Research Statement

Li Li

Research has been an integral part of my scholarly endeavors over the last decade. The opportunity to investigate many research questions has afforded me the chance to advance my learning experience and to share these developments with my students. In my five years as an Assistant Professor of UNM, I have continued research in the areas of Bayesian reliability and survival modeling, which were the topics of my Ph.D. dissertation. I have also investigated new topics, including time series, spatial statistics, and deep learning. I collaborate with epidemiologists at the UNM Comprehensive Cancer Center, which has drawn me into missing data techniques, causal inference, and mediation analysis. I enjoy learning all that I can in these areas, and I believe I can make some additional meaningful contributions to statistical methods and application fields (for example, epidemiology applications). I now present my ongoing research accomplishments and interests in more detail:

1 Reliability and survival modeling

Broadly, reliability and survival modeling are categories of research dealing with lifetime data, which typically records the length of time until events of interest, such as mechanical failures, heart attacks, and cancer relapses. Often subjects are monitored over multiple time points as well as their locations; hence, longitudinal and spatial dimensions complicate the data correlation structure. A significant number of scientific questions arise in understanding the process that generates the events and the effects of risk factors on the time distributions until events. The risk factors are numerical quantities that identify treatment or intervention codes, demographic information, geographic records, and behavioral information. Drawing proper conclusions (inferences) about the risk factors' effects relies on reasonable model assumptions. One major assumption that can affect inference results unduly is on the distribution family of time-to-event. Bayesian nonparametric methods avoid misspecifying the distribution family by placing priors on the space of densities. Despite their flexibilities in modeling distributions, Bayesian nonparametric methods has limited applications in reliability and survival modeling due to modeling complexities and computational

challenges. Besides inferences on risk factors' effects, statistical studies also typically involve diagnosing possible problems with model assumptions and making predictions for unknown responses of interest. My vision for lifetime data modeling is to develop and apply Bayesian nonparametric methods along with spatial statistics and time series techniques to accommodate complex lifetime data structures. My achievements in these categories include five first-authored publications in high quality refereed journals, with three of these among the top 5 journals in Statistics and the other two among the top 10: *Technometrics*, *Computational Statistics and Data Analysis*, *Biometrics*, *Journal of Royal Statistical Society, Series A*, and *Journal of American Statistical Association*.

One of my specific interests regards recurrent events data for biological subjects or repairable systems, each of whom may experience transient clinical or mechanical events repeatedly during the period of observation. Statistical research on recurrent events has often relied on simplified model assumptions, including the commonly assumed Poisson process, to quantify the dynamic risk of failure. However, there is an essential absence of literature on testing this assumption. I developed (with Timothy Hanson, Paul Damien, and Elmira Popova) a Bayesian nonparametric test to detect departures from this assumption and the paper was published in *Technometrics*. The test statistic is a pseudo-Bayes factor for comparing two models with the usual Poisson model being the null hypothesis and a two-stage generalization of the Poisson process model being the alternative. Both models were constructed using flexible Bayesian nonparametric tailfree priors for the unknown distributions and fitted using MCMC techniques. Here tailfree priors are constructed on sets that partition the positive reals and sets of conditional probabilities that are parameters. The tailfree priors are centered at the parametric families but they allow for substantial data-driven deviations from the centering families.

Although Poisson processes imply all interventions (repairs) bring a subject to "good as old" states, the reality is that interventions more often than not have effectivenesses other than "good as old". For this more realistic scenario, I developed (with Timothy Hanson) a novel regression model that allows a spectrum of heterogeneous repairs besides "good as new" and "good as old" repairs and the paper was published in *Computational Statistics and Data Analysis*. The proposed semiparametric regression model was based on Kijima's effective age process, to account for intervention-specific characteristics, e.g., cost of repair, type of maintenance, the importance of subcomponent, etc. Essentially the "internal age" is a function of these covariates. Linearity in the predictors was relaxed using a B-spline transformation.

Still, there has been limited work that is flexible enough to accommodate a wider spectrum of effectiveness than those implied by the Poisson process and renewal process. In addition to the challenge of precise modeling of interventions effects, we face time-varying treatments, time-

dependent treatment effects, and accumulative effects of treatments, which are all important issues in comparing a treatment regimen's effectiveness. The investigation of intervention' effects can be complicated when using observational data due to possible confounding effects from other variables (confounders). A confounder is a variable that influences both the dependent variable (failure time) and the independent variable (risk factor), causing a spurious association. I am interested in incorporating confounding-related covariate information into lifetime data modeling. More about confounding is in my description of causal inference and mediation analysis below.

Another type of lifetime data I have worked on is from large environmental and epidemiological lifetime studies. These often involve subject-specific geographic and other background information. Inferential tasks, such as modeling of trends and spatial correlation structures and estimation of underlying model parameters are of paramount importance with such data. Many widely applied methodologies fail under an explosion in data volume, and there remains a lack of flexible yet interpretable models. For example, cancer registry data sets typically record each patient's location up to a district or county due to patient confidentiality (lattice data). Analysis of lattice survival data has traditionally assumed a Cox proportional hazards model for describing covariates' effects and spatially correlated random effects to account for geographic variabilities. However, the proportional hazards model simply does not fit many data sets, and the often-used random-effects modeling lacks population-averaged interpretations for the regression coefficients. I (with Timothy Hanson and Jiajia Zhang) proposed an innovative new Bayesian semi-parametric estimation procedure for an extended hazards model that includes the proportional hazards, accelerated failure time and accelerated hazards models as formally nested cases. We also developed new Bayesian tests for each of special cases (reduced models). To account for spatial variability, we generalized the copula approach via normal transformation for point-referenced data by Li and Lin (2006) to large lattice data. Since we seek to formally test whether simpler models are adequate relative to the EH model with spatial correlation, typical spatial random effects complicate such tests. The copula approach, however, allows careful modeling of the spatial correlation, admits population-averaged covariate effects, and allows for the tests for reduced models. Spatial smoothing is introduced through a class of intrinsic autoregressive processes. We developed highly efficient MCMC algorithms applicable to large censored data sets. The work was published in *Biometrics*.

Regarding epidemiological lifetime data, there has been an increasing interest in studying the proportion of cured patients among those who had been receiving treatment for cancer. I (with Ji-Hyun Lee) developed a new cure rate model and applied it to a large New Mexico breast cancer registration data where a fraction of breast cancer patients are long-term survivors. Our model assesses treatment's effect on the probability of being cured for each individual and quantifies the

associated sub-population variability. Kim et al. (2009) proposed a latent promotion time cure rate marker model for right-censored survival data. They assumed the cure rate parameter of a targeted population to distribute over a number of ordinal levels according to probabilities governed by risk factors. We proposed to use a mixture of linear dependent tail-free processes as prior for the distribution of the cure rate parameter. The proposed approach can accommodate a richer class of distributions for the cure rate parameter, and hence avoid oversimplifying the number of ordinal cure rate levels. The algorithms developed in this work also allow the fitting of several survival models for metastatic tumor cells. The work was published in *Journal of Royal Statistical Society, Series A*.

Motivated by data gathered in an oral health study, I (with Alejandro Jara, Maria Jose Garcia-Zattera, and Timothy Hanson) developed a model for correlated time-to-event data when the responses can only be determined to lie in an interval obtained from a sequence of examination times and the determination of the occurrence of the event is subject to misclassification. The joint model for true and unobserved time-to-event data is defined in a semiparametric way, where the marginal distribution is specified by considering a tailfree prior for the baseline distribution and standard assumptions on the relationship between the predictors and the responses, such as accelerated failure time, proportional hazards and proportional odds. The joint model is completed by considering a parametric copula function. The copula approach allows careful modeling of the within-unit correlations and admits population-averaged covariate effects. A general misclassification model was discussed in detail, considering the possibility that different examiners were involved in the assessment of the occurrence of the events for a given subject across time. We provided empirical evidence that the model can be used to estimate the underlying time-to-event distribution and the misclassification parameters without any external information about the latter parameters. This work was published in *Journal of American Statistical Association*.

2 Missing data methods

Missing data often occur in data collection due to nonresponse – no information provided for one or more items or a whole unit, mistakes made in data entry, and even sample corruption. Most methods and theories were developed for full data, and they could be much more complicated if they are needed to handle missing data properly. My vision is to extend state-of-the-art models to accommodate missing data. My achievements in this category include one publication in *Statistical Modelling*.

Motivated by a smoking cessation intervention study, I (with Ji-Hyun Lee, Steven Sutton,

Vani Simmons, and Thomas Brandon) developed an innovative new Bayesian transition model for missing longitudinal binary outcomes. Data include observed smoking status at discrete follow-up assessments, with missing data in different amounts at each assessment. Smoking status in these studies is a dynamic process with individuals transitioning from smoking to abstinence, as well as abstinence to smoking, at different times during the intervention. In this work, we model changes in smoking status and examine how interventions and other covariates affect the transitions. We use a selection model for the missing outcomes, which accommodates the commonly seen missing mechanisms: missing completely at random, missing at random, and missing not at random. We proposed a Bayesian approach for fitting the joint model to observed data and imputing missing outcomes based on the selection model. This work was published in *Statistical Modelling*.

3 Casual inference and mediation analysis

Causal inference is the process of concluding a causal connection based on the conditions of the occurrence of an effect. Causality lies at the heart of many scientific research endeavors, including Statistics, Biostatistics, Epidemiology, Economics, Computer Science, Data Science, Sociology, Political Science, to name a few. Several application studies I joined, aim to investigate the effects of heavy metal exposures on children's neuro-developmental assessments. Causal analysis plays a key role in disentangling the effects of metals; for example, the effects of Arsenic, from the effects of Barium, which is a pollutant showing up frequently from the same source with Arsenic. Without controlling for the confounding effects of Barium, the estimated effects of Arsenic could be misleading; i.e., if we observe negative associations between Arsenic and neuro-developmental assessments, it is possible that the negative associations are partially induced by the negative effects of Barium. In fact, we observe many spurious associations between heavy metals and outcomes. Existing causal inference methods have high potentials in such applications. Causal discovery and methods for dealing with unmeasured confounders are still open research areas. As a step further for causal inference, mediation analysis searches for mechanisms behind causes and effects. It is quite exciting to see actual applications that unlock mechanisms from observational data. The preliminary results of the neuro-developmental assessments for the Navajo children are published in *Child: care, health and development* where we found that Navajo children had higher percentages of children at-risk for developmental delays than those from the national sample. Several manuscripts are under development about the toxic effects of heavy metals on neuro-developmental assessments.

To summarize, my research interests are mainly in Bayesian complex lifetime modeling, missing data techniques, causal inference, and mediation analysis. I have regularly published in the top

journals in Statistics, and I have great interests for continuing my current work and branching out into exciting new directions with even greater productivity.