

# Wasserstein Regression with Empirical Measures and Density Estimation for Sparse Data

Yidong Zhou, UC Davis

(Joint work with Hans-Georg Müller)

Preprint: [arXiv:2308.12540](https://arxiv.org/abs/2308.12540)

2024 Joint Statistical Meetings



# Probability Measures

Probability measures (distributions), a prevalent example of non-Euclidean data, arise in numerous applications, including the analysis of mortality, brain connectivity, financial returns, and multi-cohort studies.

How does a distribution change as a function of vector covariates?

## Wasserstein Space

- ▶ For a closed interval  $\Omega$  with Borel  $\sigma$ -algebra  $\mathcal{B}(\Omega)$ , let  $\mathcal{W}$  be the set of probability measures over  $(\Omega, \mathcal{B}(\Omega))$ , with finite second moments.
- ▶ The space  $\mathcal{W}$  is a metric space with the 2-Wasserstein metric,

$$d_{\mathcal{W}}^2(\mu_1, \mu_2) = \int_0^1 \{F_{\mu_1}^{-1}(p) - F_{\mu_2}^{-1}(p)\}^2 dp,$$

where the quantile function  $F_{\mu}^{-1}$  is the left continuous inverse of the cumulative distribution function  $F_{\mu}$ ,

$$F_{\mu}^{-1}(p) = \inf\{x \in \Omega : F_{\mu}(x) \geq p\}, \quad p \in [0, 1].$$

- ▶ Data:  $\{(X_i, \nu_i)\}_{i=1}^n$ , where  $X_i \in \mathbb{R}^p$ ,  $\nu_i \in \mathcal{W}$ .
- ▶ In practice, access to the entire distribution is typically unavailable. Instead, one has samples of independent data  $\{Y_{ij}\}_{j=1}^{N_i}$  that are generated according to the distribution  $\nu_i$ .
- ▶ Current approaches often involve a preliminary distribution estimation step, where a density estimate is substituted for the unobservable distribution.
- ▶ Related work: Bigot et al. (2018), Petersen and Müller (2019), Petersen et al. (2021).

## Limitations of Existing Methods

- ▶ The random distribution is absolutely continuous with respect to the Lebesgue measure and thus possesses a density.
- ▶ The random density is assumed to follow certain smoothness or regularity conditions to achieve a reasonable rate of convergence.
- ▶ The minimum number of observations  $\min_{1 \leq i \leq n} N_i$  is required to increase to infinity at a fast rate, typically faster than  $n$ .

What actions should be taken if the number of observations  $N_i$  varies significantly across different distributions, with some not even increasing with  $n$ ?

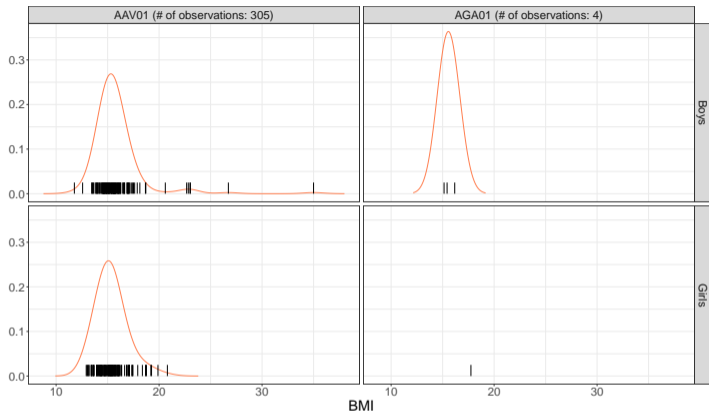
## Example: Cohort-specific BMI distribution for US preschool children

- ▶ The Environmental influences on Child Health Outcomes (ECHO) program.
- ▶ As a **multi-cohort** study, ECHO brings separate cohorts together so that researchers can access data from heterogeneous populations 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**).
- ▶ **Response**: distribution of BMI for 4-year-old children for each cohort.
- ▶ **Predictors**: average BMI of mothers, average parental education, and proportion of Asians.

## Example: Cohort-specific BMI Distribution for US Preschool Children

Table 1: 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



**Figure 1:** Kernel density estimates of BMI distributions of US preschool boys and girls for AAV01 and AGA01 cohorts. The corresponding BMI measurements are shown as ticks.



# Wasserstein Regression with Empirical Measures

Rather than substituting density estimates for the unobservable distributions, we suggest employing **empirical measures**

$\hat{\nu}_i = (1/N_i) \sum_{j=1}^{N_i} \delta_{Y_{ij}}$ , where  $N_i \geq 1$  and  $\delta_{Y_{ij}}$  denotes the Dirac measure at  $Y_{ij}$ .

# Wasserstein Regression with Empirical Measures

The proposed approach

- ▶ avoids smoothing bias and tuning parameter choice in the pre-smoothing step.
- ▶ is 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 distribution.
- ▶ achieves consistent density estimates even for distributions with sparse numbers of observations by leveraging information across the sample.

## Global Regression with Empirical Measures

- ▶ To model the relationship between distributions and vector predictors, a natural target is the conditional Fréchet mean (Petersen & Müller, 2019),

$$m(x) = \arg \min_{\mu \in \mathcal{W}} E\{d_{\mathcal{W}}^2(\nu, \mu) | X = x\}. \quad (1)$$

- ▶ Recall that for scalar responses, linear regression assumes a linear relationship between  $X$  and the conditional mean of  $Y$  given  $X$ , i.e.,

$$E(Y|X) = \beta_0 + \beta_1'X.$$

## Global Regression with Empirical Measures

Using ordinary least squares, the regression function can be alternatively characterized by

$$E(Y|X = x) = \arg \min_{y \in \mathbb{R}} E\{s_G(x)(Y - y)^2\},$$

where the weight function  $s_G(x) = 1 + (X - \theta)' \Sigma^{-1} (x - \theta)$  with  $\theta = E(X)$  and  $\Sigma = \text{Var}(X)$ .

## Global Regression with Empirical Measures

- ▶ Extending linear regression to distributional responses, the regression function is defined as

$$m_G(x) = \arg \min_{\mu \in \mathcal{W}} E\{s_G(x) d_{\mathcal{W}}^2(\nu, \mu)\}. \quad (2)$$

- ▶ Suppose that  $(X_i, \nu_i) \sim F, k = 1, \dots, n$  are independent and define

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

## Global Regression with Empirical Measures

- ▶ The regression function in (2) can be estimated by

$$\tilde{m}_G(x) = \arg \min_{\mu \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^n s_{iG}(x) d_{\mathcal{W}}^2(\nu_i, \mu), \quad (3)$$

where  $s_{iG}(x) = 1 + (X_i - \bar{X})' \hat{\Sigma}^{-1} (x - \bar{X})$ .

- ▶ Using empirical measures  $\hat{\nu}_i$  in lieu of the unobservable measures  $\nu_i$  as responses, the **global Regression with Empirical Measures (REM)** is

$$\hat{m}_G(x) = \arg \min_{\mu \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^n s_{iG}(x) d_{\mathcal{W}}^2(\hat{\nu}_i, \mu). \quad (4)$$

# Rates of Convergence

## Theorem

For a fixed  $x \in \mathbb{R}^p$ , the global REM estimate defined in (4) satisfies

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

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

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

# Cohort-specific BMI Distribution for US Preschool Children

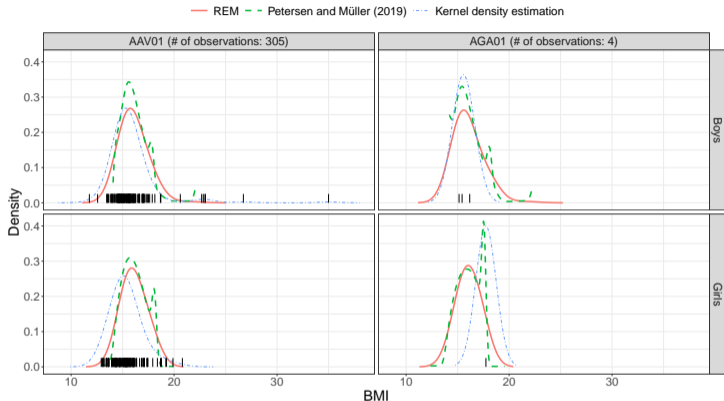


Figure 2: Fitted densities of BMI distributions of US preschool boys and girls for AAV01 and AGA01 cohorts using global REM (solid) and Petersen and Müller (2019) (dashed), along with direct kernel density estimates (dotdash). The corresponding BMI measurements are shown as ticks.



# Cohort-specific BMI Distribution for US Preschool Children

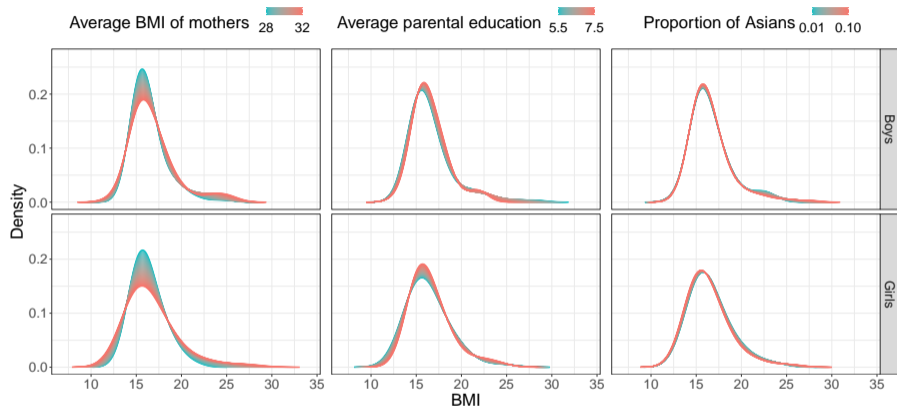






Figure 3: Predicted BMI densities of US preschool boys and girls at different predictor levels.

# References I

-  Bigot, J., Gouet, R., Klein, T., & 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(2), 2253–2289.
-  Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincaré*, 10(4), 215–310.
-  Petersen, A., Liu, X., & Divani, A. A. (2021). Wasserstein  $F$ -tests and confidence bands for the Fréchet regression of density response curves. *Annals of Statistics*, 49(1), 590–611.
-  Petersen, A., & Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *Annals of Statistics*, 47(2), 691–719.

Questions?



## Fréchet Mean and Conditional Fréchet Mean

- ▶ Consider the random pair  $(X, \nu) \sim F$ , where  $X$  takes values in  $\mathbb{R}^p$ ,  $\nu \in \mathcal{W}$  is a distribution.
- ▶ The Fréchet mean and conditional Fréchet mean are generalizations of the mean and conditional mean from the real line to general metric spaces.
- ▶ Let  $Y \in \mathbb{R}$  denote a random variable on the real line.

$$E(Y) = \arg \min_{y \in \mathbb{R}} E\{(Y - y)^2\}, \quad E(Y|X) = \arg \min_{y \in \mathbb{R}} E\{(Y - y)^2|X\}.$$

# Fréchet Mean and Conditional Fréchet Mean

$$E(Y) = \arg \min_{y \in \mathbb{R}} E\{(Y - y)^2\}, \quad E(Y|X) = \arg \min_{y \in \mathbb{R}} E\{(Y - y)^2|X\}.$$

- ▶ Mean  $\rightsquigarrow$  Fréchet mean (Fréchet, 1948):

$$E(Y) \rightsquigarrow \arg \min_{\mu \in \mathcal{W}} E\{d_{\mathcal{W}}^2(\nu, \mu)\}.$$

- ▶ Conditional mean  $\rightsquigarrow$  conditional Fréchet mean (Petersen & Müller, 2019):

$$E(Y|X) \rightsquigarrow \arg \min_{\mu \in \mathcal{W}} E\{d_{\mathcal{W}}^2(\nu, \mu)|X\}.$$

# Cohort-specific BMI Distribution for US Preschool Children

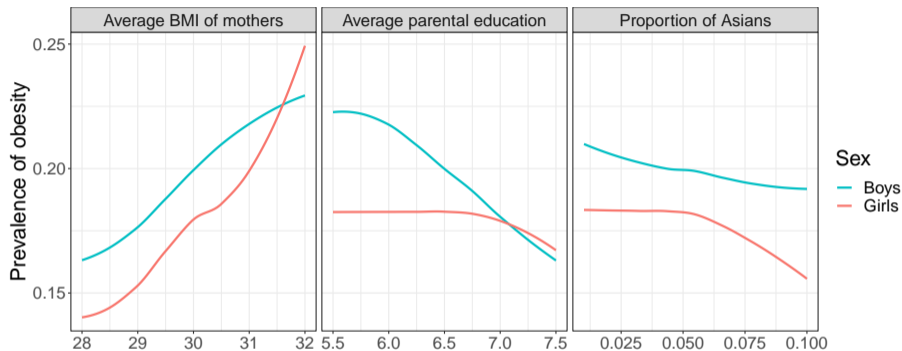


Figure 4: Prevalence of obesity for US preschool boys and girls at different predictor levels.