UC DAVIS
UNIVERSITY OF CALIFORNIA

Deep Fréchet Regression

Yidong Zhou, UC Davis

(joint work with Su I Iao and Hans-Georg Müller)

Columbia FDAWG

- Samples of random objects (non-Euclidean data) that take values in a metric space are becoming increasingly prevalent.
- Due to the absence of a vector space structure, basic statistical tools for scalar/vector data are no longer applicable.
- Examples of random objects:

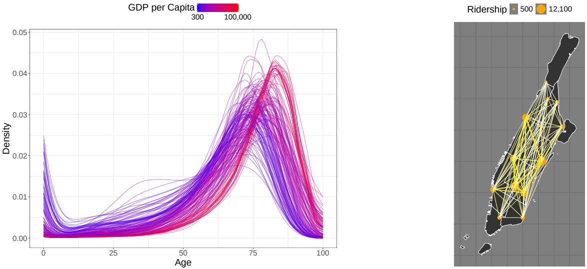


Figure 1: Left: Age-at-death densities of 162 countries in 2015. Right: Taxi traffic network on Jan 1, 2017 in Manhattan.

Motivation

- Modeling the relationship between metric space-valued responses and moderate to high-dimensional Euclidean predictors nonparametrically.

- Two primary challenges:
  - Curse of dimensionality in nonparametric regression.
  - Absence of linear structure in general metric spaces.

Petersen, A., & Müller, H. G. (2019). Fréchet regression for random objects with Euclidean predictors. *Annals of Statistics*, 47(2), 691-719.

Related work:

- Sufficient dimension reduction: Ying and Yu (2022) and Zhang et al. (2024).
- Single index models: Bhattacharjee and Müller (2023) and Ghosal et al. (2023).
- Principal component regression: Song and Han (2023).

These methods, while capable of reducing predictor dimensionality, are confined to the construction of a universal kernel or strong assumptions similar to those for classical single index and linear models.

Fréchet Mean and Conditional Fréchet Mean

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

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

$$E(Y) \rightsquigarrow \underset{\omega \in \Omega}{\arg\min}\, E\{d^2(Y, \omega)\}.$$

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

$$E(Y|\boldsymbol{X}) \rightsquigarrow \underset{\omega \in \Omega}{\arg\min}\, E\{d^2(Y, \omega)|\boldsymbol{X}\}.$$

Fréchet Regression

- To model the relationship between metric space-valued responses and vector predictors, a natural target is the conditional Fréchet mean (Petersen & Müller, 2019),

$$m(\boldsymbol{x}) = \underset{\omega \in \Omega}{\arg\min}\, E\{d^2(Y, \omega) | \boldsymbol{X} = \boldsymbol{x}\}. \tag{1}$$

- In their work, Petersen and Müller (2019) extended both linear and local linear regression to accommodate metric space-valued responses by leveraging the algebraic structure inherent in the predictor space.

Global Fréchet Regression

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

$$E(Y|\boldsymbol{X}) = \beta_0 + \boldsymbol{\beta}_1'\boldsymbol{X}.$$

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

$$E(Y|\boldsymbol{X}=\boldsymbol{x}) = \operatorname*{arg\,min}_{y\in\mathbb{R}} E\{s_G(\boldsymbol{x})(Y-y)^2\},$$

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

Global Fréchet Regression

The global Fréchet regression, extending linear regression to metric space-valued responses, is defined as

$$m_G(\boldsymbol{x}) = \underset{\omega \in \Omega}{\arg\min}\, E\{s_G(\boldsymbol{x}) d^2(Y, \omega)\}, \tag{2}$$

where the response Y is a random object residing in the metric space Ω.

Recall that the standard linear regression is

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

Global Fréchet Regression

- Suppose that $(\boldsymbol{X}_i, Y_i) \sim F, i = 1, \ldots, n$ are independent and define

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

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

$$\hat{m}_G(\boldsymbol{x}) = \operatorname*{arg\,min}_{\omega \in \Omega} \frac{1}{n}\sum_{i=1}^{n} s_{iG}(\boldsymbol{x}) d^2(Y_i, \omega), \tag{3}$$

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

Local Fréchet Regression

The local Fréchet regression, a generalization of local linear regression to metric space-valued responses, follows a similar form but employs a different weight function.

Deep Fréchet Regression

- (Ω, d): a metric space with metric $d : \Omega \times \Omega \mapsto [0, \infty)$.
- $\mathcal{M} \subset \Omega$: a manifold isometric to a subspace of $\mathbb{R}^r$.
- $\boldsymbol{Z}^0 = \psi(Y) \in \mathbb{R}^r$: the low-dimensional representation of $Y \in \mathcal{M}$.
- $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$: n independent copies of the random pair $(\boldsymbol{X}, Y)$ in $\mathbb{R}^p \times \mathcal{M}$, where $\mathcal{M} \subset \Omega$.

Deep Fréchet Regression

- (Ω, d): a metric space with metric $d : \Omega \times \Omega \mapsto [0, \infty)$.
- $\mathcal{M} \subset \Omega$: a manifold isometric to a subspace of $\mathbb{R}^r$.
- $\boldsymbol{Z}^0 = \psi(Y) \in \mathbb{R}^r$: the low-dimensional representation of $Y \in \mathcal{M}$.
- $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$: n independent copies of the random pair $(\boldsymbol{X}, Y)$ in $\mathbb{R}^p \times \mathcal{M}$, where $\mathcal{M} \subset \Omega$.

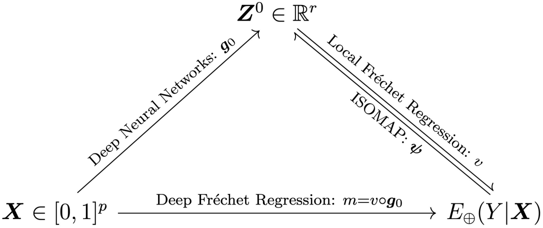


Figure 2: Schematic diagram for the deep Fréchet regression $m = v \circ \boldsymbol{g}_0$.

Manifold Learning

The representation map $\psi : \mathcal{M} \mapsto \mathbb{R}^r$ is unknown and must be estimated from the data $\{Y_i\}_{i=1}^n$, yielding $\hat{\psi}$. Suppose the manifold can be well identified at the sample points through the ISOMAP algorithm.

$$\boldsymbol{Z}_i^0 = (Z_{i1}^0, \ldots, Z_{ir}^0)^{\mathrm{T}} = \psi(Y_i), \quad \boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ir})^{\mathrm{T}} = \hat{\psi}(Y_i).$$

Manifold Learning

The representation map $\psi : \mathcal{M} \mapsto \mathbb{R}^r$ is unknown and must be estimated from the data $\{Y_i\}_{i=1}^n$, yielding $\hat{\psi}$. Suppose the manifold can be well identified at the sample points through the ISOMAP algorithm.

$$\boldsymbol{Z}_i^0 = (Z_{i1}^0, \ldots, Z_{ir}^0)^{\mathrm{T}} = \psi(Y_i), \quad \boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ir})^{\mathrm{T}} = \hat{\psi}(Y_i).$$

For each $j = 1, \ldots, r$, there exists a function $\pi_j : [0,1]^{p \times n} \mapsto \mathbb{R}$ such that

$$Z_{ij} - Z_{ij}^0 = \pi_j(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)(\epsilon_{ij} + u_{nj}), \quad i = 1, \ldots, n,$$

Manifold Learning

The representation map $\psi : \mathcal{M} \mapsto \mathbb{R}^r$ is unknown and must be estimated from the data $\{Y_i\}_{i=1}^n$, yielding $\hat{\psi}$. Suppose the manifold can be well identified at the sample points through the ISOMAP algorithm.

$$\boldsymbol{Z}_i^0 = (Z_{i1}^0, \ldots, Z_{ir}^0)^{\mathrm{T}} = \psi(Y_i), \quad \boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ir})^{\mathrm{T}} = \hat{\psi}(Y_i).$$

For each $j = 1, \ldots, r$, there exists a function $\pi_j : [0,1]^{p \times n} \mapsto \mathbb{R}$ such that

$$Z_{ij} - Z_{ij}^0 = \pi_j(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)(\epsilon_{ij} + u_{nj}), \quad i = 1, \ldots, n,$$

- the bias $u_{nj} \to 0$ as $n \to \infty$,
- $\{\epsilon_{ij}\}_{i=1}^n$ are i.i.d. (mean zero) sub-Gaussian random variables and independent of $\{\boldsymbol{X}_i\}_{i=1}^n$,
- there exists a constant C_π such that

$$\sup_{\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}} |\pi_j(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)| \leq C_\pi, \text{ for all } j.$$

Deep Neural Networks

- Observed predictors: $\boldsymbol{X}_i \in [0,1]^p$
- Estimated responses: $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ir}) = \hat{\boldsymbol{\psi}}(Y_i)$
- True responses: $\boldsymbol{Z}_i^0 = (Z_{i1}^0, \ldots, Z_{ir}^0) = \boldsymbol{\psi}(Y_i) = \boldsymbol{g}_0(X_i)$

Deep Neural Networks

- Observed predictors: $\boldsymbol{X}_i \in [0,1]^p$
- Estimated responses: $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ir}) = \hat{\psi}(Y_i)$
- True responses: $\boldsymbol{Z}_i^0 = (Z_{i1}^0, \ldots, Z_{ir}^0) = \psi(Y_i) = \boldsymbol{g}_0(X_i)$

We model the relationship between $\boldsymbol{Z}_i$ and $\boldsymbol{X}_i$ as

$$Z_{ij} = g_{0j}(\boldsymbol{X}_i) + \pi_j(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)(\epsilon_{ij} + u_{nj}), \quad j = 1, \ldots, r.$$

Deep Neural Networks

- Observed predictors: $\boldsymbol{X}_i \in [0,1]^p$
- Estimated responses: $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ir}) = \hat{\psi}(Y_i)$
- True responses: $\boldsymbol{Z}_i^0 = (Z_{i1}^0, \ldots, Z_{ir}^0) = \psi(Y_i) = \boldsymbol{g}_0(X_i)$

We model the relationship between $\boldsymbol{Z}_i$ and $\boldsymbol{X}_i$ as

$$Z_{ij} = g_{0j}(\boldsymbol{X}_i) + \pi_j(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)(\epsilon_{ij} + u_{nj}), \quad j = 1, \ldots, r.$$

The estimation of g_0 is performed by minimizing the empirical risk

$$\hat{\boldsymbol{g}} = \arg\min_{\boldsymbol{g} \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{Z}_i - \boldsymbol{g}(\boldsymbol{X}_i)\|^2,$$

where $\mathcal{G}$ is a class of neural networks.

Local Fréchet Regression

- Estimated predictors: $\hat{\boldsymbol{Z}}_i = \hat{\boldsymbol{g}}(\boldsymbol{X}_i)$,
- True unobservable predictors: $\boldsymbol{Z}_i^0 = \boldsymbol{g}_0(\boldsymbol{X}_i) = \psi(Y_i)$,
- Response: Y_i.

Local Fréchet Regression

- Estimated predictors: $\hat{\boldsymbol{Z}}_i = \hat{\boldsymbol{g}}(\boldsymbol{X}_i)$,
- True unobservable predictors: $\boldsymbol{Z}_i^0 = \boldsymbol{g}_0(\boldsymbol{X}_i) = \psi(Y_i)$,
- Response: Y_i.

This leads to an errors-in-variables version of local Fréchet regression,

$$\hat{v}_h(\boldsymbol{z}) = \underset{y \in \Omega}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \hat{w}(\hat{\boldsymbol{Z}}_i, \boldsymbol{z}, h) d^2(Y_i, y),$$

where

- $\hat{w}(\hat{\boldsymbol{Z}}_i, \boldsymbol{z}, h) = \frac{1}{\hat{\mu}_0 - \hat{\mu}_1^{\mathrm{T}} \hat{\mu}_2^{-1} \hat{\mu}_1} K_h(\hat{\boldsymbol{Z}}_i - \boldsymbol{z}) \{1 - \hat{\mu}_1^{\mathrm{T}} \hat{\mu}_2^{-1} (\hat{\boldsymbol{Z}}_i - \boldsymbol{z})\}$,
- $\hat{\mu}_k = n^{-1} \sum_{i=1}^{n} K_h(\hat{\boldsymbol{Z}}_i - \boldsymbol{z})(\hat{\boldsymbol{Z}}_i - \boldsymbol{z})^{\oplus k}$ for $k = 0, 1, 2$,
- h is the bandwidth.

Rates of Convergence

Theorem

Under certain regularity conditions, we have

$$d\{\hat{m}(\boldsymbol{X}), m(\boldsymbol{X})\} = O_p[h^{2/(\gamma_1-1)} + (nh^r)^{-1/\{2(\gamma_2-1)\}} + \{h^{-r-1}(\kappa_n \log^{3/2} n + u_n)\}^{1/(\gamma_3-1)}]$$

where

- *Deep Fréchet Regression:* $\hat{m} = \hat{v} \circ \hat{g}$,
- $\boldsymbol{X}$ *is a new predictor independent of sample* $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$,

Rates of Convergence

Theorem

Under certain regularity conditions, we have

$$d\{\hat{m}(\boldsymbol{X}), m(\boldsymbol{X})\} = O_p[h^{2/(\gamma_1-1)} + (nh^r)^{-1/\{2(\gamma_2-1)\}} + \{h^{-r-1}(\kappa_n \log^{3/2} n + u_n)\}^{1/(\gamma_3-1)}]$$

where

- *Deep Fréchet Regression:* $\hat{m} = \hat{v} \circ \hat{g}$,
- $\boldsymbol{X}$ *is a new predictor independent of sample* $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$,
- κ_n *reflects smoothness and intrinsic dimension of true function* g_0,
- $u_n^2 = \max_{j=1,\dots,r} |u_{nj}|$ *is the the vanishing bias produced by ISOMAP,*
- γ_j, $j = 1, 2, 3$, *relates with the curvature of the metric space.*
- $\gamma_1 = \gamma_2 = \gamma_3 = 2$ *for the Wasserstein space, the space of networks, and the space of covariance/correlation matrices.*

Simulation Studies

- We consider sample sizes $n = 100, 200, 500, 1000$, with $Q = 500$ Monte Carlo runs.
- Predictor $\boldsymbol{X} \in \mathbb{R}^9$.
- Metric space-valued responses Y: distributions and networks.
- The intrinsic dimension of the manifold $\mathcal{M}$: $r = 2$.

Distributions

Table 1: Average mean squared prediction error of deep Fréchet regression (DFR), global Fréchet regression (GFR) (Petersen & Müller, 2019) and sufficient dimension reduction (SDR) (Zhang et al., 2024) for distributional responses.

n	DFR	GFR	SDR
100	34.000	52.573	43.598
200	20.976	48.140	26.128
500	12.544	45.712	18.742
1000	7.874	46.418	15.719

Networks

Table 2: Average mean squared prediction error of deep Fréchet regression (DFR), global Fréchet regression (GFR) (Zhou & Müller, 2022) and sufficient dimension reduction (SDR) (Zhang et al., 2024) for network responses.

n	DFR	GFR	SDR
100	88.025	97.994	94.843
200	52.404	91.872	73.035
500	21.486	88.940	59.609
1000	11.701	87.416	56.257

Comparison of Different Manifold Learning Methods

Table 3: Median mean squared prediction error of deep Fréchet regression using various manifold learning techniques, including ISOMAP (Tenenbaum et al., 2000), t-SNE (Van der Maaten & Hinton, 2008), UMAP (McInnes & Healy, 2018), Laplacian eigenmaps (LE) (Belkin & Niyogi, 2003) and diffusion maps (DM) (Coifman & Lafon, 2006).

n	ISOMAP	tSNE	UMAP	LE	DM
100	26.792	30.773	38.193	36.015	29.772
200	16.182	17.010	21.715	21.960	18.063
500	8.363	9.790	11.018	12.635	9.931
1000	4.674	6.228	5.998	8.707	6.298

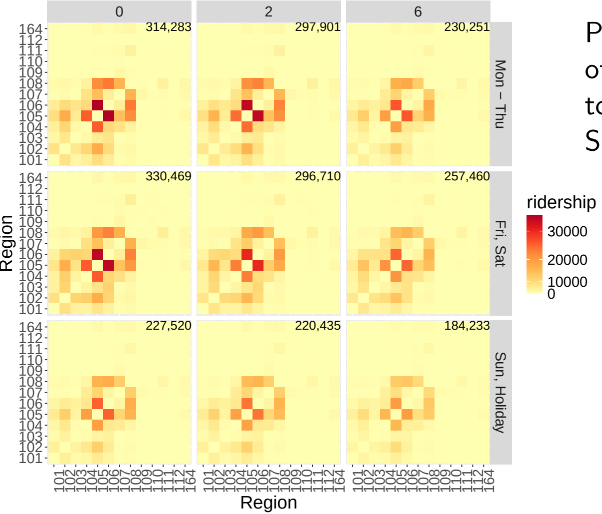
New York Yellow Taxi Data

- **Responses**: daily traffic networks from Jan 1, 2017 to Dec 31, 2019 in Manhattan.
- **15 Predictors**: daily weather information, indicators for days of the week or holiday, and daily trip features averaged over each day.

Table 4: Average mean squared prediction error of deep Fréchet regression (DFR), global Fréchet regression (GFR) (Zhou & Müller, 2022) and sufficient dimension reduction (SDR) (Zhang et al., 2024) for New York yellow taxi data.

DFR	GFR	SDR
0.0074	0.0136	0.0166

New York Yellow Taxi Data



Predicted networks at different levels of precipitation (inches) on Monday to Thursday, Friday to Saturday, and Sunday or holiday.

- 105, 106, 107, and 108: residential areas with popular destinations such as Penn Station and Grand Central Terminal.
- 102, 103, and 104: Popular bars, restaurants, chain stores, and high-end art galleries and museums.

Human Mortality Data

- **Responses**: Age-at-death distributions of 162 countries in 2015.
- **9 Predictors**: demographic (population density, sex ratio, mean childbearing age), economic (GDP, GVA, CPI, unemployment rate, health expenditure), and environmental (arable land) factors in 2015.

Table 5: Average mean squared prediction error of deep Fréchet regression (DFR), global Fréchet regression (GFR) (Petersen & Müller, 2019) and sufficient dimension reduction (SDR) (Zhang et al., 2024) for human mortality data.

DFR	GFR	SDR
26.377	31.322	27.602

Human Mortality Data

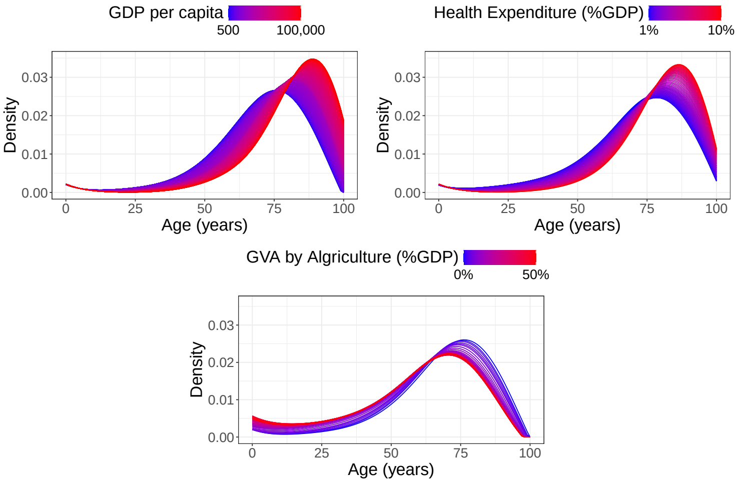


Figure 3: Age-at-death densities at different levels of GDP, health expenditure and gross value added (GVA) by agriculture (%GDP).

Contribution

- Using deep learning for Fréchet regression, highlighting its efficacy as a powerful tool for the regression of metric-space-valued responses.
- Allowing high-dimensional predictors for metric-space-valued responses residing in a low-dimensional manifold.

Contribution

- Using deep learning for Fréchet regression, highlighting its efficacy as a powerful tool for the regression of metric-space-valued responses.
- Allowing high-dimensional predictors for metric-space-valued responses residing in a low-dimensional manifold.
- Investigating the convergence rate of deep neural networks in the presence of dependent sub-Gaussian noise accompanied by bias.

Contribution

- Using deep learning for Fréchet regression, highlighting its efficacy as a powerful tool for the regression of metric-space-valued responses.
- Allowing high-dimensional predictors for metric-space-valued responses residing in a low-dimensional manifold.
- Investigating the convergence rate of deep neural networks in the presence of dependent sub-Gaussian noise accompanied by bias.
- Establishing convergence rate in the presence of errors in predictors for local Fréchet regression,
- Demonstrating how local Fréchet regression can be used to map the low-dimensional representation back to the original metric space.

References I

- Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*, 1373–1396.
- Bhattacharjee, S., & Müller, H.-G. (2023). Single index Fréchet regression. *Annals of Statistics*, *51*(4), 1770–1798.
- Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, *21*(1), 5–30.
- Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincaré*, *10*(4), 215–310.
- Ghosal, A., Meiring, W., & Petersen, A. (2023). Fréchet single index models for object response regression. *Electronic Journal of Statistics*, *17*(1), 1074–1112.

References II

- McInnes, L., & Healy, J. (2018).UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Petersen, A., & Müller, H.-G. (2019).Fréchet regression for random objects with Euclidean predictors. *Annals of Statistics*, *47*(2), 691–719.
- Song, D., & Han, K. (2023).Errors-in-variables Fréchet regression with low-rank covariate approximation. *Advances in Neural Information Processing Systems*, *36*, 80575–80607.
- Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000).A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319–2323.
- Van der Maaten, L., & Hinton, G. (2008).Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.
- Ying, C., & Yu, Z. (2022).Fréchet sufficient dimension reduction for random objects. *Biometrika*, *109*(4), 975–992.

References III

Zhang, Q., Xue, L., & Li, B. (2024). Dimension reduction for Fréchet regression. *Journal of the American Statistical Association, 119*(548), 2733–2747.

Zhou, Y., & Müller, H.-G. (2022). Network regression with graph Laplacians. *Journal of Machine Learning Research, 23*, 1–41.

Questions?