RESEARCH STATEMENT

Fei Jiang (fj2@rice.edu)

My researches involve the statistical methodology developments in the experimental design and observational data analysis areas. The classical devices for studying the areas are developed from the frequentist's points of view, while more and more Bayesian methods have emerged along with the development of computational technologies. Working with Dr. J. Jack Lee, Dr. Peter Mueller, and Dr. Yanyuan Ma, I have experienced the theoretical and practical developments of the subjects from both the frequentist and Bayesian perspectives.

For the experimental designs, I focus on developing adaptive randomization procedures for clinical trials. The studies introduce new models and procedures to enhance the trial ethics and efficiency . The resulted trials reduce the risks of patients to expose to the treatments with unsure effects, and allow the new efficacious treatments to be approved sooner.

The adaptive randomization designs allow the decision makings occur at interim stages to alter the trial procedures. The typical decisions include whether to stop or continue the trial based on the observations, and to which treatment arms the new enrolled patients should be assigned, etc. The former decision is made in a group sequential design, while the latter one is made in an adaptive allocation design. Upon achieving required Type I and II error rates, the group sequential designs focus on stopping trials early to reduce the enrolled sample sizes, and in turn enhance the population benefits, while the adaptive allocation designs emphasize on assigning better treatments to patients, and as a result augment the individual benefits. The two goals are conflict with each other. The adaptive allocation procedures assign patients unevenly to the treatment arms which increases the variations and in turn reduce the information contained in the samples. On the contrary, the group sequential designs prefer balanced allocations in which the information can be maximized at each analysis stage. My researches investigate the approaches of combining the two conflict procedures to maximize the overall trial utilities under the Bayesian framework.

For the observational data analysis, I focus on developing semiparametric models as well as estimation methods for handling various data structures, including survival data and repeated measurement data. The model and estimation procedures are robust, efficient and practically preferable to analyze the high dimensional data.

The semiparametric methods are used when population distributions are partially unknown. In these situations, parametric and nonparametric methods are not preferable. The classical parametric method replies heavily on the correctness of the models, while the nonparametric methods loss the efficiency by disregarding any additional population information. The semiparametric approaches overcome the drawbacks. The semiparametric models specify the structures of the known part. The corresponding estimation procedures can process efficiently without the full information about population distributions. The properties of the semiparametric methods are crucial to the current data analysis, especially when the dimensions of the parameter space is large. My studies investigate the frequentist semiparametric methods thoroughly.

From the studies, I develop deep understandings of the Bayesian and frequentist methods. The Bayesian and frequentist methods have their unique merits as well as disadvantages. Bayesian methods are more beneficial for an adaptive study with new samples come over time, while frequentist methods are more suitable for analyzing the data from an observational study. In an adaptive experimental study with new data come regularly, the Bayesian methods update the inferences seamlessly with the new arrived data, and the resulted the posterior distributions automatically include all the previous informations. Nevertheless, the frequentist methods repeat the inferences on the observed data. The repeated analyses violate independent sampling assumptions, and reduce the efficiency of trials. On the other hand, the frequentist inferences are theoretically grounded and have lower requirements on the computational capacity. However, the Bayesian inferences depend extensively on the algorithms, which are computational intensive and often difficult to be justified theoretically.

The specific applications of the techniques in the the experimental design and observational data analysis areas are described in the following sections, which include the detailed discussions of my projects.

Summary of the research works

The Bayesian decision-theoretic approach.

Bayesian decision-theoretic approach has been studied in the Economical areas. The ultimate goal is to maximize utilities by choosing proper decisions. The applications in the biology fields are first introduced by Lewis and Berry (1994). Lewis and Berry (1994) introduced a framework of the Bayesian decision-theoretic method, and illustrated its application to an animal study and clinical trials. The paper shows that the designs using the Bayesian decision-theoretic framework require smaller sample sizes on average than those using classical frequentist methods. Since then, the Bayesian decision-theoretic method has been used in clinical trials. Gausche et al. (2000) applied a Bayesian decision-theoretic method to evaluate the outcomes associated with the use of endotracheal intubation in pediatric patients in out-of-hospital emergency

settings. Gausche et al. (2000) chose to use a Bayesian decision-theoretic method in their study, because it allows more infrequent interim analyses than the traditional methods. In another example, Young et al. (2004) applied the method to design a clinical trial on the prophylactic use of phenytoin. Since Bayesian decision-theoretic designs are optimal with respect to the defined utility functions, they perform better than other designs in maximizing the utility functions. The studies demonstrate the advantages of the Bayesian decision-theoretic approaches and motivate us to extend the use of the methods to design combined adaptive allocation and group sequential clinical trials.

Project 1:

This project is a join work with Dr. J. Jack Lee and Dr. Peter Mueller, and has been published in Statistics in Medicine in January 2013.

The project proposes a class of phase II clinical trial designs with sequential stopping and adaptive treatment allocation to evaluate treatment efficacy. Our work is based on two-arm (control and experimental treatment) designs with binary endpoints. Our overall goal is to construct more efficient and ethical randomized phase II trials by reducing the average sample sizes and increasing the percentage of patients assigned to the better treatment arms of the trials. The designs combine the Bayesian decision-theoretic sequential approach with adaptive randomization procedures in order to achieve simultaneous goals of improved efficiency and ethics. The design parameters represent the costs of different decisions, e.g., the decisions for stopping or continuing the trials. The parameters enable us to incorporate the actual costs of the decisions in practice. The proposed designs allow the clinical trials to stop early for either efficacy or futility. Furthermore, the designs assign more patients to better treatment arms by applying adaptive randomization procedures. We develop an algorithm based on the constrained backward induction and forward simulation to implement the designs. The algorithm overcomes the computational difficulty of the backward induction method, thereby making our approach practicable. The designs result in trials with desirable operating characteristics under the simulated settings. Moreover, the designs are robust with respect to the response rate of the control group.

The semiparametric modeling and estimation.

Current statistical analyzes have two tendencies. For one, the statistical analyzes prefer to use as fewer assumptions as possible to reduce the dependency on the external information. For the other, statistical models tend to extract the information regarding the interested parameters from massive data, which are usually finite dimensional. These two trends motivate the uses of semiparametric methods, which produce robust fits to data with no requirement on correctly specifying the full population distributions.

Unlike the parametric models, which are indexed by finite dimensional parameters, and non-parametric models whose parameter spaces are unified infinite dimensional, the semiparametric models contain both the finite and infinite components. Usually, the parameters of interest are defined to be finite dimensional, while the nuisance parameters are infinite dimensional. With specific models assumptions, the parameter estimations focus on estimating the parameters of interest consistently and efficiently. Bickel et al. (1998) introduced the influence function method for the semiparametric estimations. The work provides theoretic background from differential geometry and functional analysis perspectives for the semiparametric estimations. Tsiatis (2006) introduced the semiparametric methods for practical applications. The work introduces the semiparametric models include restricted moment model, Cox proportional hazard model, etc. Additionally, the book explains the available semiparametric estimations, like the M-estimators, using the influence function concept. Tsiatis (2006) introduce the application of the semiparametric techniques for analyzing various data, e.g., the data with known moment structures, and the coarse data.

According to Bickel et al. (1998) and Tsiatis (2006), with specific model assumptions, the procedures for the semiparametric estimation can be summarized as follows.

1: Find the score functions for the parameters of interest

2: Find the tangent spaces $\mathcal{T}$ and Λ for the parameters of interets and nuisance parameters.

3: Project the score functions to the $\Lambda^{\perp}$, the space orthogonal to Λ to obtain the efficient score functions.

The roots of the score functions are the estimators for the interested parameters.

The semi-parametric method is widely applicable because of the low requirements on the model assumptions and the robustness of the estimations. I focus on the subject, and did extensive researches in my next two projects.

Project 2: A semiparametric method for survival analysis, with application to seamless phase II/III clinical trial design

Summary of the project. This project is a join work with Dr. Yanyuan Ma and Dr. J. Jack Lee. The project involves the utilizations of the kernel regression and imputation techniques. The U-statistics and martingale devices are

used for deriving the large sample properties of the estimators. The work has been submitted to Journal of the American Statistical Association (JASA).

In this paper, we propose a semiparametric framework to describe survival data, where only the dependence of the mean and variance of the survival outcomes on the covariates are specified through a restricted moment model. The semiparametric model provides a better fit than the classic parametric accelerated failure time model when applied to the BATTLE trial data, which motivated this work. We use a second-order semiparametric efficient score combined with a nonparametric imputation device to estimate parameters. The imputation method provides a general strategy for extending to survival data the efficient score method that was developed for complete data. Theoretic results show that the estimators are consistent and asymptotically normal. Compared with the optimal weighted least squares method, the proposed approach improves the efficiency of the parameter estimation as long as the third moment of the error distribution is nonzero. We apply the model and estimation procedure to design a seamless phase II/III clinical trial. Compared with its parametric counterpart of the Weibull survival distribution assumption, the proposed design requires a smaller sample size on average to achieve the pre-specified error rates. This work provides a flexible and robust method to analyze survival data that benefits both data analysis and clinical trial design.

Project 3: The single index model method for analyzing longitudinal data

Summary of the methods used in the project. This project is a join work with Dr. Yanyuan Ma and Dr. Yuanjia Wang. The project is an application of single-index model to analyze repeated measurement data. The parameter estimation approach is a combination of the kernel and Bspline techniques. Because the single-index model overcomes the curse of dimensionality, the model and succeeding estimation methodologies are particularly useful for high dimensional data analysis.

In this project, we describe a semiparametric single-index score model to analyze the repeated measurement disease data from the studies in which the multiple disease related risk scores are collected at each measurement. The single-index risk score combines relevant individual scores to an overall score, namely the total score, which is more interpretable than the individual scores in practice. In addition, the semiparametric single-index score is more flexible than other commonly used total scores in epidemiology. We develop a combined kernel-Bspline method for the succeeding parameter estimation. The method improves a Bspline method to accommodate the situation where the measurement times are random and continuous. We apply the profile procedures to implement the estimation. Asymptotic theory is presented for the resulted estimators. We study the finite sample properties of the method on simulated data sets. The desirable properties of the estimators allow us to apply the method for analyzing the repeated measurement data from the Cooperative Huntingtons Observational Research Trial (COHORT) on Huntington's disease (HD).

Future research plan and summary.

My future researches will continue focus on the Bayesian and frequentist statistical methodology developments, and the applications on the clinical trial design and data analysis. The efforts would also be devoted to develop the methods for high dimensional data analysis, which are particularly useful when huge informations are available. Additionally, the studies will be extended to different areas including biomedical science, finance and computer science.

In summary, the experiences on the researches have enhanced my theoretical and practical abilities. The unique skills acquired during my Ph.D. studies in Rice University are invaluable for continuing my contribution to the statistics field.

References

Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. (1998), *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins series in the mathematical sciences, Springer.

Gausche, M., Lewis, R. J., Stratton, S. J., Haynes, B. E., Gunter, C. S., Goodrich, S. M., Poore, P. D., McCollough, M. D., Henderson, D. P., Pratt, F. D., and Seidel, J. S. (2000), "Effect of out-of-hospital pediatric endotracheal intubation on survival and neurological outcome: a controlled clinical trial." *JAMA: The Journal Of The American Medical Association*, 283, 783 – 790.

Lewis, R. J. and Berry, D. A. (1994), "Group Sequential Clinical Trials: A Classical Evaluation of Bayesian Decision-Theoretic Designs," *Journal of the American Statistical Association*, 89, 1528–1534.

Tsiatis, A. (2006), *Semiparametric theory and missing data*, Springer series in statistics, Springer.

Young, K. D., Okada, P. J., Sokolove, P. E., Palchak, M. J., Panacek, E. A., Baren, J. M., Huff, K. R., McBride, D. Q., Inkelis, S. H., and Lewis, R. J. (2004), "A randomized, double-blinded, placebo-controlled trial of phenytoin for the prevention of early posttraumatic seizures in children with moderate to severe blunt head injury." *Ann Emerg Med*, 43, 435–446.