Research Statement

Fangyi Wang
Department of Statistics, The Ohio State University

Modern scientific research increasingly demands **reliable and efficient uncertainty quantification** alongside accurate predictions, as large-scale, high-dimensional data with complex structures and black-box machine learning models become ubiquitous. A primary example is *functional data*, where observations typically depict changes of measurements over time and appear in a variety of fields: physiological signals in biomedical studies, environmental factors like temperature, daily traffic flow in civil engineering, etc. While conventional uncertainty quantification techniques, such as closed-form variance estimates and bootstrap methods, may suffice for simple parametric models, they become theoretically or computationally prohibitive when dealing with inherent challenges in functional data: infinite-dimensionality, complex dependencies, and irregular noises.

These challenges are further exacerbated by the existence of **two distinct sources of variation in functional data**: *amplitude* (variation along the y-axis) and *phase* (variation along the x-axis). For example, consider the Berkeley growth rate functions in Figure 1(a), which are the first derivative of measurements on height in centimeters for 93 children from age 1 to 18 [1]. Individuals can have different number and magnitude of growth spurts (amplitude) occurring at different ages (phase). These two weaved variation can cause significant issues for statistical inference: simply averaging growth rate curves at each time point (Figure 1(a), cyan line) fails to provide a good estimation for the general trend. This motivates the *registration* process, which "warps" the functions' domains to align their key features with respect to time. However, registration is often treated as a pre-processing step, with subsequent statistical modeling and inference applied to "well-aligned" data. Meanwhile, registration uncertainty is often ignored in downstream tasks.

My research addresses these fundamental challenges in registration and prediction for partially observed functional data [4]. The core innovation lies in integrating registration into *conformal prediction*, a framework providing finite-sample coverage guarantees without strong distributional assumptions [2]. More broadly, I also explored reliable recovery and uncertainty quantification of the fixed effect function in *Bayesian functional mixed model* that explicitly models object-level phase variation [3]. In addition, I have been collaborating on two application-driven projects. (i) Using neighborhood disinvestment score (NDS) based on Google street view streetscapes auditing, we build a model for *spatio-temporal universal kriging* of NDS and study its effect on colorectal cancer (CRC) survival. (ii) We prove the suboptimality of classification and regression trees (CART) for minimizing *latent probability classification risk* and fine-tune last node splits for further risk reduction, resulting in targeting more vulnerable subpopulations in policymaking [5]. In these cases, distribution-free and finite-sample valid uncertainty quantification remains largely unexplored, leading to rich future research directions.

I. Joint registration and conformal prediction for partially observed functional data.

Consider predicting a child's growth rate trajectory given observations up to age 9 (Figure 1(a), red line) and another $n = 92$ complete growth curves. Let $f_1, \dots, f_n$ denote fully observed growth rate functions and $f_{n+1}^{\mathcal{J}}$ denote partially observed test function on subinterval $\mathcal{J} = [1, 9]$, assuming $f_1, \dots, f_{n+1} \overset{iid}{\sim} \mathbb{P}_{\mathcal{F}}$, where $\mathbb{P}_{\mathcal{F}}$ is some probability distribution on the function space. To get finite-sample valid pointwise prediction intervals (PIs) for f_{n+1} without introducing further distributional assumptions, we adapt conformal prediction to functional data. In this context, the key methodological challenge is constructing *exchangeable* predictor-response pairs (X_i, Y_i), which are not naturally defined for functional data. Our solution is to treat the truncated observations from age 1 to 9 as predictors, i.e., $X_i = f_i^{\mathcal{J}}$ and the growth rate at any age $t \in [1, 18]$ as responses, i.e., $Y_i(t) = f_i(t)$ for $i =$

$1, \ldots, n+1$. Then standard conformal prediction algorithm can be applied repeatedly using $\{X_i, Y_i(t)\}$ for a uniform grid of $t \in [1, 18]$.

Figure 1(b) shows the prediction results: target function f_{n+1} in black, pointwise PIs in red and a point prediction in blue, which is taken to be midpoint of PIs at each time point. While the PIs provide reasonable coverage, the prediction band fails to capture the three growth spurts, which are the key geometric features of the growth rate function. Consequently, the point prediction, i.e., blue curve, also misses the key features. This is because naive functional conformal prediction (FCP) does not separate amplitude and phase variation. At a given age, some children are having growth spurts while others are not. Therefore, $Y_i(t)$ fails to provide enough useful information for predicting $Y_{n+1}(t)$. This is analogous to the case where the cross-sectional mean fails to recover the growth rate trend.

This limitation motivates incorporating registration, which separates amplitude and phase variation, to improve prediction accuracy. To see how it works, Figure 1(c) shows an example of two functions f_1 (blue) and f_2 (red) that have two peaks and one valley. Before registration, the peaks and valley are not aligned, indicating phase variation exists. After registering f_2 to f_1, as shown in black, these geometric features are well aligned with respect to time, suggesting phase variation has been removed. Registration is performed in a pairwise manner with a chosen template, e.g., f_1 in the previous example. For simultaneous registration of multiple functions, *sample Karcher mean* is often used as the template, which is analogous to sample average in the multivariate setting. To incorporate registration into FCP, we leverage *split conformal* methods. Specifically, we split data into training and calibration set, use training set to estimate sample Karcher mean, then register calibration functions to this template. The new responses are taken pointwise from registered functions with amplitude variation disentangled from phase. This construction maintains exchangeability while enabling conformal prediction to capture aligned geometric features.

Figure 1(d) shows sample Karcher mean of training set (cyan) and registered calibration functions. Comparing to those before registration in Figure 1(a), one can easily see the well-aligned growth spurts across observations. The corresponding prediction results are presented in Figure 1(e). The point prediction (blue) is very close to the underlying target function (black) and it clearly captures the three growth spurts. Moreover, the prediction band maintains coverage while being much tighter and informative than that in Figure 1(c). Summarily, our joint registration and conformal prediction framework provides valid, efficient and flexible uncertainty quantification for functional data with partial observation. Further, it opens a variety of exciting research directions. For instance, I am currently generalizing the method from two aspects: (i) a joint prediction for the entire function through basis expansion; (ii) joint conformal prediction and classification for 2-D shapes, with application to fossil bovid teeth. These will significantly improve the applicability and efficiency of the current framework.

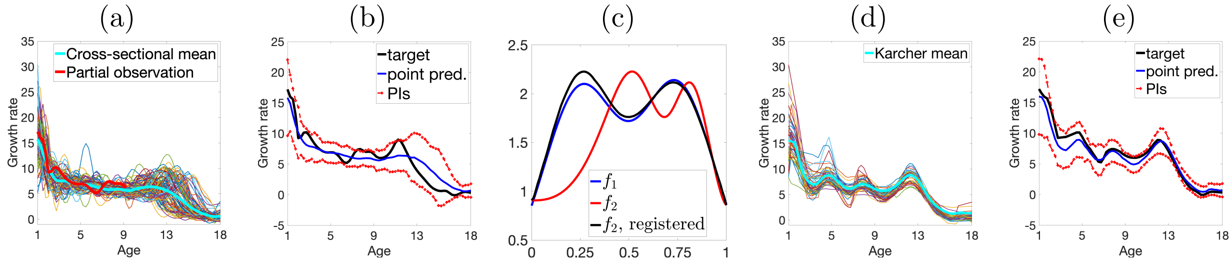

Figure 1: (a) Berkeley growth rate functions with cross-sectional mean (cyan) and a partial observation (red). (b) Target function (black), point prediction (blue) and pointwise PIs (red) for naive FCP. (c) Registration illustration, f_1 in blue, f_2 in red and f_2 after registration in black. (d) Karcher mean (cyan) and registered calibration functions. (e) Same as (b), but for joint registration and FCP.

II. Probabilistic size-and-shape functional mixed models.

My second research direction addresses reliable recovery and uncertainty quantification of population-level fixed effect function in functional mixed models. Again consider Berkeley growth data: while individual children deviate from the average growth pattern in both magnitude (amplitude) and timing (phase), we are interested in recovering the underlying population growth trend. Nevertheless, since amplitude and phase variations are confounded with measurement error, this problem is notoriously difficult in standard functional mixed models. Our work addresses this by answering *what* can be reliably recovered: instead of the fixed effect function itself, we target its *size-and-shape*. This geometric property is preserved under *norm-preserving action*, which can be interpreted as a rotation of coordinates in an infinite-dimensional Hilbert space. Specifically, we propose a Bayesian functional mixed model with two object-level random effects: a "size-and-shape–preserving" component γ_i that explains timing variability of growth spurts and a "size-and-shape–altering" component v_i that captures changes in number and magnitude of growth spurts. For inference purpose, fixed effect and random component v_i are expressed using suitable orthonormal basis of the Hilbert space. Also, we place informative priors on the random component γ_i, which allows for regularization of the posterior distribution of the fixed effect. Notably, inference for γ_i can be viewed as an automatic identification of a data optimal rotation of the chosen basis that best captures population and subject level variations.

Numerical experiments suggest our model successfully recovers representative geometric features of fixed effect that the state-of-the-art benchmark smooths away, enabling better scientific understanding of underlying population processes. For example, on Berkeley growth data, the posterior mean for fixed effect uncovers two growth spurts, a small initial one followed by a larger pubertal one, which agrees with previous literature.

III. Applied statistical methods for policy-relevant research.

Spatio-temporal universal kriging for neighborhood disinvestment in Columbus area. We study the effect of neighborhood disinvestment conditions on CRC prognosis and survival. Our finding suggests there exist seasonal and temporal effects for NDS, where winter has the worst neighborhood conditions comparing to other seasons. Meanwhile, spatial trends suggest urban areas have much higher NDS comparing to rural areas and precise spatial coordinates are more informative comparing to coarser spatial clusters. Using the kriging results, we average the predicted NDS over different time lags before CRC diagnosis and examine its association with survival time. Results show a consistent stage-specific signal: pre-diagnosis NDS relates to survival primarily for stage-2 patients, suggesting an "intermediate window" where neighborhood disinvestment may play a role.

Improving CART for binary latent probability classification. Policymakers often use CART to partition and target subpopulations whose outcome probability exceeds a threshold, e.g., 50%. We formalize this problem of "latent probability classification" (LPC) by defining the LPC misclassification risk and prove CART's split rules do not minimize this risk. For further risk reduction, we redefine the cost function in CART's final splits so that the zero-risk rules can be identified. When applied to real-world datasets, our proposed method generates policies that target more vulnerable subpopulations that CART fails to identify.

Challenges in uncertainty quantification. For spatio-temporal kriging and tree-based models, uncertainty quantification becomes challenging due to complex data structure and recursive nature of the method. For instance, the validity and accuracy of classic kriging variance estimates rely on correct specification of the covariance structure and become unstable when data is irregular and sparse. To apply conformal prediction, one needs a careful definition and evaluation of exchangeability for spatio-temporal data. For tree-based models, although data exchangeability is straightforward, conformal PIs may be overly conservative for binary classification. All these inspire interesting future research.

References

[1] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer, 2005.

[2] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[3] F. Wang, K. Bharath, O. Chkrebtii, and S. Kurtek. Probabilistic size-and-shape functional mixed models. *Advances in Neural Information Processing Systems*, 37:50031–50061, 2024.

[4] F. Wang, S. Kurtek, and Y. Zhang. Joint registration and conformal prediction for partially observed functional data. *arXiv preprint arXiv:2502.15000*, 2025.

[5] L. B. Wang, Z. Jiao, and F. Wang. Modifying final splits of classification tree for fine-tuning subpopulation target in policy making. *arXiv preprint arXiv:2502.15072*, 2025.