Statement of Purpose (Harvard Ph.D. in Statistics)

Conducting research is like climbing mountains; we are always testing how far we can go, and in fact the farther we go, the more we see. Unknown knowledge, just like the undiscovered views over the mountain peak, inspires my strong desire and curiosity. Pursing a doctorate is the ideal path for me to realize my dream and fulfill my interest in scientific research.

In my first two years of undergraduate study, I received excellent theoretical training through a wide range of pure math courses, such as real analysis, complex analysis, and functional analysis. More important than achieving a top GPA rank, I gained understanding about how to think critically and raise meaningful questions. In my sophomore summer, I participated in the Second Joint Biostatistics Symposium and learned the language of clinical experimental design, drug efficiency test and epidemic models, which motivated me to pursue more knowledge in statistics.

In my junior year, I successfully finished several projects and received a good training in data analysis, simulation and optimization. Several projects utilizing real data, such as heart disease prediction, online advertisement clicking, and air pollution data, familiarized me to regression and classification algorithms as well as how to conduct statistical inference. The project of utilizing MCEM to derive MLE taught me that the simulation is not only a useful tool in method evaluation, but also an important method itself when closed-form formula is not available. The optimization projects, such as fast implementation of quantile regression and algorithm design for MRF, showed me that appropriate optimization algorithm will render models more practical and applicable. Through these projects and several graduate courses, I became well-prepared and motivated to delve deeper into the world of statistics.

Last summer, I worked on machine learning and statistical modeling with Professor Ying Nian Wu at UCLA. We designed a probabilistic model and incorporated statistical learning algorithms to tackle the image template learning and image classification problem. Our method first conducted unsupervised dictionary learning for image templates, and then utilized the templates to encode images into feature vectors for classification. To improve interpretability of the model and its parameters, I revised the dictionary learning algorithm inversely and showed its equivalence to a clustering process for high density region detection. During the templates' detector training, we were confronted with a lack of positive samples. To handle this, I recalled the resampling method for imbalanced data and adapted its equivalent version of weighting loss function for our problem. Then, to overcome over fitting in high dimensional settings (about 10 4 dimensions), I attached an L2 penalty term to the loss function (using an L1 penalty when sparsity is required). Through this research, I mastered basic principles of statistical learning and came to understand the three essential aspects of modeling: performance, interpretability and efficient implementation. Moreover, this experience at UCLA taught me how to adapt to a new academic environment and new research topics quickly. My other research project was in applied probability with Professor Chenxu Liat Peking University on realized variance calculation of diffusion processes. It is known that PDE method could calculate realized variance for affine diffusion case. Our contribution is to utilize the operator method to resolve the problem under a general diffusion model with jumps. I adapted an operator to integrate the effect of Brownian motion and Poisson jump, and achieved a closed-form asymptotic expansion for the realized variance. Furthermore, being interested on statistical inference on diffusion process, I conducted some literature survey on my own. I was attracted by Samuel Kou's paper of the parameter estimation for diffusion process with concept of data augmentation. And I found that with the assumption of single jump per time interval, we could generalize the framework to cover the jump diffusion case. Also, further research is deserved in extending Kay Giesecke's result on exact (bias-free) sampling of jump diffusion process to a multi-factor framework to better match the observed data. Through this research, I gained a solid training on probability and stochastic analysis, and mastered an asymptotic view for model evaluation.

Through these experiences, I developed an interest in: high-dimensional statistical analysis, Bayesian statistics and statistical learning. In particular, I am interested in Bayesian perspective of statistical models. For instance, in high-dimensional settings, Bayesian priors provide an easy way to incorporate previous knowledge as well as a better interpretation for regularization term. Also, I find the popular idea of latent variables to be very similar to that of data augmentation. While most learning models follow the roadmap from objective function to optimization algorithm and only provide us with an optimal solution or MLE, the Bayesian perspective provides a closed-form or sampling scheme of posterior distribution of responses, parameters or models, which is more panoramic and insightful. As to the computational aspects of Bayesian statistics, through literature survey, I find that techniques to further accelerate MCMC, the posterior approximation method, and the stochastic optimization algorithms incited by the idea of annealing are three attractive research topics for me. With my computation background, I also like to investigate more on this area in my graduate career. Last summer, I visited Harvard University's Statistics Department after my summer intern at UCLA. I was deeply impressed by the welcoming and professional nature of the professors and students, and by the excellent academic environment of the department. Professor Pillai suggested me a reading list on the topics of Bayesian statistics, and Professor Liu illustrated to me that department values interaction between application and theory. All of these convince me that this is the ideal setting for the realization of my academic goals in Bayesian statistics, computational statistics and statistical modeling. As an enthusiastic, creative, and committed researcher, I hope to join Harvard in order to contribute to the department and the greater statistics community.