Research Statement

Fangyi Wang
Department of Statistics, The Ohio State University

My research focuses on statistical modeling and inference for functional data, which is challenging due to their infinite-dimensionality, complex dependencies and irregular noises. In addition, there exists two sources of variation in functional data, namely *amplitude* (variation along the x-axis) and *phase* (variation along the y-axis). I developed methods for **conformal prediction for partially observed functional data** and **probabilistic functional mixed effect model** that explicitly considered these two sources of variation, which results in more accurate inferential results and reveals important data features. In addition, I finished two application-driven projects. (i) We built and compared **spatio-temporal kriging models** for neighborhood disinvestment based on Google Streetview imagery auditing and test its associations with colon and rectum cancer survival. (ii) 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.

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 are important for disease diagnosis and prevention. Figure 1(a) shows growth rate functions from Berkeley growth data, with a (simulated) partial observation highlighted in red. Different children have different number and magnitude 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 coverage validity. *Conformal prediction* is well-suited for this task, which 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, then repeatedly apply standard 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 same age fail to provide accurate information.

This motivates 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, their geometric features, i.e., two peaks and one valley, are much better aligned. 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. We leveraged *split conformal* methods, splitting data into training and calibration set, using training set to estimate sample Karcher mean, then register calibration functions to it. The new responses are taken pointwise from registered functions. This construction maintains exchangeability while enabling conformal prediction to capture aligned geometric features. The corresponding

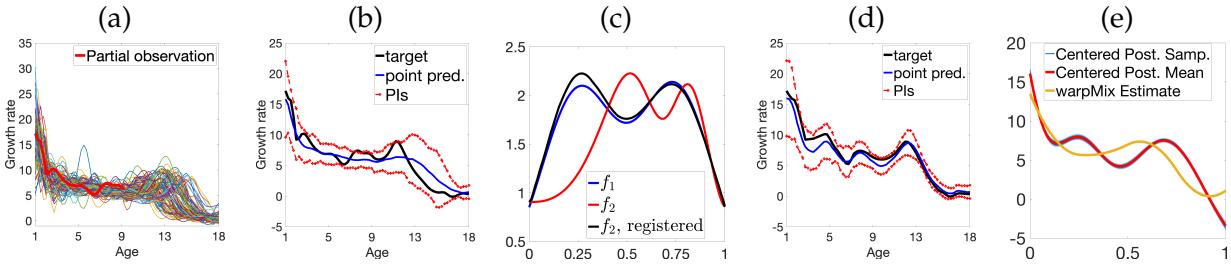

Figure 1: (a) Berkeley growth rate functions with a partial observation (red). (b) Target function (black), point prediction (blue) and pointwise PIs (red) for functional conformal prediction. (c) Registration illustration, f_1 in blue, f_2 in red and f_2 after registration in black. (d) Same as (b),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

prediction results are presented in Figure 1(d). 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 is much tighter and more informative. I am currently developing conformal prediction framework for shapes of planar closed curves. Instead of a pointwise procedure, we use basis expansion and perform predictions on basis coefficients.

Future research directions:

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

(ii) Our proposed method satisfies coverage validity and improves prediction accuracy through registration, but I want to get a better understanding of the theoretical properties of the PI length. Can we further improve accuracy and what is the best we can get?

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

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

Reliable recovery and uncertainty quantification of population-level fixed effect function in functional mixed models is a challenging problem. 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 variation are confounded with measurement error, this problem is difficult in 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 *norm-preserving action*, 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 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 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 posterior mean (red) of the

fixed effect function when applying the proposed model 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 for 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 our model successfully recovers representative geometric features of fixed effect that the benchmark smooths away.

Future research directions: (i) Develop data-driven methods for selecting the number of basis. (ii) Extend the framework to sparse/fragmented functions or higher dimensional objects such as curves and surfaces. (iii) Explore more efficient algorithms for MCMC computations.

Applied statistical modeling in public health and social sciences.

Spatio-temporal modeling for neighborhood disinvestment. *In preparation*

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 large spatial correlation in neighborhood disinvestment (ND) 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) Develop methods that incorporate residential mobility patterns of CRC cases used in the external validation analysis. (ii) Jointly perform prediction and validation and quantify the corresponding uncertainty of each step.

Improving (KD-)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 and the knowledge distillation (KD) method, whose student model is a CART, are suboptimal for this task. Adapting concepts from decision theory, we proposed methods that modify the final splits of (KD-)CART and generate split rules that strictly dominant (KD-) CART's rules. When applied to real-world datasets, our proposed methods generates policies that target more vulnerable subpopulations that (KD-)CART fails to identify.

Future research directions: (i) Extend the framework to policy learning while keeping the interpretability of tree based methods. (ii) Modify the procedure to incorporate potential budget, fairness, or coverage constraints. (iii) Further explore asymptotic and finite-sample statistical properties of the proposed estimator of the split rule. (iv) 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.