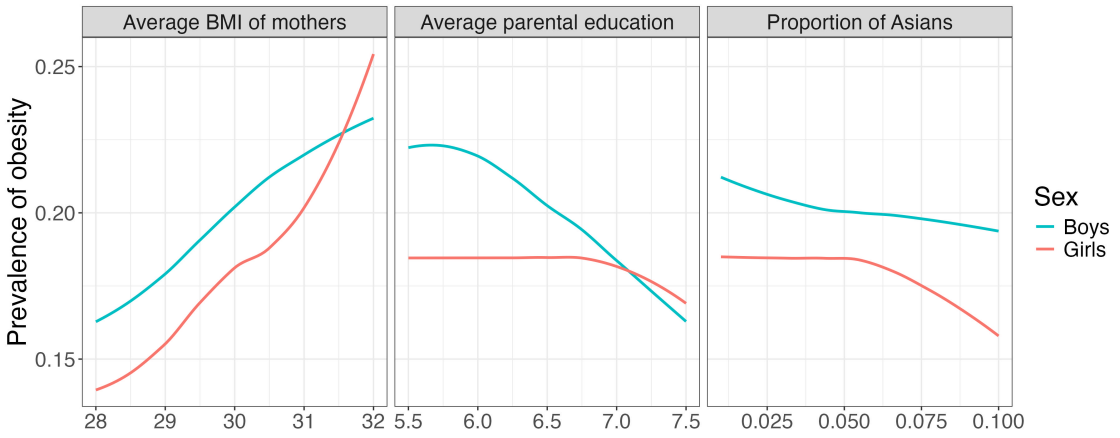

Figure 5. Prevalence of obesity for US preschool boys and girls at different predictor levels.

Table 1

Four simulation settings, where F_ν^{-1} represents the quantile function of the generated random response. The distribution parameters of the random response depend on the predictor Z as indicated for Settings I–IV. An additional transport map T, uniformly sampled from the collection of maps $T_k(\alpha) = \alpha - \sin(k\alpha)/|k|$ for $k \in \{-2, -1, 1, 2\}$, is applied to the resulting random responses in Settings III and IV

	Type	Setting
I	global model	$F_\nu^{-1}(\cdot) = \eta + \sigma\Phi^{-1}(\cdot)$, where $\eta\|Z \sim N(\eta_0 + \alpha Z, \tau^2)$ and $\sigma\|Z \sim \text{Gamma}[\{\sigma_0 + \beta Z\}^2/\kappa, \kappa/\{\sigma_0 + \beta Z\}]$
II	local model	$F_\nu^{-1}(\cdot) = \eta + \sigma\Phi^{-1}(\cdot)$, where $\eta\|Z \sim N(\eta_0 + \alpha \sin(\pi Z), \tau^2)$ and $\sigma\|Z \sim \text{Gamma}[\{\sigma_0 + \beta \sin(\pi Z)\}^2/\kappa, \kappa/\{\sigma_0 + \beta \sin(\pi Z)\}]$
III	global model with random transport	$F_\nu^{-1}(\cdot) = T \circ \{\eta + \sigma\Phi^{-1}(\cdot)\}$, where $\eta\|Z \sim N(\eta_0 + \alpha Z, \tau^2)$ and $\sigma\|Z \sim \text{Gamma}[\{\sigma_0 + \beta Z\}^2/\kappa, \kappa/\{\sigma_0 + \beta Z\}]$
IV	local model with random transport	$F_\nu^{-1}(\cdot) = T \circ \{\eta + \sigma\Phi^{-1}(\cdot)\}$, where $\eta\|Z \sim N(\eta_0 + \alpha \sin(\pi Z), \tau^2)$ and $\sigma\|Z \sim \text{Gamma}[\{\sigma_0 + \beta \sin(\pi Z)\}^2/\kappa, \kappa/\{\sigma_0 + \beta \sin(\pi Z)\}]$

Table 2
Number of weight and height measurements for each cohort

Cohort	AAA01	AAD01	AAE01	AAF01	AAJ01	AAU01	AAV01	AAW02
Boys	139	70	77	35	9	10	160	12
Girls	101	70	81	15	7	8	145	8
Total	240	140	158	50	16	18	305	20
AAX04	AAX06	ADA01	AGA01	AJA02	AJA03	AKA01	AKA02	ALA01
83	124	8	3	44	6	7	76	134
67	110	7	1	38	14	3	71	128
150	234	15	4	82	20	10	147	262

Biometrics XX, 1–8 DOI: xx
May 2024

Supplementary Materials for "Wasserstein regression with empirical measures and density estimation for sparse data" by Yidong Zhou and Hans-Georg Müller

Yidong Zhou

Department of Statistics, University of California, Davis, Davis, CA 95616, U.S.A.

email: ydzhou@ucdavis.edu

and

Hans-Georg Müller

Department of Statistics, University of California, Davis, Davis, CA 95616, U.S.A.

email: hgmueller@ucdavis.edu

Key words: distributional data analysis, Fréchet mean, multi-cohort study, optimal transport, sample of distributions, Wasserstein distance.

This paper has been submitted for consideration for publication in *Biometrics*

Web Appendix A: Distributional Condition on Numbers of Observations

In practice, different study designs and data collection processes may lead to varying underlying distributions for the number of observations N. As demonstrated in Theorems 1 and 2, the rates of convergence for the proposed methods depend solely on the expectation of $N^{-1/2}$, which is bounded by $\{E(N^{-1})\}^{1/2}$, without necessitating a specific distribution for N. This generality makes the proposed methods suitable for a wide range of real data applications. In the following two lemmas, we show that two common discrete distributions, the negative binomial distribution and the Poisson distribution, with suitable parameter choices satisfy the distributional condition (C1) that $E(N^{-1}) \to 0$ as $n \to \infty$.

LEMMA 1: *Suppose the random variable U is negative binomial distributed with a fixed number of successes $r > 2$ and a success probability $p_n > 0$ that varies with n such that $p_n \to 0$ as $n \to \infty$. Specifically, the probability mass function of U is*

$$P(U = k) = \binom{k+r-1}{k}(1-p_n)^k p_n^r, \quad k = 0, 1, 2, \ldots.$$

If the number of observations is $N = U + 1$, then $E(N^{-1}) \leqslant p_n \to 0$ as $n \to \infty$.

Proof. Consider

$$\begin{aligned}
E\left(\frac{r-1}{U+r-1}\right) &= \sum_{k=0}^{\infty} \frac{r-1}{k+r-1}\binom{k+r-1}{k} p_n^r (1-p_n)^k \\
&= \sum_{k=0}^{\infty} \frac{r-1}{k+r-1}\binom{k+r-1}{r-1} p_n^r (1-p_n)^k \\
&\overset{\text{(i)}}{=} \sum_{k=0}^{\infty} \binom{k+r-2}{r-2} p_n^r (1-p_n)^k \\
&= \sum_{k=0}^{\infty} \binom{k+r-2}{k} p_n^r (1-p_n)^k \\
&= p_n \sum_{k=0}^{\infty} \binom{k+r-2}{k} p_n^{r-1} (1-p_n)^k \\
&\overset{\text{(ii)}}{=} p_n.
\end{aligned}$$

Here, (i) is valid since $r \geqslant 2$ and

$$\frac{k}{n}\binom{n}{k} = \binom{n-1}{k-1}$$

for $n \geqslant k \geqslant 1$. Additionally, (ii) follows from the fact that

$$\binom{k+r-2}{r-2} p_n^{r-1}(1-p_n)^k$$

represents the probability mass function of a negative binomial random variable with parameters $r-1$ and p_n, which sums to 1.

Furthermore, note that

$$\frac{1}{U+1} - \frac{r-1}{U+r-1} = \frac{U(2-r)}{(U+1)(U+r-1)} \leqslant 0$$

due to $r \geqslant 2$ and $U \geqslant 0$. Consequently,

$$E\left(\frac{1}{U+1}\right) \leqslant E\left(\frac{r-1}{U+r-1}\right) = p_n.$$

It follows that $E(N^{-1}) \leqslant p_n \to 0$ as $n \to \infty$.

LEMMA 2: *Suppose the random variable U is Poisson distributed with parameter λ_n, where $\lambda_n \to \infty$ as $n \to \infty$. If the number of observations $N = U+1$, then $E(N^{-1}) \leqslant \lambda_n^{-1} \to 0$ as $n \to \infty$.*

Proof. Consider

$$\begin{aligned}
E\left(\frac{1}{U+1}\right) &= \sum_{k=0}^{\infty} \frac{1}{k+1} e^{-\lambda_n} \frac{(\lambda_n)^k}{k!} \\
&= \sum_{k=0}^{\infty} e^{-\lambda_n} \frac{(\lambda_n)^k}{(k+1)!} \\
&= \lambda_n^{-1} \sum_{k=0}^{\infty} e^{-\lambda_n} \frac{(\lambda_n)^{k+1}}{(k+1)!} \\
&= \lambda_n^{-1} \sum_{k=1}^{\infty} e^{-\lambda_n} \frac{(\lambda_n)^k}{k!} \\
&= \lambda_n^{-1}(1 - e^{-\lambda_n}) \\
&\leqslant \lambda_n^{-1}.
\end{aligned}$$

It follows that $E(N^{-1}) \leqslant \lambda_n^{-1} \to 0$ as $n \to \infty$.

Web Appendix B: Proofs of Theoretical Results

Proof of Theorem 1

Proof. We first establish the pointwise rate of convergence for the global REM estimate. By the triangle inequality, we have

$$d_{\mathcal{W}}\{\widehat{m}_G(z), m_G(z)\} \leqslant d_{\mathcal{W}}\{\widehat{m}_G(z), \widetilde{m}_G(z)\} + d_{\mathcal{W}}\{\widetilde{m}_G(z), m_G(z)\}. \tag{1}$$

The second term on the right-hand side, corresponding to the pointwise rate of convergence for global Fréchet regression with fully observed measures, is $O_p(n^{-1/2})$ by Proposition 1 and Theorem 2 in Petersen and Müller (2019).

To analyze the first term, we will utilize simplified expressions of $\widetilde{m}_G(z)$ and $\widehat{m}_G(z)$. Let $\langle \cdot, \cdot \rangle_{L^2}$, $\| \cdot \|_{L^2}$, and $d_{L^2}(\cdot, \cdot)$ be the inner product, norm and distance on the Hilbert space $L^2(0,1)$. For any $\mu \in \mathcal{W}$, the map $Q : \mu \mapsto F_\mu^{-1}$ is an isometry from $\mathcal{W}$ to the subset of $L^2(0,1)$ formed by equivalence classes of left-continuous nondecreasing functions on $(0,1)$. The Wasserstein space $\mathcal{W}$ can thus be viewed as a subset of $L^2(0,1)$, which has been shown to be convex and closed (Bigot et al., 2017).

Define

$$\widetilde{B}_G(z) = \frac{1}{n}\sum_{i=1}^n s_{iG}(z) F_{\nu_i}^{-1}, \quad \widehat{B}_G(z) = \frac{1}{n}\sum_{i=1}^n s_{iG}(z) F_{\widehat{\nu}_i}^{-1},$$

where $F_{\nu_i}^{-1}$ and $F_{\widehat{\nu}_i}^{-1}$ are the quantile functions of ν_i and $\widehat{\nu}_i$, respectively. Since $n^{-1}\sum_{i=1}^n s_{iG}(z) = 1$, it follows that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n s_{iG}(z) d_{\mathcal{W}}^2(\nu_i, \mu) &= \frac{1}{n}\sum_{i=1}^n s_{iG}(z)[d_{L^2}^2\{F_{\nu_i}^{-1}, \widetilde{B}_G(z)\} + d_{L^2}^2\{\widetilde{B}_G(z), F_\mu^{-1}\} \\
&\quad + 2\langle F_{\nu_i}^{-1} - \widetilde{B}_G(z), \widetilde{B}_G(z) - F_\mu^{-1}\rangle_{L^2}] \\
&= \frac{1}{n}\sum_{i=1}^n s_{iG}(z) d_{L^2}^2\{F_{\nu_i}^{-1}, \widetilde{B}_G(z)\} + d_{L^2}^2\{\widetilde{B}_G(z), F_\mu^{-1}\} \\
&\quad + 2 \cdot \frac{1}{n}\sum_{i=1}^n s_{iG}(z)\langle F_{\nu_i}^{-1} - \widetilde{B}_G(z), \widetilde{B}_G(z) - F_\mu^{-1}\rangle_{L^2},
\end{aligned}$$

where the last term

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} s_{iG}(z)\langle F_{\nu_i}^{-1} - \widetilde{B}_G(z), \widetilde{B}_G(z) - F_{\mu}^{-1}\rangle_{L^2} &= \langle \frac{1}{n}\sum_{i=1}^{n} s_{iG}(z) F_{\nu_i}^{-1} - \widetilde{B}_G(z), \widetilde{B}_G(z) - F_{\mu}^{-1}\rangle_{L^2} \\ &= \langle \widetilde{B}_G(z) - \widetilde{B}_G(z), \widetilde{B}_G(z) - F_{\mu}^{-1}\rangle_{L^2} \\ &= 0.\end{aligned}$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^{n} s_{iG}(z) d_{\mathcal{W}}^2(\nu_i, \mu) = \frac{1}{n}\sum_{i=1}^{n} s_{iG}(z) d_{L^2}^2\{F_{\nu_i}^{-1}, \widetilde{B}_G(z)\} + d_{L^2}^2\{\widetilde{B}_G(z), F_{\mu}^{-1}\},$$

whence

$$\begin{aligned}\widetilde{m}_G(z) &= \arg\min_{\mu\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} s_{iG}(z) d_{\mathcal{W}}^2(\nu_i, \mu) \\ &= \arg\min_{\mu\in\mathcal{W}} d_{L^2}^2\{\widetilde{B}_G(z), F_{\mu}^{-1}\}.\end{aligned}$$

One can similarly show that

$$\widehat{m}_G(z) = \arg\min_{\mu\in\mathcal{W}} d_{L^2}^2\{\widehat{B}_G(z), F_{\mu}^{-1}\}.$$

By the convexity and closedness of $\mathcal{W}$, the minimizers $\widetilde{m}_G(z)$ and $\widehat{m}_G(z)$, viewed as projections onto $\mathcal{W}$, exist and are unique for any $z \in \mathbb{R}^p$ (Deutsch, 2001, chap. 3).

Now consider the first term in (1). The contractive property of the projection onto a closed and convex subset in the Hilbert space $L^2(0,1)$ implies that

$$\begin{aligned}d_{\mathcal{W}}\{\widehat{m}_G(z), \widetilde{m}_G(z)\} &= d_{\mathcal{W}}[\arg\min_{\mu\in\mathcal{W}} d_{L^2}^2\{\widehat{B}_G(z), F_{\mu}^{-1}\}, \arg\min_{\mu\in\mathcal{W}} d_{L^2}^2\{\widetilde{B}_G(z), F_{\mu}^{-1}\}] \\ &\leqslant d_{L^2}\{\widehat{B}_G(z), \widetilde{B}_G(z)\} \\ &= d_{L^2}\{\frac{1}{n}\sum_{i=1}^{n} s_{iG}(z) F_{\widehat{\nu}_i}^{-1}, \frac{1}{n}\sum_{i=1}^{n} s_{iG}(z) F_{\nu_i}^{-1}\}.\end{aligned}$$

By the triangle inequality, we have that

$$\begin{aligned}d_{\mathcal{W}}\{\widehat{m}_G(z), \widetilde{m}_G(z)\} &\leqslant \frac{1}{n}\sum_{i=1}^{n} |s_{iG}(z)| d_{L^2}(F_{\widehat{\nu}_i}^{-1}, F_{\nu_i}^{-1}) \\ &= \frac{1}{n}\sum_{i=1}^{n} |s_{iG}(z)| d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i).\end{aligned}$$

Note that $s_{iG}(z) = s_i(z) + W_{0n}(z) + Z_i^{\mathrm{T}} W_{1n}(z)$, where

$$s_i(z) = 1 + (Z_i - \theta)^{\mathrm{T}} \Sigma^{-1}(z - \theta),$$

$$W_{0n}(z) = \theta^{\mathrm{T}} \Sigma^{-1}(z - \theta) - \overline{Z}\widehat{\Sigma}^{-1}(z - \overline{Z}),$$

$$W_{1n}(z) = \widehat{\Sigma}^{-1}(z - \overline{Z}) - \Sigma^{-1}(z - \theta).$$

Both $W_{0n}(z)$ and $\|W_{1n}(z)\|_E$ are $O_p(n^{-1/2})$ by the central limit theorem. It follows that

$$\begin{aligned} d_{\mathcal{W}}\{\widehat{m}_G(z), \widetilde{m}_G(z)\} &\leqslant \frac{1}{n}\sum_{i=1}^{n} |s_i(z) + W_{0n}(z) + Z_i^{\mathrm{T}} W_{1n}(z)| d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i) \\ &\leqslant \frac{1}{n}\sum_{i=1}^{n} |s_i(z)| d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} |W_{0n}(z)| d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i) + \frac{1}{n}\sum_{i=1}^{n} |Z_i^{\mathrm{T}} W_{1n}(z)| d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i). \end{aligned} \tag{2}$$

By the Cauchy-Schwarz inequality, for the first term of (2) we have

$$\begin{aligned} E\{\frac{1}{n}\sum_{i=1}^{n} |s_i(z)| d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i)\} &= \frac{1}{n}\sum_{i=1}^{n} E\{|s_i(z)| d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i)\} \\ &\leqslant \frac{1}{n}\sum_{i=1}^{n} [E\{|s_i(z)|^2\}]^{1/2} [E\{d_{\mathcal{W}}^2(\widehat{\nu}_i, \nu_i)\}]^{1/2}. \end{aligned} \tag{3}$$

By Theorem 7.9 in Bobkov and Ledoux (2019), it follows that

$$E\{d_{\mathcal{W}}^2(\widehat{\nu}_i, \nu_i) | N_i\} \leqslant \frac{1}{2\sqrt{N_i}}.$$

Under Condition (C1), taking expectation with respect to N_i yields

$$E\{d_{\mathcal{W}}^2(\widehat{\nu}_i, \nu_i)\} \leqslant E(\frac{1}{2\sqrt{N_i}}) = E(\frac{1}{2\sqrt{N}}).$$

Since $[E\{|s_i(z)|^2\}]^{1/2}$ are finite, it follows from (3) that

$$E\{\frac{1}{n}\sum_{i=1}^{n} |s_i(z)| d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i)\} = O(\{E(N^{-1/2})\}^{1/2}),$$

which implies

$$\frac{1}{n}\sum_{i=1}^{n} |s_i(z)| d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i) = O_p(\{E(N^{-1/2})\}^{1/2}).$$

By (2) and the fact that $W_{0n}(z)$ and $\|W_{1n}(z)\|_E$ are $O_p(n^{-1/2})$, one has

$$d_{\mathcal{W}}\{\widehat{m}_G(z), \widetilde{m}_G(z)\} = O_p(\{E(N^{-1/2})\}^{1/2}).$$

Combined with (1), we conclude that

$$d_{\mathcal{W}}\{\widehat{m}_G(z), m_G(z)\} = O_p(n^{-1/2} + \{E(N^{-1/2})\}^{1/2}).$$

For the uniform result over $\|z\|_E \leqslant B$, use the fact that $W_{0n}(z)$, $\|W_{1n}(z)\|_E$ are both $O_p(n^{-1/2})$, and $[E\{|s_i(z)|^2\}]^{1/2}$ is $O(1)$, uniformly over $\|z\|_E \leqslant B$. Applying similar arguments leads to

$$\sup_{\|z\|_E \leqslant B} d_{\mathcal{W}}\{\widehat{m}_G(z), \widetilde{m}_G(z)\} = O_p(\{E(N^{-1/2})\}^{1/2}).$$

By Theorem 2 in Petersen and Müller (2019), one has

$$\sup_{\|z\|_E \leqslant B} d_{\mathcal{W}}\{\widetilde{m}_G(z), m_G(z)\} = O_p(n^{-1/\{2(1+\varepsilon)\}})$$

for any $\varepsilon > 0$. Again by the triangle inequality, we conclude that

$$\sup_{\|z\|_E \leqslant B} d_{\mathcal{W}}\{\widehat{m}_G(z), m_G(z)\} = O_p(n^{-1/\{2(1+\varepsilon)\}} + \{E(N^{-1/2})\}^{1/2}).$$

Proof of Theorem 2

Proof. We first establish the pointwise rate of convergence for the local REM estimate. By the triangle inequality, we have

$$d_{\mathcal{W}}\{\widehat{m}_{L,h}(z), m(z)\} \leqslant d_{\mathcal{W}}\{\widehat{m}_{L,h}(z), \widetilde{m}_{L,h}(z)\} + d_{\mathcal{W}}\{\widetilde{m}_{L,h}(z), m(z)\}. \tag{4}$$

The second term on the right-hand side, corresponding to the pointwise rate of convergence for local Fréchet regression with fully observed measures, is $O_p(n^{-2/5})$ under Conditions A1 and A2 by Proposition 1 and Corollary 1 in Petersen and Müller (2019).

To analyze the first term, we will utilize simplified expressions of $\widetilde{m}_{L,h}(z)$ and $\widehat{m}_{L,h}(z)$. By similar arguments as in the proof of Theorem 1 and the fact that $n^{-1}\sum_{i=1}^n s_{iL}(z,h) = 1$, one has

$$\widetilde{m}_{L,h}(z) = \operatorname*{arg\,min}_{\mu \in \mathcal{W}} d^2_{L^2}\{\widetilde{B}_{L,h}(z), F^{-1}_\mu\},$$

$$\widehat{m}_{L,h}(z) = \operatorname*{arg\,min}_{\mu \in \mathcal{W}} d^2_{L^2}\{\widehat{B}_{L,h}(z), F^{-1}_\mu\},$$

where

$$\widetilde{B}_{L,h}(z) = \frac{1}{n}\sum_{i=1}^{n} s_{iL}(z,h)F^{-1}_{\nu_i}, \quad \widehat{B}_{L,h}(z) = \frac{1}{n}\sum_{i=1}^{n} s_{iL}(z,h)F^{-1}_{\widehat{\nu}_i},$$

with $F^{-1}_{\nu_i}$ and $F^{-1}_{\widehat{\nu}_i}$ being the quantile functions of ν_i and $\widehat{\nu}_i$, respectively. By the convexity and closedness of $\mathcal{W}$, the minimizers $\widetilde{m}_{L,h}(z)$ and $\widehat{m}_{L,h}(z)$, viewed as projections onto $\mathcal{W}$, exist and are unique for any $z \in \mathbb{R}$ (Deutsch, 2001, chap. 3).

Note that $s_{iL}(z,h) = s_i(z,h) + W_{0n}K_h(Z_i - z) + W_{1n}K_h(Z_i - z)(Z_i - z)$, where

$$\begin{aligned} s_i(z,h) &= \sigma_0^{-2}K_h(Z_i - z)\{\mu_2 - \mu_1(Z_i - z)\}, \\ W_{0n} &= \frac{\widehat{\mu}_2}{\widehat{\sigma}_0^2} - \frac{\mu_2}{\sigma_0^2}, W_{1n} = \frac{\widehat{\mu}_1}{\widehat{\sigma}_0^2} - \frac{\mu_1}{\sigma_0^2}. \end{aligned}$$

From the proof of Lemma 2 in Petersen and Müller (2019), we have $s_i(z,h) = O_p(1)$, $W_{0n} = O_p\{(nh)^{-1/2}\}$, and $W_{1n} = O_p\{(nh^3)^{-1/2}\}$. Similar to the proof of Theorem 1 one can show that

$$\begin{aligned} d_{\mathcal{W}}\{\widehat{m}_{L,h}(z), \widetilde{m}_{L,h}(z)\} \leqslant \frac{1}{n}\sum_{i=1}^{n} |s_i(z,h)| d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i) + |W_{0n}| \cdot \frac{1}{n}\sum_{i=1}^{n} K_h(Z_i - z) d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i) \\ + |W_{1n}| \cdot \frac{1}{n}\sum_{i=1}^{n} K_h(Z_i - z)|Z_i - z| d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i). \end{aligned} \tag{5}$$

In the proof of Theorem 1, we have shown that

$$E\{d^2_{\mathcal{W}}(\widehat{\nu}_i, \nu_i)\} \leqslant E(\frac{1}{2\sqrt{N}}).$$

It follows that

$$d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i) = O_p(\{E(N^{-1/2})\}^{1/2}).$$

We hence have

$$\frac{1}{n}\sum_{i=1}^{n} K_h(Z_i - z)|Z_i - z|^k d_{\mathcal{W}}(\widehat{\nu}_i, \nu_i) = O_p(h^k\{E(N^{-1/2})\}^{1/2})$$

for $k = 0, 1$. Since $s_i(z,h) = O_p(1)$, $W_{0n} = O_p\{(nh)^{-1/2}\}$, and $W_{1n} = O_p\{(nh^3)^{-1/2}\}$, it follows from (5) that

$$d_{\mathcal{W}}\{\widehat{m}_{L,h}(z), \widetilde{m}_{L,h}(z)\} = O_p(\{E(N^{-1/2})\}^{1/2}).$$

With (4), we conclude that

$$d_{\mathcal{W}}\{\widehat{m}_{L,h}(z), m_{L,h}(z)\} = O_p(n^{-2/5} + \{E(N^{-1/2})\}^{1/2}).$$

For the uniform result over $z \in \mathcal{T}$, we use the fact that $s_i(z,h) = O_p(1)$, $W_{0n} = O_p\{(nh)^{-1/2}\}$, and $W_{1n} = O_p\{(nh^3)^{-1/2}\}$, uniformly over $z \in \mathcal{T}$. Applying similar arguments leads to

$$\sup_{z\in\mathcal{T}} d_{\mathcal{W}}\{\widehat{m}_{L,h}(z), \widetilde{m}_{L,h}(z)\} = O_p(\{E(N^{-1/2})\}^{1/2}).$$

By Theorem 1 in Chen and Müller (2022), under Conditions A3 and A4 one has

$$\sup_{z\in\mathcal{T}} d_{\mathcal{W}}\{\widetilde{m}_{L,h}(z), m(z)\} = O_p(n^{-1/(3+\varepsilon)})$$

for any $\varepsilon > 0$. Again by the triangle inequality, we conclude that

$$\sup_{z\in\mathcal{T}} d_{\mathcal{W}}\{\widehat{m}_{L,h}(z), m(z)\} = O_p(n^{-1/(3+\varepsilon)} + \{E(N^{-1/2})\}^{1/2}).$$

References

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

Bobkov, S. and Ledoux, M. (2019). *One-dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances*, volume 261. American Mathematical Society.

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

Deutsch, F. (2001). *Best Approximation in Inner Product Spaces*, volume 7. Springer.

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