STATEMENT OF RESEARCH INTERESTS

My work spans several major areas of statistics and is driven by an interest in developing statistically sound solutions to real-world problems. My methodological research interests can be categorized into the two broad areas: analysis of incomplete multivariate data and stochastic simulation. Although my concentration has been on biostatistics in recent years, the fundamental principles of these research topics can be applied to a wide range of disciplines including finance, econometrics, mathematics, engineering, marketing, quantitative social and behavioral sciences, among many others. Having engineering and business administration degrees, followed by a doctoral level training from a mathematical statistics department, and currently working in a public-health oriented biostatistics department, give me a good opportunity for striking a delicate balance of technique and judgment. In other words, methodological advancements guided by applied context of the problem ideally suits to my background.

As far as missing data are concerned, I focus on multiple imputation (MI) inferences where missing observations are replaced by a set of plausible values that are drawn from a predictive distribution. Once multiple versions of the completed data sets are obtained, one can proceed with analyzing them with standard complete-data methods, and incorporating the results into a single inferential summary. As a result, with MI, uncertainty due to missing data is formally taken into account in the modeling process. In many studies, incompleteness arises due to a myriad of reasons that are beyond the control of researchers, and adequately addressing missing data issues in a statistically defensible manner is crucially important. I have been working on MI-based methodologies—especially in incomplete longitudinal data context—as well as random number generation over the years that led to more than 50 methodological, peer-reviewed articles published in major statistical journals.

From stochastic simulation point of view, I am deeply interested in random number generation, statistical distribution theory, and modern computational statistics topics such as EM-related algorithms, MCMC (data augmentation, Gibbs sampler), density estimation, bootstrapping, smoothing, Bayesian computing and non-linear optimization to the extent they apply to missing-data problems, and more broadly to develop new techniques for data generation within the stochastic simulation paradigm. I plan to maintain a substantial degree of exposure to the simulation notion in

the context of incomplete data in the future. I find the idea of simulation -- where we describe a real phenomenon by generating an environment within which stochastic process under consideration is assumed to operate-- exciting. Creating mirror images and imperfect proxies of the perceived truth, and iteratively refining and re-defining the truth have always been thrilling to me.

My intermediate research goal is relating random number generation/statistical distribution theory and MI, which I believe are closely connected. The essential step in parametric MI is filling in the missing data by drawing from the conditional distribution of the missing data given the observed data. This usually entails positing a model for the data and using it to derive this conditional distribution. For continuous data, multivariate normality among the variables has been perceived as a convenient assumption since the conditional distribution of the missing data given the observed data is then also multivariate normal. Recently, extending the practice of MI from normality to more general classes of densities has begun to receive attention. Considering the restrictive nature of the normality assumption, employing a distributional setup that spans a broader spectrum of symmetry-elongation behavior in the imputation process may provide a reasonable way to handle non-Gaussian continuous data. In this regard, generalized distributions are sensible alternatives because of the ability of accommodating a variety of distributional shapes. Imputation under generalized distributions such as Tukey's classes, the Burr family, the Johnson Family, the Pearson family, generalized lambda and Beta families, and Fleishman polynomials, that include many standard densities as exact or approximate special cases, appears to have significant potential to capture real-data trends. I have recently placed a specific research emphasis on augmenting power polynomials that have been developed for characterizing the behavior of continuous distributions to multivariate discrete distributions. I believe that this line of research has prospects to be very fruitful. On a related note, one can take advantage of well-studied computational and conceptual merits of Gaussian imputation model, yet still adequately address non-normal features (skewness, heavy tails, flat regions, multimodality, skip patterns, etc.) of data by probability integral transformation that makes a connection between quantile functions. This promising possible solution has not been investigated from an imputation standpoint. I would like to direct a major portion of my intellectual energy to this, along with richer families of parametric distributions that I mention above.

Thinking beyond the realm of normality also has potential of leading to substantial advances in modeling non-normal random effects, especially in longitudinal settings. Estimating random effects nonparametrically, and subsequently expressing the empirical cumulative distribution function in terms of normal quantile function through probability integral transformation might be a viable approach. Non-normal MI models and random effects can be studied simultaneously considering their shared characteristics, and that is what I plan to do in the near future.

I will also build a unified framework for concurrently generating data sets that include all four major kinds of variables (i.e., binary, ordinal, count and continuous) when the marginal distributions and a feasible association structure in the form of Pearson correlations are specified for simulation purposes. Stochastic simulation is an indispensable part and a major focus of scientific inquiry. Model building, estimation and testing typically require verification via simulation to assess the validity, reliability and plausibility of inferential techniques, to evaluate how well the implemented models capture the specified true population values, and how reasonably these models respond to departures from underlying assumptions, among other things. Describing a real notion by creating mirror images and imperfect proxies of the perceived underlying truth in a repeated manner allows researchers to study the performance of their methods through simulated data replicates that mimic the real data characteristics of interest in any given setting. Resulting accuracy and precision measures regarding the parameters under consideration signal if the procedure works properly, and may suggest remedial action to minimize the discrepancies between expectation and reality. Simulations have been commonly employed in a wide spectrum of research fields including the physical, medical, social and managerial sciences as well as in engineering, in order to better comprehend today's increasingly sophisticated issues. A central aspect of every simulation study is the quantification of the model components and parameters that jointly define a scientific process. When this quantification cannot be performed via deterministic tools, researchers resort to random number generation (RNG) as a starting point in finding simulation-based answers to address the stochastic nature of the problem. Such problems typically involve variables of many different types on a structural level; i.e., causal and correlational interdependencies are a function of a mixture of binary, ordinal, count and continuous variables, which act simultaneously to characterize the mechanisms that collectively delineate a phenomenon. Although many RNG algorithms have appeared in the literature, a major

limitation is that they were designed for only one type of data. To the best of our knowledge, this project is the first systematic attempt and comprehensive investigation for mixed data generation. As such, it represents a substantial augmentation of the existing methods, and it has potential to advance scientific research in a meaningful way. The potential broader impact of this framework is that methodologists and practitioners in many disciplines, who may lack statistical and/or computational dexterity, will be able to simulate mixed data with relative ease due to the planned R package that will be made publicly available soon. The creation of this R package will extend and complement our previously developed packages.