Research Statement

Hang Zhou (hgzhou@ucdavis.edu)

From my Ph.D. research in functional data analysis to my current work with object data and machine learning, I have consistently focused on **leveraging the intrinsic structure of data to develop statistical models and address scientific problems**. Advancements in data collection techniques have led to the emergence of complex structured data in various fields, including neuroimaging, health science, energy, transportation and information science. The technical and theoretical challenges presented by these applications have motivated me to explore and formulate generalized statistical methodologies that are compatible with the natural geometry of the data and to develop a unified theoretical framework.

Research Accomplishments

Topic 1: Optimal rate and phase transition in modeling discretely observed functional data

Functional data analysis is a significant field in statistics that treats data as random functions defined over a continuum, often time. In reality, data are collected at discrete time points with measurement errors. Understanding how these discrete observations and measurement errors affect convergence has become a major line of research and has been well studied for mean and covariance functions. After two decades of research, a key problem remains unsolved: the optimal convergence and phase transition for models involving inverse problems, such as functional linear regression and functional generalized regression. In [1], we proposed a new estimation strategy and obtained the optimal convergence rate for the functional linear regression model. In the subsequent work [2], we focused on the perturbation analysis of the covariance operator and derived a unified theory for a diverging number of eigencomponents, addressing an important open problem on which there has been little substantial progress since Peter Hall's 2006 result on the estimation of a fixed number of eigenfunctions. The results and techniques from these two papers fill a critical gap and provide a novel toolkit for the perturbation analysis of covariance operators with noisy, discretely observed functional data. They can be applied to models involving inverse problems, dimension reduction, and data representation based on functional principal component analysis.

Topic 2: Statistical inference for distributional data

Modern data collection technologies have rapidly evolved, leading to the widespread emergence of distribution-valued data and stochastic processes in various complex applications. By leveraging the geometric structure of Wasserstein space, we developed a variety of statistical frameworks compatible with optimal transport. In [3], we studied the canonical correlation for distribution-valued functional data and extended the concepts of tensor Hilbert spaces and tangent fields to Wasserstein space. In [4], we proposed an intrinsic representation model for distribution-valued processes, where the processes may be fully or only partially observed. In [5], we propose a generalized notion of the Wasserstein-Fréchet integral of conditional distributions, leading to a useful and interpretable concept of conditional distribution with the systematic effect removed. Our proposed methods have connected statistical methodologies with scientific applications in neuroimaging, studies of human longevity and diabetes, COVID-19 case analysis, and children early development.

Topic 3: Conformal prediction for object data

Object data residing in general metric spaces paired with Euclidean predictors have gained increasing interest in modern data analysis, and various statistical approaches for analyzing such data have been developed over the past few years. In [6], we introduced conditional distance profiles and developed a conformal inference framework for object-valued responses. Unlike most existing methods that project data onto extrinsic linear spaces, the proposed conditional profile score is an intrinsic approach for projecting the object data onto the Wasserstein space. We also developed a theoretical framework to establish uniform convergence rates for the local linear estimator involving function classes defined on metric spaces.

Topic 4: Model-agnostic algorithms and theoretical foundations in statistical learning

Machine learning methods have recently surged in popularity and are now a significant area in statistical research. My coauthors and I have focused on developing interpretable machine learning algorithms with statistical guarantees. In [7], we introduced veracity scores that distinguish between genuine errors and natural data fluctuations, and proposed model-agnostic approaches that can utilize any type of regression model to identify errors in responses. In [8], we investigated the non-parametric regression problem with repeated measurements using fully connected deep neural networks, and derived a novel theoretical framework that allows for arbitrary sampling frequency.

Topic 5: Modeling dynamical systems and its applications in information retrieval

With the development of modern data analysis techniques, the distinctions between statistics, machine learning, data mining, and data science have become increasingly blurred. Together with my collaborators in computer science and electrical engineering, I have developed algorithms and computational tools designed to solve complex interdisciplinary scientific problems. In [9], we study the problem of semi-supervised 2D and 3D cross-modal retrieval. In [10], we investigate the modeling of multi-agent dynamical systems, and introduce prototype decomposition from contextual knowledge into a continuous graph ODE framework. In [11], we address the problem of effective rigid dynamics modeling and propose a novel approach that effectively estimates collision occurrences and updates the states of both mesh node and object representations during evolution.

Future Research Agenda

In my future research, I aim to develop interpretable statistical models tailored for complex data and to tackle various scientific problems. My work will focus on two main areas and their respective applications:

Goal 1: Statistical inference and applications for object data

Current inference methods for object data often rely on projecting the data onto a linear space, leading to transformations that can be metric-distorting or non-bijective. To address this, I plan to utilize the concept of the distance profile to establish a unified intrinsic statistical inference framework for object data, encompassing independent testing, outlier detection and classification.

Goal 2: Theoretical framework for functional nonparametric regression with deep neural networks

While the learnability of deep neural networks is well-studied, their potential as a dimension reduction tool in functional non-parametric regression remains unclear. I am investigating how to achieve a polynomial rate of convergence for this problem by assuming a specific intrinsic dimension of the regression function and considering the smoothness of the functional inputs.

Throughout my academic career, I have collaborated extensively with researchers from various backgrounds, including both junior and senior colleagues at institutions such as Rutgers University, UCLA, NUS, Carnegie Mellon University, and industry partners like Clean Lab. These collaborations have demonstrated the value of interdisciplinary communication in solving complex problems. I look forward to expanding my collaborative network at your university.

References

[1] **Hang Zhou**, Fang Yao, and Huiming Zhang. Functional linear regression for discretely observed data: from ideal to reality. *Biometrika*, 110(2):381–393, 2023.

[2] **Hang Zhou**, Dongyi Wei, and Fang Yao. Theory of functional principal component analysis for noisy and discretely observed data. *In revision at Annals of Statistics*, 2024.

[3] **Hang Zhou**, Zhenhua Lin, and Fang Yao. Intrinsic wasserstein correlation analysis. *Statistica Sinica*, in press, 2024+.

[4] **Hang Zhou** and Hans-Georg Müller. Optimal transport representations and functional principal components for distribution-valued processes, 2023.

[5] Álvaro Gajardo, Hans-Georg Müller, and **Hang Zhou**. Wasserstein-fréchet integral of conditional distributions. *In minor revision at Electronic Journal of Statistics,*, 2024+.

[6] **Hang Zhou** and Hans-Georg Müller. Conformal inference for random objects. *In major revision at Annals of Statistics*, 2024+.

[7] **Hang Zhou**, Jonas Mueller, Mayank Kumar, Jane-Ling Wang, and Jing Lei. Detecting errors in a numerical response via any regression model. *Journal of Data-centric Machine Learning Research*, 2024.

[8] Shunxing yan, **Hang Zhou**, and Fang Yao. Deep regression for repeated measurements. *In major revision at Journal of American Statistical Association,*, 2024+.

[9] Fan Zhang, **Hang Zhou**, Xian-Sheng Hua, Chong Chen, and Xiao Luo. Hope: A hierarchical perspective for semi-supervised 2d-3d cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[10] Xiao Luo, Yiyang Gu, Huiyu Jiang, **Hang Zhou**, Jinsheng Huang, Wei Ju, Zhiping Xiao, Ming Zhang, and Yizhou Sun. Pgode: Towards high-quality system dynamics modeling. In *The Forty-first International Conference on Machine Learning*.

[11] Jingyang Yuan, Gongbo Sun, Zhiping Xiao, **Hang Zhou**, Xiao Luo, Junyu Luo, Yusheng Zhao, Wei Ju, and Ming Zhang. EGODE: An event-attended graph ODE framework for modeling rigid dynamics. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.