Applicant: Wenjin Zhang Program: Ph.D. in Statistics

Academic Statement of Purpose

The lifelong pursuit of finding a balance between truth, beauty, and the real world has been my guiding spirit, shaping my early ambition for a career in statistical research. In my view, statistical research is a pursuit worthy of a lifetime's dedication, requiring not only a sensitivity to beauty but also a rigorous commitment to precision and a passion for addressing real-world issues. With this conviction, I undertook an extensive and progressing education at the **University of Science and Technology of China** (USTC), where my growing enthusiasm for statistics fueled my desire to delve deeper. Thus, I am determined to apply to the **Ph.D. program in Statistics at Carnegie Mellon University** as my next station.

Statistics requires finding the most elegant, clear pathways to solve problems, calling for the ability to perceive the full landscape of mathematics and create works of beauty. As the only sophomore in my university's mathematics major to complete **graduate courses** in algebra in both semesters, I developed an early grasp of the broader structure of modern mathematics. During these courses, I devoted considerable effort to correcting errors in proofs within the texts, honing my skills in identifying and repairing logical gaps. Naturally, I scored at least 97 in both graduate courses. After passing rigorous written and oral exams, I was selected to participate in the "Algebraic and Number Theory" summer school organized by the **Chinese Academy of Sciences**, earning a certificate upon completing the final exam. In my junior year, despite a demanding course load of six statistics courses, along with courses in optimization, machine learning, and mathematics, including graduate ones, I achieved the highest GPA in the Department of Probability and Mathematical Statistics for that academic year and earned the **National Scholarship**. These experiences greatly strengthened my mathematical foundation and deepened my commitment to the pursuit of truth.

The greatest allure of statistics, compared to pure mathematics, is its deep roots in real-world problems, with broad and profound influence. Driven by a desire to apply class knowledge to real-world scenarios, I reached out to **Prof. Weijing Tang** at Carnegie Mellon University in November 2023, later collaborating with **Prof. Yinqiu He** from the University of Wisconsin-Madison, and quickly immersed myself in research on community size inference. I implemented our method in R, **simplifying the sampling steps** based on the inherent properties of the **Stochastic Block Model** (SBM) and **extending this framework** to the **Degree-Corrected Block Model** (DCBM) by controlling the degree of nodes. In analyzing experimental results, I meticulously tracked each stage of this complex process, which enabled me to propose and refine methods along the way. For instance, after noting the effectiveness of spectral clustering for initialization, I examined its mechanism and hypothesized that majority voting could yield similar outcomes by preserving cluster structure—an idea confirmed through testing. With extensive simulations, our method has shown promising results in achieving high coverage. We are currently addressing challenges in refining candidate sets to reduce the size of the confidence set and further enhance our approach.

During this research, I became captivated by the unique appeal of statistics, which deepened my desire to explore more. To pursue this goal, I initiated a collaboration with **Prof. Yuqi Gu** at Columbia University to further investigate a method I had encountered in previous work. We discovered that the **stepwise Goodness-of-Fit** (StGoF) method by Jin et al. (2022) not only performs effectively for graph-based network models but also shows significant potential for estimating the number of components in finite mixture models. Through this research, I systematically tested the method across models of increasing complexity—from the **Latent Block Model**, the **Latent Class Model**, to the **sub-exponential mixtures—providing both experimental validation and theoretical proof** under certain conditions. Originally designed for DCBM, StGoF required tailored adaptations for each model, so I adjusted the method to obtain a reasonable estimation of the signal matrix to accommodate each model. In our case, the models of interest lack degree heterogeneity but involve many more parameters, prompting me to try techniques from multiple fields to refine the original proof and enhance its elegance. For example, I employed the pigeonhole principle to construct lower bounds for matrix singular values in one lemma. My experience in troubleshooting proofs, along with a comprehensive understanding of mathematics, was instrumental in completing the full proof, which spanned dozens of pages.

Throughout my research, a commitment to truth-seeking has always been and continues to be a guiding force in my work. While proving a lemma in this project, I **uncovered an academic oversight** that may have gone unnoticed for 34 years. In the 1990 book ***Matrix Perturbation Theory*** by Professors Stewart and Sun, Theorem 3.9 contains an incorrect inequality without an accompanying proof. This theorem was later cited in Lemma 2.1 of the 2021 book *Spectral Methods for Data Science: A Statistical Perspective*, which presents a correct statement but references the erroneous 1990 theorem in its proof. After verifying this issue with my advisor, I reached out to Prof. Jianqing Fan at Princeton University and Prof. Cong Ma at the University of Chicago, who acknowledged the error. I am grateful for their openness and understanding, and I am also gratified to have contributed a small correction to the field of statistics.

As my research deepened, I developed a growing interest in various fields within statistics. While collaborating with **Prof. Yuqi Gu**, **Prof. Yixin Wang** from the University of Michigan later joined our collaboration, and together we embarked on a project focused on causal inference for latent attributes. Through a comprehensive literature review and code implementation, I observed that the **Greedy Equivalence Search** (GES) method was the best fit for our setting, as we needed a causal discovery method as part of our approach. After relaxing the conditions under which the BIC method ensures consistency, I successfully **proved the consistency of our method**, building on the theoretical guarantees in the original GES paper. Our approach shows promise when dealing with a limited number of latent attributes. To improve scalability and practical applicability, we are currently applying a Stochastic Approximation EM (SAEM) algorithm tailored to our setting to verify the results. This project has strengthened my commitment to finding elegant, truthful solutions that not only uphold mathematical beauty but also adapt to real-world needs.

My lifelong pursuit of balance between truth, beauty, and the real world continues to inspire my journey in statistical research. **Carnegie Mellon University** holds special significance for me, as it offers the opportunity to work with distinguished faculty in statistics and machine learning, including **Prof. Weijing Tang**, **Prof. Jiashun Jin**, and **Prof. Yuejie Chi**, and I would be honored to work under their guidance. With strong motivation and a strong research foundation, I am eager to contribute to and grow within Carnegie Mellon University's academic community, striving for a meaningful path in statistics. I sincerely hope for serious consideration of my application.