Statement of Purpose

Shuofan Zhang, Statistics

It is my desire to pursue a Ph.D. in Statistics at Rice University as part of my long-term professional goal of becoming an academic researcher. I have a strong interest in data visualization and data mining.

It was the first statistics course of my master's program taught by Professor Dianne Cook that piqued my interest. Part of the assessments was an interesting competition on Kaggle[1] predicting housing prices. My random forest regression outperformed all my classmates' linear regressions in terms of out-of-sample mean squared error. Studying more on the machine learning technique, I realized its similarity to estimation problems in econometrics and became curious about its possible applications to this field. I decided to pursue my interest in applying statistical methodologies and transferred my specialization from actuarial studies to applied econometrics.

In the year that followed, I received the *Monash Business School Student Excellence Award* for achieving the highest mark in seven of my eight courses. In April 2018, I was chosen as one of the four representatives of Monash University to participate in the *Econometric Game*[2], where teams of postgraduate students examine a research topic and deliver academic papers for competition. Thirty prestigious universities were represented, including Harvard University and the University of Cambridge. The research topic considered the detrimental effects of an individual's unemployment on that individual's happiness, as well as on a group's wellbeing. The dataset contained more than one thousand variables. We first decided to choose the explanatory variables based on empirical results from relevant literature, but realized this approach was inefficient and may omit important information. Hence, I suggested the Least Absolute Shrinkage and Selection Operator (LASSO) to conduct variable selection, with the tuning parameter λ chosen by cross validation. With the variables filtered, our team constructed an ordered probit model with the raw responses to the survey item capturing overall life satisfaction as the dependent variable. Under the assumption of the homogeneous spillover effects amongst individuals in a group, we estimated the multiplier between the effects on an individual and a group. This great experience significantly stimulated my interest in research and helped me understand the power of statistics.

As an attempt to explore the potential of statistical techniques, my master's thesis employed deep learning to facilitate the hypothesis test design, which was supervised by Professor Dianne Cook. The derivation of hypothesis tests and their asymptotic distributions constitute a considerable part of the

[1] Kaggle is an online community of data scientists, owned by Google, Inc. See https://www.kaggle.com/

[2] The Econometric Game is hosted by the University of Amsterdam. See http://econometricgame.nl/

statistics literature. However, the derivation is often complex and the resulting test may lack power. For example, the commonly used unit root tests in time series all suffer from low power in distinguishing the unit root null from stationary alternatives. In my thesis, I trained a binary deep learning classifier to test the null of no structure against linear patterns in a scatter plot, as an alternative to the conventional t-test. Tested on a large unseen dataset, the power $(1-\beta)$ of the classifier was always close to the t-test holding the type I error (α) constant. Given that the t-test is known as the uniformly most powerful test under such experimental settings according to the Neyman–Pearson lemma, this finding implies that the deep learning model has the potential to approach the unknown best test in more complicated situations. The study was then extended to test the null of homoscedasticity against heteroscedasticity using the re-trained classifier. A small dataset of human evaluations was collected using the "lineup" protocol (Majumder, Hofmann, and Cook 2013) and a specific form of the White test was applied to provide a reference level of the test accuracy. The classifier achieved much higher accuracy than both the White test and the human evaluations. Using the data plot as the test statistic instead of a single quantity, this approach could avoid complex derivations, while exploiting the useful information in the plots (see Anscombe (1973) for examples). Although the training design could be challenging, these results encourage future research. For instance, by replacing the scatter plot with time plots one can use this approach to perform a unit root test. I am currently writing a first-author paper on these results to submit to *Statistical Analysis and Data Mining*. This study was also presented by Professor Cook as the 50^{th} *Belz* Lecture for the *Statistical Society of Victoria*[3].

After completing my master's degree, I was excited to accept an offer from Professor Heather Anderson and Professor Farshid Vahid as a research assistant to work on extensions to a paper studying high-dimensional predictive regression with the LASSO (Koo et al. 2016). In this project, I reviewed the literature to investigate the compatibility (or restricted eigenvalue) condition and its implication of choosing the tuning parameter λ for the ℓ_1 norm to achieve the prediction optimality, while taking into account the potential consequences of inconsistency for variable selection by the LASSO (Bühlmann and Van De Geer 2011). I have been self-studying real analysis by reading the book *Principles of Mathematical Analysis* (Rudin 1976) to better understand the relevant concepts. Comparison between the LASSO, the adaptive LASSO and the group LASSO is under consideration. In addition, the out-of-sample mean squared errors of forecasting GDP growth and inflation rate using the LASSO on 146 macroeconomics variables are compared against other approaches including an autoregressive model and a principal component analysis. The potential co-integrating relationships in the selected variables are being studied. Meanwhile, I am co-authoring a paper with the Learning and Teaching team at Monash University which measures student levels of perceptions of live-streaming, a new technology implemented in the lectures. Our study adapted the CRiSP[4] questionnaire which was validated by a

[3]The Victorian Branch of the Statistical Society of Australia Inc. See https://www.statsoc.org.au/branches/victoria/

[4]CRiSP is the name of classroom response system perceptions questionnaire. (Richardson et al. 2015)

combination of factor analyses. Our results revealed three reliable scales: acceptance, usability, and confidence. Following the validation results, I investigated the correlations between the three scales and the self-reported study attitudes using the estimated factor scores.

Although I am open to a variety of topics in statistics, there are several professors at Rice University whose projects are especially appealing to me: Professors Scott, Merényi and Allen. After reading some of their papers, I believe their work is closely aligned with my skills and interests and that Rice will be a great environment for me to thrive.

References

Anscombe, FJ. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21.

Bühlmann, Peter, and Sara Van De Geer. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.

Koo, Bonsoo, Heather M Anderson, Myung Hwan Seo, and Wenying Yao. 2016. "High-Dimensional Predictive Regression in the Presence of Cointegration." https://ssrn.com/abstract=2851677.

Majumder, Mahbubul, Heike Hofmann, and Cook Dianne. 2013. "Validation of Visual Statistical Inference, Applied to Linear Models." *Journal of the American Statistical Association* 108 (503). Taylor & Francis Group: 942–56.

Rudin, Walter. 1976. "Principles of Mathematical Analysis (International Series in Pure & Applied Mathematics)." McGraw-Hill Publishing Co.