Research Statement

Luo Xiao

I am interested both in applied and theoretical statistical problems, which is reflected in my choice of doctoral training in statistics at Cornell and postdoctoral training in biostatistics at Johns Hopkins. The emergence of the ever larger and more complex data in real world problems raises statistical challenges that require theoretical, methodological and computational development and innovation. Addressing these challenges has motivated my research and shaped my publication record.

My research interests include: 1) smoothing and functional data analysis; 2) high dimensional statistics; 3) real data applications in wearable computing, children growth, and brain imaging. Below I describe the specific areas that I have worked on and some directions that I intend to pursue in the future.

Bivariate smoothing and functional data models

My first research project during my doctoral studies was about bivariate spline smoothing. In the published paper *Fast bivariate P-splines: the sandwich smoother*, I proposed a fast penalized spline method, the sandwich smoother, for smoothing bivariate data on a regular grid. The sandwich smoother provides the next level of computational scalability for bivariate smoothers and has significant computational advantages over competing methods. This is achieved, essentially, by transforming the technical problem of bivariate smoothing into a short sequence of univariate smoothing steps. I also derived the asymptotic distribution of the sandwich smoother and proved its rate-optimality.

At Hopkins, I worked on the scientific problem of quantifying the circadian rhythm of physical activity and its relation with age. The physical activity data from the Baltimore Longitudinal Study on Aging has minute-by-minute counts for more than 700 subjects, with each subject having multiple days of data. While this can be naturally formulated as a bivariate smoothing problem, accounting for the correlation structure and the large data size is very challenging for any smoothing method. In the "to appear" paper *Quantifying the life-time circadian rhythm of physical activity: a covariate-dependent functional approach*, I extended the sandwich smoother for correlated functional data and designed a fast algorithm for selecting the smoothing parameters. Moreover, to quantify the subject-level and day-to-day within-subject random circadian rhythms of physical activity, I have developed an age-dependent multi-level fPCA, where covariance functions were estimated by a novel trivariate spline smoother.

In the "to appear" paper *Fast covariance estimation for high-dimensional functional data*, I developed a fast spline smoothing method for covariance operators for functional data

with thousands of data points per curve. In particular, I extended the method to structured functional data including multilevel or longitudinal structures. The method was successfully applied to the Sleep Heart Health Study (SHHS) sleep EEG data.

Currently I am working on extending and analyzing theoretically existing models for structured functional data, inspired by the structure of new studies that are funded by NIA, NIMH and NINDS. One research project is to design a penalized spline smoother for estimating covariance operators using irregularly spaced functional data. Another project is to develop functional regression models where the parameter of interest is a bivariate smooth function.

Regression models for multiple outcomes or multivariate responses

During my doctoral studies at Cornell I also worked on quantifying the effects of prenatal exposure to mercury on the functioning of the central nervous system of children in the Seychelles Child Development Study (SCDS). The SCDS data include 20 outcomes measured on 9-year old children that can be classified broadly in four outcome classes or "domains": cognition, memory, motor, and social behavior. Previous analyses and scientific theory suggested that these outcomes may belong to more than one of these domains, rather than a single domain, as was frequently assumed for modeling. Our goal was to examine the effects of exposure and other covariates when outcomes may belong to more than one domain and domain assignment was unknown. I proposed a linear mixed model with a sparse component to model the domain-membership vectors of outcomes. To implement a Bayesian MCMC approach, I designed a novel prior for the domain-membership vectors and used a random-walk Metropolis-Hasting sampler.

After my doctoral studies I have developed an interest in reduced rank models for multivariate responses. In the "to appear" paper *Learning regulatory programs by threshold SVD regression*, I proposed a new model where both predictor selection and response selection were conducted through thresholding the singular value decomposition of a coefficient matrix. I carefully calibrated the estimation algorithm to handle large numbers of predictors and responses. The model was used to infer the regulatory relationships between different genome-wide measurements from the complex biological systems where the numbers of predictors and responses far exceeded the sample size.

Currently I am working on extending reduced rank models to multivariate binary responses, and to longitudinally observed multivariate responses. This line of research has suggested many future research directions that I am eager to pursue. On the methodological side, there is a great need for developing inferential frameworks for reduced rank models, especially in high dimensional data settings. On the data application side, the methods were applied to genetic data, but they are also useful in some brain imaging

applications that I have started to work on.

Covariance matrix estimation

Covariance estimation is an important problem in high dimensional statistics. In the "to appear" paper *On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA*, I studied properties of the sample covariance matrix as an estimator of population matrices of reduced effective rank. The effective rank of a matrix is the ratio of its trace to its largest singular value, and provides a measure of complexity. I established sharp finite sample bounds on the operator and Frobenius norm of the estimation error. In particular, I showed that when the effective rank is smaller than the sample size (up to logarithmic factors), the sample covariance matrix could still serve as an accurate estimator, even if the dimension of the population matrix is larger than the sample size. Examples of covariance matrices of reduced effective rank can be found in functional data models and in finite factor models.

In two other "under review" papers, I considered estimators of bandable covariances, whose elements decay to zero as the distance from the diagonal increases. In the paper *On the theoretic and practical merits of the banding estimator for large covariance matrices*, I proved that the banding estimator proposed by Bickel and Levina achieves rate-optimality under the operator norm, for a class of approximately banded covariance matrices. In addition, I proposed a Stein's unbiased risk estimate (Sure)-type approach for selecting the bandwidth for the banding estimator. In the paper *Convex banding of covariance matrices*, a new sparse estimator of high-dimensional bandable covariance matrices was proposed and I showed that it is theoretically optimal.

Research on wearable computing

During my postdoctoral studies, I have been involved and got interested in wearable computing research motivated by data collected from body-worn sensors. Accelerometry data, in raw format, are high-frequency triaxial accelerations, which are proxies of intensity and direction of movement. In processed format, data are often measured as counts at the minute level and are proxies of the average activity intensity in each minute. Accelerometry data provide objective and detailed measurements of physical activity and have been widely used in observational studies and clinical trials.

In the "tentatively accepted" paper *Movement prediction using accelerometers in a human population*, I introduced statistical methods for predicting the types of human activity at sub-second resolution using raw triaxial accelerometry data. The major innovation is that I used labeled activity data from some subjects using movelets, a type of dictionary learning, to predict the activity labels of other subjects. Prediction results based on other people's labeled dictionaries performed almost as well as those obtained using their own

labeled dictionaries. These findings indicate that prediction of activity types for data collected during natural activities of daily living may actually be possible.

I have also worked directly with Jennifer Schrack (Assistant Professor at Johns Hopkins epidemiology) and Luigi Ferrucci (Chief of Longitudinal Studies Section in NIA) on a comparative study of activity between men and women across the life span. This resulted in the manuscript *Quantification of gender differences in sedentary behavior measured by accelerometers* that will be submitted soon.

Following the above line of research, I have been working on several projects on accelerometry data including (1) developing statistical methods for identifying walking in the accelerometry data collected from real life; and (2) quantifying different types of activities in terms of accelerometry counts.

Other ongoing collaborative works

I have been working directly with William Checkley (Assistant Professor of medicine at Johns Hopkins) on quantifying children growth of height and weight. In one project I am using nonlinear mixed effects model to predict children growth. In another project, I am using historical functional linear regression models to quantify the historical effect of weight on height.

I am also involved in developing PCA methods for matrix-variate data with application to fMRI.