Research Statement

Fangyi Wang
Department of Statistics, The Ohio State University

My research focuses on statistical modeling and inference for functional data, where each observation is a curve varying over a continuum, e.g., daily temperature or children's growth rate. Analyzing functional data is challenging due to its infinite-dimensionality, complex dependencies within and across observations and irregular noise. In addition, there exists two sources of variation in functional data, namely *amplitude* (variation along the y-axis) and *phase* (variation along the x-axis). During my graduate studies, I developed (i) methods for **conformal prediction of partially observed functional data** and (ii) a **probabilistic functional mixed effect model**, both of which explicitly considered these two sources of variation. As a result, the developed methods are able to achieve more accurate inferential results and reveal important data features.

In addition, I finished two application-driven projects. In the first, We built and compared **spatio-temporal kriging models** for neighborhood disinvestment based on Google Streetview imagery auditing; we also tested for associations with colon and rectum cancer survival. In the second, we proved the suboptimality of classification and regression trees (CART) for **latent probability classification** and proposed methods that modify the final splits of CART, resulting in better performance while maintaining interpretability for policymaking. These sparked my interests in collaborating with domain experts and solving scientific problems with appropriate statistical tools.

Joint registration and conformal prediction for partially observed functional data *Major revision, JCGS*

Accurate prediction of future trajectories given historical functional data is an important question in many applications. For instance, given complete, fully observed children's growth rate functions from age 1 to 18, one may be interested in predicting a nine-year-old child's growth rate trajectory as the magnitude and timing of growth spurts (local maxima in growth rate functions) are important for disease diagnosis and prevention. Figure 1(a) shows growth rate functions from the Berkeley growth data, with a (simulated) partial observation highlighted in red. Different children have different numbers and magnitudes of growth spurts (amplitude variation) that occur at different ages (phase variation). The goal is to construct pointwise prediction intervals (PIs) for the partial observation with a coverage validity guarantee. *Conformal prediction* is well-suited for this task as it provides PIs that have a finite-sample coverage guarantee without imposing strong assumptions on the data generating process. In the functional data context, applying conformal prediction requires a careful construction of exchangeable predictor-response data pairs, which are not naturally defined. Our solution was to treat the partial observations from age 1 to 9 as predictors and the function values at any specific age as responses. Based on this construction, we then applied conformal prediction on a fixed, uniform grid of time points from age 1 to 18. Figure 1(b) shows the prediction results. While the PIs (red) provide reasonable coverage for the true target function (black), 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 (blue), which is taken to be midpoint of PIs at each time point, also misses the key features. This is because amplitude and phase variation are entangled and growth rate function values of other children at the same age fail to provide accurate information for prediction.

This motivated incorporating *registration*, which separates amplitude and phase variation, into functional conformal prediction to improve prediction accuracy. Figure 1(c) shows an example of registration, where f_2 (red) is being registered to f_1 (blue), with the registered function shown in black. After registration, the functions' geometric features, i.e., two peaks and one valley, are much better aligned. Registration is performed in a pairwise manner with respect to a chosen template, e.g., f_1 in the previous example. For simultaneous registration of multiple functions, a *sample Karcher mean* is often used as the template. We leveraged *split conformal* methods to incorporate registration. We first split the data into training and calibration sets. We used the training set to estimate the sample Karcher mean, and then registered calibration functions to it. The new responses were defined pointwise based on the registered functions in the calibration set. This construction maintains exchangeability while enabling conformal prediction to capture aligned geometric features. The corresponding prediction results are presented in Figure 1(d). The

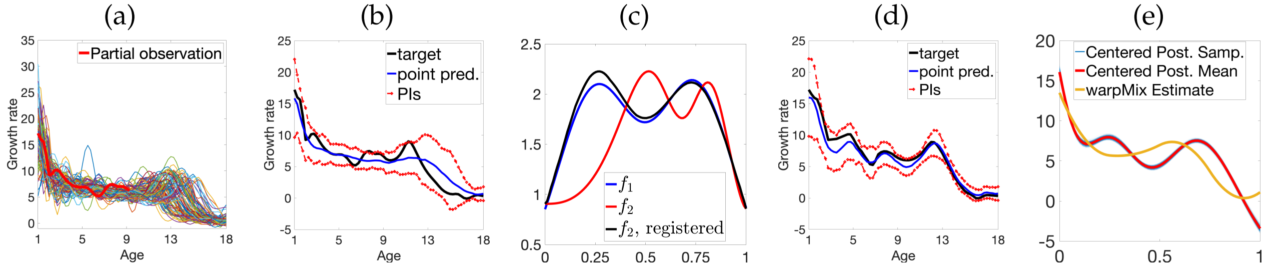

Figure 1: (a) Berkeley growth rate functions with a partial observation (red). (b) Target function (black), point prediction (blue) and pointwise PIs (red) based on functional conformal prediction. (c) Illustration of registration: f_1 in blue, f_2 in red and f_2 after registration in black. (d) Same as (b), but based on joint registration and functional conformal prediction. (e) Posterior samples (blue) and posterior mean (red) of fixed effect function, and `warpMix` estimate (yellow).

point prediction (blue) is very close to the underlying target function (black) and it clearly captures the three growth spurts. Moreover, the prediction band is much tighter and more informative. I am currently developing a conformal prediction framework for shapes of planar closed curves. Instead of a pointwise procedure, we use a basis expansion in an appropriate tangent space and perform predictions via basis coefficients.

Future research directions:

(i) When data comes from a heterogeneous population with significant amplitude variation across subpopulations, an overall Karcher mean may not be a good representative of any of the subpopulations. As a result, registration and prediction become less effective. I plan to incorporate mixture models into the current framework to account for population heterogeneity.

(ii) The proposed method satisfies coverage validity and improves prediction accuracy through registration. However, I want to get a better understanding of the theoretical properties of the PI lengths to answer the following question: can we further improve prediction accuracy and what is the best result we can hope for?

(iii) I'd like to generalize the framework to more complicated data structures, such as spatial functional data or functional time series, where the i.i.d. assumption is no longer satisfied.

Probabilistic size-and-shape functional mixed models *NeurIPS 2024*

Reliable recovery and uncertainty quantification of a population-level fixed effect function in functional mixed models is a challenging problem. Again, consider the 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 variation are confounded with measurement error, this task is difficult using standard functional mixed models. We addressed this by answering *what* can be reliably recovered: instead of the fixed effect function itself, we targeted its *size-and-shape*. This geometric property is preserved under the *norm-preserving action* of the group of phase functions, which can be interpreted as a rotation of coordinates in an infinite-dimensional Hilbert space. Specifically, we proposed 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 the number and magnitude of growth spurts. For inference, we expressed the fixed effect and random components v_i using suitable orthonormal bases of the Hilbert space. Also, we placed 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. Figure 1(e) shows posterior samples (blue) and the posterior mean (red) of the fixed effect function when the proposed model was applied to Berkeley growth data. Estimation using `warpMix`, a state-of-the-art frequentist functional mixed model proposed by Claeskens et al. [2021], is shown in yellow for comparison. The posterior mean of the fixed effect uncovers two growth spurts, a small initial one followed by a larger pubertal one, whereas `warpMix` is only able to recover one small growth spurt. These results suggest that our model successfully recovers representative geometric features of the fixed effect function that the benchmark smooths away.

Future research directions:

(i) I plan to develop data-driven methods for selecting the type and number of basis functions for the fixed effect and size-and-shape altering random effects.

(ii) I will extend the framework to sparse/fragmented functional or higher dimensional objects such as curves and surfaces.

(iii) I plan to explore more efficient MCMC algorithms for posterior sampling or variational methods.

Applied statistical modeling in public health and social sciences

Spatio-temporal modeling for neighborhood disinvestment ***Under review***

Neighborhood disinvestment, characterized by physical disorder in the built environment, has been linked to health behaviors and outcomes, including cancer survival. While previous studies have focused on spatial characteristics of built environment sampling schemes, temporal dimensions of disinvestment remain underexplored, despite potential relevance for long-latency outcomes such as colon and rectum cancer (CRC). We described a neighborhood auditing procedure based on Google Streetview imagery in Franklin County, Ohio, developed and compared spatio-temporal kriging models of neighborhood disinvestment, and examined time-lagged associations between disinvestment and CRC survival. Our analysis revealed a large spatial correlation in neighborhood disinvestment within a short distance and moderate, long-lasting temporal dependence, with precise location information and classic spatio-temporal kriging providing the most accurate predictions. We found evidence that higher disinvestment prior to diagnosis was significantly associated with reduced survival time, particularly among patients diagnosed at regional stage.

Future research directions:

(i) I will develop methods that incorporate residential mobility patterns of CRC cases used in the external validation analysis, which requires accounting for spatial correlation in the survival model.

(ii) I plan to jointly perform prediction and validation, and quantify the corresponding uncertainty of each step.

Improving CART for latent probability classification ***Under review***

Policymakers often use recursive binary split rules to partition populations based on binary outcomes and target subpopulations whose probability of the binary event exceeds a threshold. We formalized the latent probability classification (LPC) problem and proved classic CART are suboptimal for this task. Adapting concepts from decision theory, we proposed methods that modify the final splits of CART and generate split rules that strictly dominate CART's rules. When applied to real-world datasets, our proposed methods generate policies that target more vulnerable subpopulations that CART fails to identify.

Future research directions:

(i) We will extend the framework to policy learning while keeping the interpretability of tree-based methods.

(ii) We will modify the procedure to incorporate potential budget, fairness, or coverage constraints.

(iii) We plan to further explore asymptotic and finite-sample statistical properties of the proposed estimator of the split rule.

(iv) We will develop methods for uncertainty quantification.

References

G. Claeskens, E. Devijver, and I. Gijbels. Nonlinear mixed effects modeling and warping for functional data using B-splines. *Electronic Journal of Statistics*, 15(2):5245–5282, 2021.