Research Statement

My research interests broadly fall into these areas:

1. **Statistics on non-Euclidean object valued data-** with applications in brain imaging, mortality distributions, child neurological development, traffic networks, microbiome data, and genetics. A sub-domain involves studying **time-varying metric space data in functional and longitudinal analysis**, e.g., dynamic networks and distribution objects.

2. mathematical and computational methods in **reproducing kernel Hilbert spaces**- 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**.

1 Random object data analysis

Data taking values in metric spaces, often referred to as *random objects*, are increasingly common in real-world applications in the form of graph Laplacians, covariance matrices, probability distributions, and compositional vectors. Analyzing random objects, like the mentioned examples, poses complexity due to their departure from Euclidean space properties and common statistical concepts designed for Euclidean and functional data typically residing in Hilbert spaces.

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, $\boldsymbol{\theta}_0$. We assume 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. In our work, we offer a novel 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.

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. Unlike traditional mixed-effects models, 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. These geodesics can be recovered from noisy observations by making a connection between the geodesic path and the path obtained by global Fréchet regression for random objects. We analyze the asymptotic convergence of proposed estimates and offer illustrations using resting-state 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.

Nonlinear global Fréchet regression for object-data via weak conditional expectation

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, the state-of-the-art globally linear Fréchet regression emerges as a special case. 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. 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, the goal is to develop doubly debiased estimates for distribution-valued outcomes and continuous treatments, incorporating 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

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*. In contrast with conventional epidemiological models, the proposed model involves interpretable time-varying rate parameters and latent unobservable compartments. The model fitting strategy is built upon *nonparametric smoothing and profiling ideas*, with confidence bands for the parameters obtained through residual bootstrap. 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.

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. In a 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.

1. I'm interested in studying causal relationships between 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.
2. The connection between object data analysis and theoretical machine learning is of potential interest. Currently, I am working on a nonparametric approach to contextual bandit problems with finite arms, utilizing index Frćhet regression models for inference via a kernelized version of the $\epsilon-$greedy strategy.
3. 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

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.