主 論 文 要 旨

報告番号	甲 乙 第　　　号	氏 名	加藤 諒

主 論 文 題 名 :

Studies in Semiparametric Causal Inference and Missing Data Analysis

（内容の要旨）

This thesis proposes and applies the semiparametric methods for causal inference and missing data. Causal inference and missing data are inherently the same problem. In this thesis, we apply semiparametric causal inference method to the social science field using propensity score, propose a new semiparametric missing data imputation method, and propose a new semiparametric causal inference method based on instrumental variable.

Chapter 1 introduces the problems of causal inference and missing data. Since this thesis often uses semiparametric Bayes modeling, we introduce Dirichlet process mixture modeling and related literature. Organization of this thesis is also provided.

In Chapter 2, we introduce the application of semiparametric causal inference methodology to social science field, especially the positive accounting and auditing research. Even though Japan is a developed country with the second largest economy in the world as of 2011 and has a unique business culture and power dynamic among audit firms, there remains a dearth of literature investigating the Japanese audit market. This chapter applied semiparametric causal inference method of propensity score matching, and discusses the features of the Japanese audit market and attempts to verify the relationship between accruals-based audit quality and auditor size in Japan. The empirical results show no relationship between audit quality and auditor size in the Japanese audit market, after client characteristics effects have been properly controlled.

In Chapter 3, we propose a new semiparametric Bayes multiple imputation approach that can deal with continuous and discrete variables. Issues regarding missing data are critical in observational and experimental research, as they induce

主 論 文 要 旨

loss of information and biased result. Recently, for datasets with mixed continuous and discrete variables in various study areas, multiple imputation by chained equation (MICE) has been more widely used, although MICE may yield severely biased estimates. Our proposed method enables us to overcome the shortcomings of multiple imputation by MICE; they must satisfy strong conditions (known as compatibility) to guarantee that obtained estimators are consistent. Our exhaustive simulation studies show that the coverage probability of 95 % interval calculated using MICE can be less than 1 %, while the MSE of the proposed one can be less than one-fiftieth. We also applied our method to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, and the results are consistent with those of the previous research works that used panel data other than ADNI database, whereas the existing methods such as MICE, resulted in entirely inconsistent results.

In Chapter 4, we develop a new semiparametric Bayes instrumental variables estimation method. We employ the form of the regression function of the reduced-form equation and the disturbances are modelled nonparametrically to achieve better predictive power of the endogenous variables, whereas we use parametric formulation in the structural equation, which is of interest in inference. Our simulation studies show that under small sample size the proposed method obtains more efficient estimates and very precise credible intervals compared with existing IV methods. The existing methods fail to reject the null hypothesis with higher probability, due to larger variance of the estimators. Moreover, the mean squared error in the proposed method may be less than 1/30 of that in the existing procedures even in the presence of weak instruments. We applied our proposed method to a Mendelian randomization dataset where a large number of instruments are available and semiparametric specification is appropriate. This is a weak instrument case; hence, the non-Bayesian IV approach yields inefficient estimates. We obtained statistically significant results that cannot be obtained by the existing methods, including standard Bayesian IV.

Chapter 5 considers the case where instrumental variable (IV) are available to infer the effect of interested variable to the outcome (or the causal effect), but some components of IV are missing with the missing mechanism of not missing at

random (NMAR). Although NMAR requires the analysis to prespecify the missing mechanism, it is unknown for us and what is worse, it is generally not identified. We use the IV distribution of original population as an auxiliary information, and show that missing mechanism can be represented as identifiable nonparametric generalized additive model. We also introduce MCMC algorithm that impute the missing values and simultaneously estimate parameters of interested. Simulation studies show that our proposed method yield the smallest MSE compared with Bayesian imputation without population level information, MICE, and complete case analysis under the situation where original distribution of the IV follows log-normal distribution and missing not at random.

Chapter 6 provides conclusion with a short discussion.