,

Research Statement

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

My primary research centers around analyzing functional and non-Euclidean data situated in general metric spaces. Data taking values in metric spaces, which we refer to as *random objects*, are becoming increasingly common in many real-life applications. Often such data objects appear as samples of graph Laplacians, covariance matrices, and probability density functions among others, with examples in brain imaging data, traffic networks, distribution valued data, and high-dimensional genetics data. A motivating data example is the functional connectivity correlation network of fMRI signals represented as correlation matrices between the different nodes of the brain, which reside in a space that is not linear and there is no concept of direction. However, the connectivity correlation matrices can be perceived as random objects in a metric space, endowed with a suitable metric (see e.g., Figure 1).

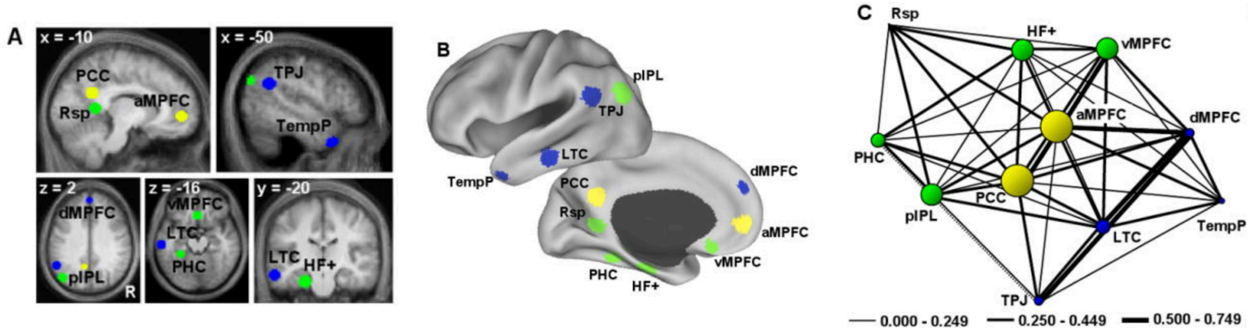

Figure 1: Image taken from Andrews-Hanna et al. (2010). Construction of a functional connectivity correlation network between 11 regions of interest in the brain. A. Eleven a priori regions within the default network are shown overlain on transverse slices colored according to the subsystems revealed 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.

Complexity in analyzing random objects, including the examples mentioned above, arises from the fact that they do not take values in Euclidean space; not even locally, and do not even conform to usual vector space structures or operations. Thus many common statistical notions and constructs devised for Euclidean and functional data- which are usually considered as random elements taking values in a Hilbert space, cannot be applicable to such object-valued data due to their nonlinear nature. On the other hand, classical functional data analysis deals with data that consist of samples of real-valued functions in a Hilbert space as infinite-dimensional elements. In the recent decade, this area has been moving towards analyzing more complex time-indexed data, including time-varying random objects.

My Ph.D. work was focused on developing broadly applicable regression methods and inference techniques to analyze object data and utilizing novel methodologies to solve scientific problems in the biological and social sciences. Metric spaces that are rich enough to admit a kernel embedding are also of special interest to construct an underlying *reproducing kernel Hilbert spaces (RKHS)*. In my postdoctoral research, I have been exploring this connection to develop a nonlinear global version of Fréchet regression for random objects via the notion of weak conditional expectation.

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

1 Random object data analysis

1.1 Single index Fréchet regression

In our paper Bhattacharjee and Müller (2023b), which will appear in the *Annals of Statistics*, we define the single index Fréchet regression (IFR) model for a response Y taking values in the metric space (Ω, d) by projecting a general multivariate Euclidean predictor $\mathbf{X}$ onto a desired direction vector as

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

where $\mathbb{E}_{\oplus}$ denotes the conditional Fréchet mean of Y given $\mathbf{X}$ as a generalization of $\mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ to metric spaces (Petersen and Müller, 2019). The conditional Fréchet mean is assumed to be a function of θ_0 in such a way that the distribution of Y only depends on $\mathbf{X}$ through the index $\mathbf{X}^\top \boldsymbol{\theta}_0$ thus tying into the sufficient dimension reduction (SDR) literature. While Fréchet regression Petersen and Müller (2019) has proved useful for modeling the conditional mean of such 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. As a consequence, distributional results for Fréchet regression have been elusive. In our work, we provide an inferential paradigm, where $\boldsymbol{\theta}_0$ can be used to substitute for the inherent absence of parameters in Fréchet regression. Specifically, the asymptotic distribution of suitable estimates of these parameters is derived with the asymptotic covariance matrix estimated by Bootstrap. This then can be utilized to test linear hypotheses for the parameters, subject to an identifiability condition. 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.

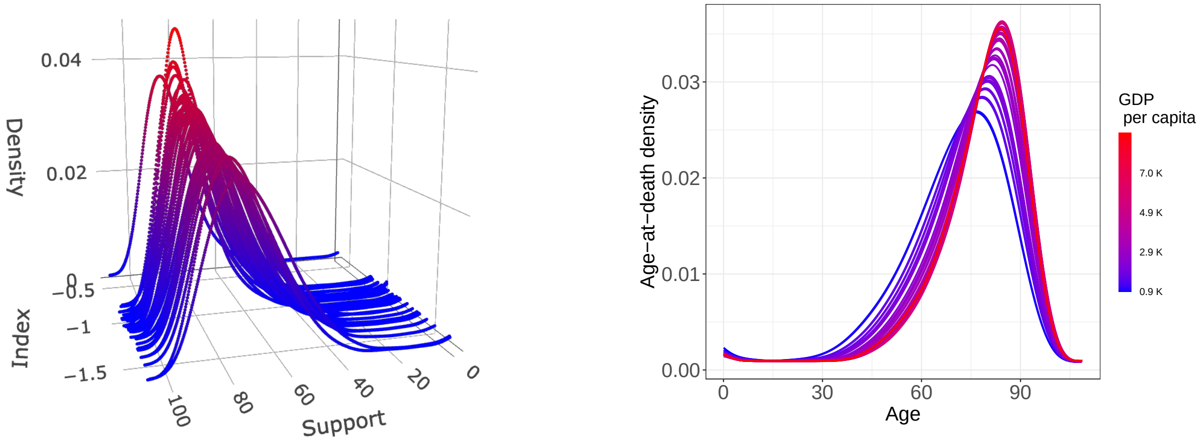

Figure 2: The left panel shows the data visualization for age-at-death densities for 40 countries at the calendar year 2010. The middle panel illustrates the change in density by changing the value of a significant predictor:"GDP per capita" from low (blue) to high (red) when the other predictors are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

1.2 Geodesic mixed-effects models for repeatedly observed/longitudinal random objects

In Bhattacharjee and Müller (2023a), which is currently under review at the *Journal of American Statistical Association (JASA)*, we introduce mixed effect modeling for repeated measurement of random object data. In such settings, the classical additive error model and distributional assumptions are unattainable, which are typical in the Euclidean versions of mixed-effects modeling. We extend mixed-effects regression for data in geodesic spaces without either global linear or local linear (Riemannian) structure, where the underlying mean response trajectories are geodesics in the metric space and the deviations of the observations from the model are quantified by perturbation maps or transports. A key finding is that the geodesic trajectories assumption for the case of random objects is a natural extension of the linearity assumption in the standard Euclidean scenario to the case of general geodesic metric spaces. 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. The effect of baseline Euclidean covariates on the geodesic paths is modeled by another object regression step. We study the asymptotic convergence of the proposed estimates and provide illustrations through simulations and real-data applications for resting-state functional Magnetic Resonance Imaging (fMRI) data from the *Alzheimer's Disease Neuroimaging (ADNI) study.*

In our related work Bhattacharjee and Müller (2022), published in the *Electronic Journal of Statistics*, we developed a time-varying regression framework for paired stochastic processes of real covariates and object responses as a function of time by extending the notion of conditional Fréchet means to a concurrent-time framework.

1.3 Nonlinear global Fréchet regression for random objects via weak conditional expectation

We introduce a nonlinear global regression model for object-valued predictor and response tuples. We propose the notion of a weak conditional Fréchet mean to aid the object-on-object regression framework with a special emphasis on distribution-on-distribution regression. One of the main contributions is to establish a connection between the conditional Fréchet mean and the weak conditional Fréchet mean, the latter being a generalization of the former. The motivation is based on Carleman operators and their inducing functions on the inherent metric of the space. The state-of-the-art globally linear Fréchet regression approach by Petersen and Müller (2019) emerges as a special case of the proposed model. We require that the metric space where the predictors reside admits a reproducing kernel Hilbert space embedding that is rich enough to characterize the joint probability distribution of the responses and the predictors, while the intrinsic geometry of the metric space where the responses lie is utilized to study the asymptotic convergence of the proposed estimates. We are preparing to submit this work to the *Annals of Statistics*.

1.4 Causal inference for distributional data with continuous treatments

The motivation for this work is provided by an ongoing collaboration on large-scale data analysis such as Medicare data, where one potential goal is to assess whether and in what magnitude exposure to air pollution is causally linked to adverse health outcomes. Understanding causal relationships between treatments and outcomes beyond a regression paradigm is one of the most important aims of modern science. On the other hand, in many modern applications, the interest lies in the causal effect on the distributions themselves, represented by distribution functions, such as shapes, curves, and images, which contain richer information than the single summary measure such as the mean or the quantiles. Due to its inherent connection with optimal transport, the Wasserstein geometry of the space of distributions enhances interpretability and statistical performance in applications. In the conjunction of distributional data analysis and causal inference, specifically, moving beyond Euclidean data in the potential outcome framework, we aim to develop double debiased estimates for distribution-valued outcomes and continuous treatments, in the presence of possibly continuous confounders.

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

2 Statistical modeling and analysis of the COVID-19 data using tools from functional data analysis and nonparametrics

The following projects describe my collaborative work in the Statistical modeling and analysis of COVID-19 data using tools from functional data analysis and nonparametrics. Figure ?? illustrates a few applications in this context.

2.1 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 ($\hat{A}_t$). 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.

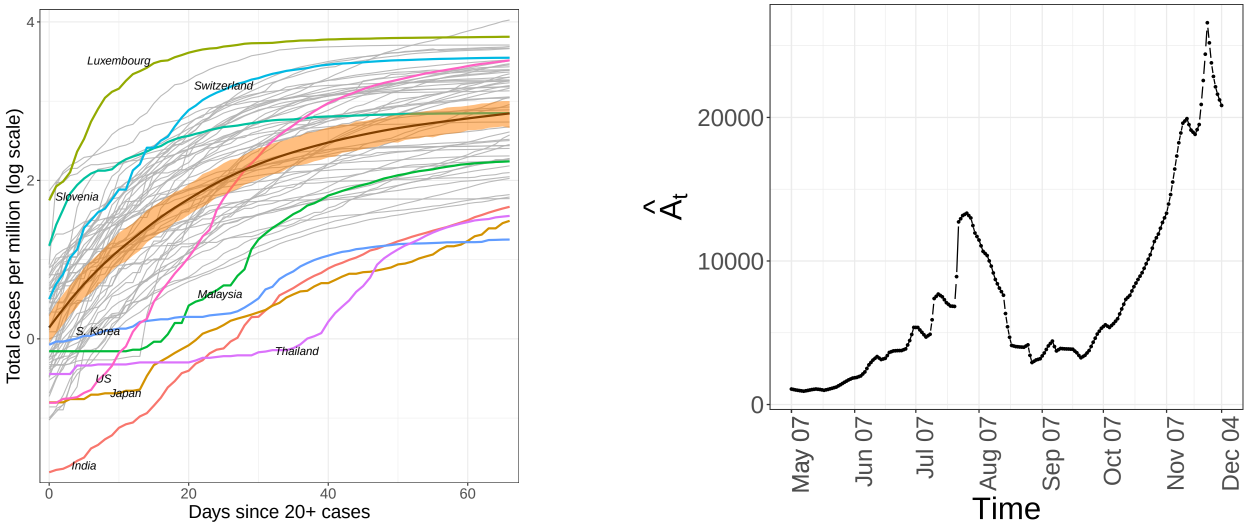

Figure 3: The left panel shows the trajectories of total case count per million individuals on the log scale, with the smoothed mean curves marked by bold black lines and the orange ribbons representing pointwise 95% bootstrap confidence bands for the overall mean functions (Carroll et al., 2020). The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

2.2 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 increasing abundance of non-Euclidean data arising in various scientific areas calls for coherent innovations in statistical methods and theories. My background in dimension reduction and object data analysis makes me suitable for this task. I hope to continue my existing collaborations and initiate new ones, working on projects that overlap with my areas of expertise.

3.1 Causal inference for random object data

Beyond the analysis of regression association that involves metric space-valued data, exploring the causal relationship of random objects is a potential area of interest for me. The potential outcome framework provides an ideal setting in the context of both modeling and covariate balancing approaches to weighting in observational studies. Along similar lines of thought, tests for homogeneity and independence via kernel mean embeddings of complex object data can be used to test for causal counterfactual effects. Such methods can be utilized to study the cause-effect relationship between the evolution of brain connectivity and cognitive behavior for a sample of individuals with neuro-atypical brains, for example.

3.2 Index models for contextual bandit problems

I am also interested in establishing connections between the realms of object data analysis and theoretical machine learning. For example, in nonparametric versions of contextual bandit problems, in a setting with finitely many arms, the index Fréchet regression models can facilitate inference through a kernelized version of the ϵ-greedy strategy.

3.3 Geodesic set regression

When developing dimension reduction methods in Hilbert space, we seek to project the data onto a subspace with a lower dimension. Unlike Hilbert spaces, there is no natural way to define projection via inner product in the Wasserstein space of univariate distributions $\mathcal{W}_2$. However, borrowing the pseudo-Riemannian structure of $\mathcal{W}_2$ (Bigot et al., 2017), we can define the geodesic set as a generalization of convex sets in Euclidean space. As generalizations of single-index and multi-index models, we can express a sufficient geodesic set reduction problem and use a forward regression approach by maximizing the total variation in the responses explained by the geodesic set projection.

3.4 Residual analysis for object regression

Another topic of interest is the development of visualization tools and diagnostic plots for any object regression method. Detection of outliers and identifying the lack of fit for the model are of key importance among other interpretations in a regression context for validating the model. The problem is both interesting and challenging on its own.

Other potential future research ideas that are of interest to me include developing dimensionality reduction tools such as principal component analysis (PCA) for metric space-valued data to generate low-dimensional approximation for intrinsically high- or infinite-dimensional data while preserving as much of the variation in the data as possible, studying methods for modeling functional, especially sparsely observed longitudinal metric space valued data, such as distributions and networks, supervised classification problems for intra-hub connectivity distributions in brains by the Wasserstein metric, and so on.

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., 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*.

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, Probabilités et Statistiques*, 53:1–26.

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.