,

Research Statement

Satarupa Bhattacharjee
Department of Statistics, Pennsylvania State University
Email: sfb5992@psu.edu

My research centers around **statistical analysis of functional and non-Euclidean data**. Data taking values in metric spaces, which we refer to as *random objects*, are increasingly common in real-world applications, such as graph Laplacians, covariance matrices, probability distributions, and compositional vectors with examples in various domains like brain imaging, social networks, income histograms, microbiome data, and genetics. For instance, I investigate the functional connectivity correlation network of fMRI signals, which resides in a non-linear space without inherent direction but can be treated as random objects in a metric space with an appropriate metric (see e.g., Figure 1). c On the other hand, classical

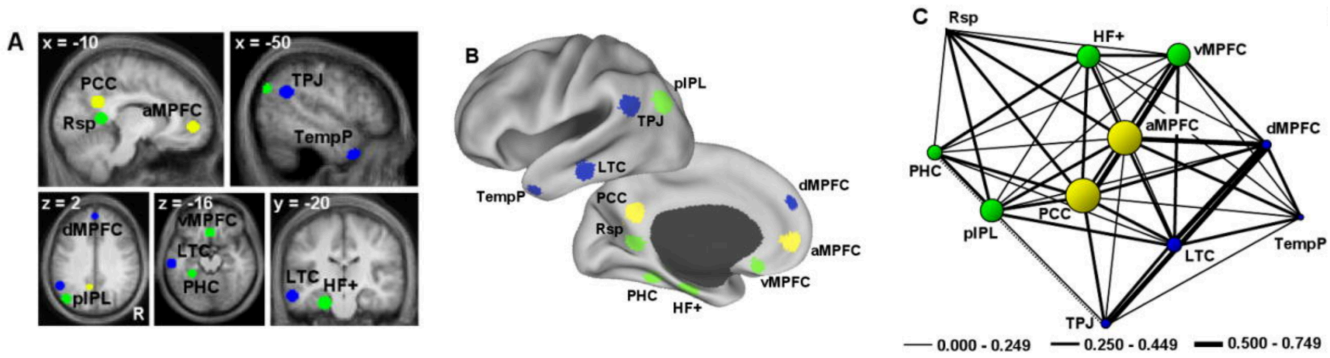

Figure 1: Image taken from Andrews-Hanna et al. (2010). Construction of a functional connectivity correlation network between 11 regions of interest in the brain. A. Eleven a priori regions within the default network are shown overlain on transverse slices colored according to the subsystems revealed in C. B. Regions are also projected onto a surface template. C. Functional correlation strengths between the 11 regions were extracted. The thickness of the lines reflects the strength of the correlation between regions.

functional data analysis typically deals with real-valued functions in Hilbert spaces, but in recent years, it has expanded to analyze more intricate time-indexed data, including time-varying random objects. So far, I have worked on the following main areas:

1. **Statistics on non-Euclidean object valued data-** developing versatile statistical methods for data in abstract metric spaces, with applications in brain imaging, mortality distributions, child neurological development, traffic networks, and genetics. A sub-domain involves studying **time-varying metric space data in functional and longitudinal analysis**, with a focus on dynamic networks and distribution objects.

2. **Reproducing kernel Hilbert spaces**- creating mathematical and computational methods for handling infinite-dimensional data, particularly in relation to *metric geometry* and *sufficient dimension reduction.*

3. **Causal inference** in conjunction with *random object and distributional data analysis.*

4. Statistical modeling and analysis of the COVID-19 data using tools from **functional data analysis and nonparametrics**.

In the following, I will describe the projects I have completed, some ongoing research, and my future research plans.

1 Random object data analysis

Single index Fréchet regression

In our paper Bhattacharjee and Müller (2023b), which is accepted in the *Annals of Statistics*, we define the Single Index Fréchet Regression (IFR) model for a response variable Y in the metric space (Ω, d). This model involves projecting a general multivariate Euclidean predictor $\mathbf{X}$ onto a specific direction vector as

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}:\boldsymbol{\theta}^\top\boldsymbol{\theta}=1}{\operatorname{argmin}}\ \mathbb{E}[d^2(Y, m_\oplus(\mathbf{X}^\top\boldsymbol{\theta}, \boldsymbol{\theta}_0))], \text{ where } m_\oplus(t, \theta_0) = \underset{\omega\in(\Omega,d)}{\operatorname{argmin}}\ \mathbb{E}(d^2(Y, \omega)|\mathbf{X}^\top\boldsymbol{\theta}_0 = t). \tag{1}$$

Here $\mathbb{E}_\oplus$ denotes the conditional Fréchet mean of Y given $\mathbf{X}$ as a generalization of the conditional mean $\mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ to metric spaces (Petersen and Müller, 2019). We assume the conditional Fréchet mean depends on θ_0 such that the distribution of Y depends on $\mathbf{X}$ only through the index $\mathbf{X}^\top\boldsymbol{\theta}_0$ thus linking it to sufficient dimension reduction (SDR) literature. While Fréchet regression Petersen and Müller (2019) is effective for modeling the conditional mean of random objects given multivariate Euclidean vectors, it does not provide for regression parameters such as slopes or intercepts, since the metric space-valued responses are not amenable to linear operations. In our work, we offer an inferential approach, using $\boldsymbol{\theta}_0$ to substitute for the inherent absence of parameters in Fréchet regression. We derive the asymptotic distribution of suitable estimates of these parameters with the asymptotic covariance matrix estimated by Bootstrap, which allows us to test linear hypotheses for the parameters, as long as an identifiability condition is met. Figure 2 illustrates the scope and interpretability of the proposed framework in the context of analyzing the *Age-at-Death Distributions* constructed from the Human Mortality Data.

This work received the *Best Student Paper Award* of the Nonparametric Statistics Section of the American Statistical Association (ASA) in 2022.

Geodesic mixed-effects models for repeatedly observed/longitudinal random objects and time-concurrent regression

In Bhattacharjee and Müller (2023a), currently under review at the *Journal of American Statistical Association (JASA)*, we introduce mixed-effects modeling for handling repeated measurements of random object data in geodesic spaces, which lack global or local linear structures. Unlike traditional mixed-effects models in Euclidean spaces, additive errors or specific distributional characteristics are unattainable for metric space-valued data. Instead, we assume the mean response trajectories are geodesics in the metric space, and deviations from the model are quantified by perturbation maps or transports. A key finding is that the geodesic trajectories assumption for random objects is a natural extension of the linearity assumption in standard Euclidean scenarios within general geodesic metric spaces. These geodesics can be recovered from noisy observations by exploiting a connection between the geodesic path and the path obtained by global Fréchet regression for random objects. Baseline Euclidean covariates' impact

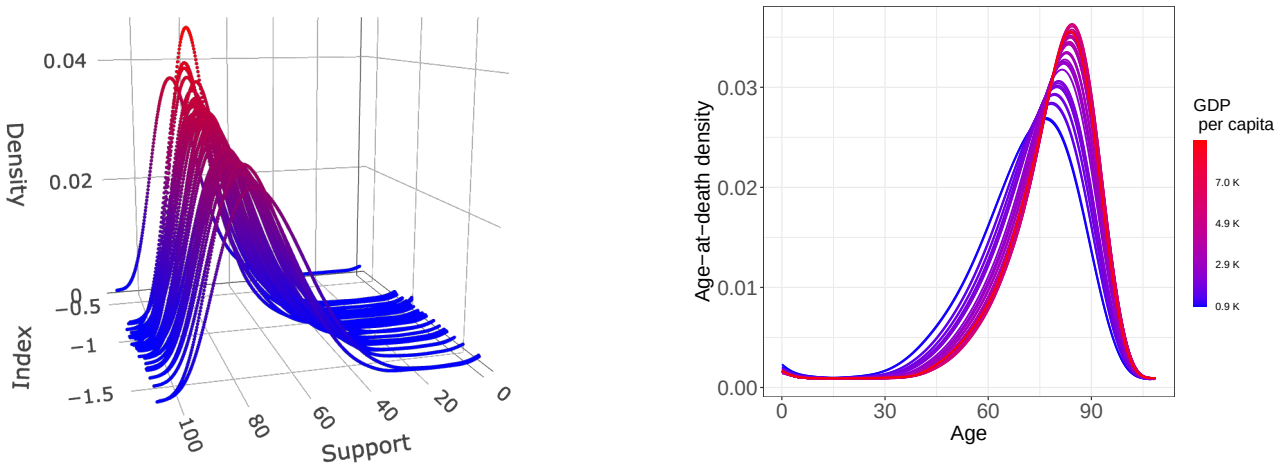

Figure 2: The left panel shows the data visualization for age-at-death densities for 40 countries at the calendar year 2010. The middle panel illustrates the change in density by changing the value of a significant predictor:"GDP per capita" from low (blue) to high (red) when the other predictors are fixed at their mean level.

on geodesic paths is modeled through another object regression step. Our work includes an analysis of the asymptotic convergence of our estimates and offers illustrations using simulations and real-data applications with resting-state functional Magnetic Resonance Imaging (fMRI) data from the Alzheimer's Disease Neuroimaging (ADNI) study.

In our previous work Bhattacharjee and Müller (2022) published in the Electronic Journal of Statistics, we introduced a time-varying regression framework for paired stochastic processes involving real covariates and object responses over time. We extended the concept of conditional Fréchet means to accommodate this concurrent-time framework.

Nonlinear global Fréchet regression for random objects via weak conditional expectation

In Bhattacharjee et al. (2023), we introduce a nonlinear global regression model for object-valued predictor and response pairs, emphasizing distribution-on-distribution regression. We propose the concept of a weak conditional Fréchet mean, which is a generalization of the conditional Fréchet mean, and establish a connection between them based on Carleman operators and their inducing functions on the space's metric. Our model includes the state-of-the-art globally linear Fréchet regression as a special case. To apply our model, the predictor's metric space must have a rich enough reproducing kernel Hilbert space embedding to characterize the joint probability distribution, while the intrinsic geometry of the metric space where the responses lie, is used for studying the proposed estimate's asymptotic convergence. We are preparing to submit this work to the *Annals of Statistics*.

Causal inference for distributional data with continuous treatments

The motivation for this work stems from ongoing research on large-scale data analysis, like Medicare data. A key objective is to assess the potential causal relationship between exposure to air pollution and adverse health outcomes, going beyond traditional regression methods. Modern science aims to

understand causality in treatments and outcomes. In many applications, the interest is in the causal effects on distribution functions, such as shapes, curves, and images, which offer richer information than single summary measures like means or quantiles. Leveraging Wasserstein geometry, which is tied to optimal transport, enhances interpretability and statistical performance in such applications. In the intersection of distributional data analysis and causal inference, especially when dealing with non-Euclidean data in the potential outcome framework, the goal is to develop doubly debiased estimates for distribution-valued outcomes and continuous treatments, even in the presence of continuous confounding factors.

This is an ongoing project, which we are preparing to submit to *Biometrics* soon.

2 Application of functional data analysis and nonparametric methods

Time-dynamics of the COVID-19 pandemic: Inference and mitigation strategy

In our paper Bhattacharjee et al. (2022a) published in *Nature- Scientific Reports*, the evolution of the COVID-19 pandemic is described through a *time-dependent stochastic dynamic model* with multiple compartments through a system of difference equations. In contrast with conventional epidemiological models, the proposed model involves interpretable time-varying rate parameters and latent unobservable compartments such as the number of asymptomatic but infected individuals over time, $\hat{A}_t$ (see Figure 3). The model fitting strategy is built upon *nonparametric smoothing and profiling ideas*, with confidence bands for the parameters obtained through a residual bootstrap procedure.

As a subsequent work, our paper Bhattacharjee et al. (2022b) which features as a chapter in the book *Managing Complexity and COVID-19, Taylor & Francis (Routledge, UK)*, we propose a comprehensive network model to determine an optimal intervention strategy from a policy perspective.

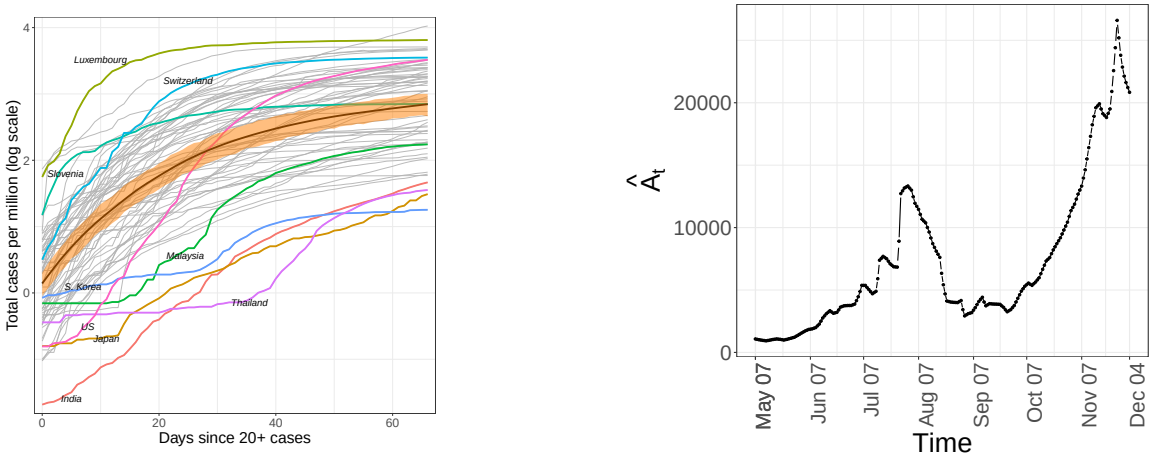

Figure 3: The left panel shows the trajectories of total case count per million individuals on the log scale, with the smoothed mean curves marked by bold black lines and the orange ribbons representing pointwise 95% bootstrap confidence bands for the overall mean functions (Carroll et al., 2020). The right panel shows the temporal pattern for the estimated asymptomatic individuals in Utah for a given period of time (Bhattacharjee et al., 2022a).

Functional data analysis on the time-dynamics of COVID-19

In our contribution Carroll et al. (2020), published in *Scientific Reports -Nature*, we apply tools from functional data analysis to model and forecast the trajectories of COVID-19 cases and deaths across countries

longitudinally. Our framework quantifies the effects of demographic covariates and social mobility indices on doubling rates and case fatality rates through a time-varying regression model.

In our related paper Dubey et al. (2022), published in *Journal of Mathematical Analysis and Applications*, we use a functional regression model with a history index from a sample of random trajectories obeying an unknown random differential equation model with delay.

3 Ongoing and Future work

The growing prevalence of non-Euclidean data in diverse scientific fields demands fresh statistical approaches. With expertise in dimension reduction and object data analysis, I'm well-equipped for this challenge. I aim to maintain current collaborations and explore new ones, focusing on projects aligning with my areas of specialization.

Causal inference for random object data

I'm interested in studying causal relationships with random objects, particularly using the potential outcome framework in the context of modeling and covariate balancing in observational studies. Additionally, tests for homogeneity and independence using kernel mean embeddings of complex object data can be used to examine causal counterfactual effects. These methods could be applied, for instance, to understand the causal link between brain connectivity evolution and cognitive behavior in individuals with neuro-atypical brains.

Index models for contextual bandit problems

I am interested in exploring connections between object data analysis and theoretical machine learning. Currently, I am working on a nonparametric approach to contextual bandit problems with finite arms, utilizing index Fréchet regression models for inference via a kernelized version of the $\epsilon-$greedy strategy.

Geodesic set regression

In Hilbert space, dimension reduction involves projecting data to a lower-dimensional subspace. In the Wasserstein space of univariate distributions $\mathcal{W}_2$, where inner product-based projection isn't natural, we leverage its pseudo-Riemannian structure to define geodesic sets, allowing us to express a geodesic set reduction problem as a generalization of single-index and multi-index models. We employ a forward regression approach to maximize the explained total variation in responses by projecting onto the geodesic set.

Residual analysis for object regression

I'm also interested in creating visualization tools and diagnostic plots for object regression methods, particularly for outlier detection and assessing model fit, which are crucial aspects of model validation within a regression context. This problem presents both interest and challenges in its own right.

Some other potential future research interests of mine include developing dimensionality reduction techniques like PCA for metric space-valued data, studying methods for modeling sparsely observed

longitudinal metric space data (e.g., distributions and networks), and exploring supervised classification for intra-hub connectivity distributions in brains using the Wasserstein metric.

References

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., and Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron*, 65(4):550–562.

Bhattacharjee, S., Li, B., and Xue, L. (2023). Nonlinear global Fréchet regression for random objects via weak conditional expectation. *arXiv preprint*.

Bhattacharjee, S., Liao, S., Paul, D., and Chaudhuri, S. (2022a). Inference on the dynamics of covid-19 in the united states. *Nature- Scientific Reports*, 12(1):2253.

Bhattacharjee, S., Liao, S., Paul, D., and Chaudhuri, S. (2022b). Taming the pandemic by doing the mundane. In *Managing Complexity and COVID-19*, pages 62–82. Routledge.

Bhattacharjee, S. and Müller, H.-G. (2022). Concurrent object regression. *Electronic Journal of Statistics*, 16(2):4031–4089.

Bhattacharjee, S. and Müller, H.-G. (2023a). Geodesic mixed effects models for repeatedly observed/longitudinal random objects. *arXiv preprint arXiv:2307.05726*.

Bhattacharjee, S. and Müller, H.-G. (2023b). Single index Fréchet regression. *arXiv preprint arXiv:2108.05437*.

Carroll, C., Bhattacharjee, S., Chen, Y., Dubey, P., Fan, J., Gajardo, Á., Zhou, X., Müller, H.-G., and Wang, J.-L. (2020). Time dynamics of covid-19. *Nature- Scientific Reports*, 10(1):21040.

Dubey, P., Chen, Y., Gajardo, Á., Bhattacharjee, S., Carroll, C., Zhou, Y., Chen, H., and Müller, H.-G. (2022). Learning delay dynamics for multivariate stochastic processes, with application to the prediction of the growth rate of covid-19 cases in the united states. *Journal of Mathematical Analysis and Applications*, 514(2):125677.

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