Summary of Studies in Semiparametric Causal Inference and Missing Data Analysis

by

Ryo Kato

Graduate school of Economics

KEIO UNIVERSITY

Chapter 1. Introduction

Causal inference and missing data

Inferring causal effects of some interested treatments is a fundamental goal in many disciplines. In observational studies, such as social and behavioral science, researchers seek to conduct quasi-experiments, where randomized controlled trials (RCTs) are not feasible. The potential outcome framework is often applied to the causal inference, in which one of the outcomes (treatment or control) is observed and the rest are missing. Therefore, causal inference is inherently a problem of missing data.

A frequently applied methodology for estimating causal effects is the propensity score analysis developed by Rosenbaum and Rubin (1983); the number of articles published that use the propensity score method is rising exponentially (Thoemmes and Kim, 2011). Propensity score is a conditional probability that indicates how likely it is for each participant to be assigned to the treatment group given the covariates representing its features. Given the propensity score, each participant is randomly assigned to the treatment or control group, enabling us to infer the causal effect as if RCTs are carried out.

Although propensity score analysis is not robust enough to the effect of the unobserved confounding, the instrumental variables (IV) method can be a strong technique if one can employ sufficient IV. Under the recent situation of several kinds of data, the IV method has become increasingly important in many empirical fields. Therefore, while initially, IV estimation and its application were restricted to empirical economics, it has begun to be applied in other fields such as epidemiology for causal inference. In particular, IV approaches are employed when it is not feasible to carry out RCTs or the standard causal inference methodology, which assumes that no unobserved confounding exists. If it is possible to find sufficient IV predictive of endogenous variables, having no direct impact on the outcome, and independent of the unobserved confounders, then the effect of the unobserved confounders can be controlled. Therefore, introducing sufficient IVs itself can be a great invention, and many researchers are trying to find them.

On the other hand, issues regarding missing data are critical in observational and experimental research as they induce loss of information and biased results. Unfortunately, the missing data problem is ubiquitous. The National Research Council (2010) published a report including recommendations on treating missing data in medical science research, indicating that researchers should employ as many confounders as possible in order to obtain valid estimates. However, when they employ a greater number of covariates, the number of observations with at least one missing component increases. Additionally, if a researcher is interested in using a regression model containing missing components in covariates, a complete case analysis, which is thought to be the most applied "method" for treating the missing data, results in biased estimates in many cases.

In this thesis, we propose and apply the methods for causal inference and missing data. As stated above, causal inference and missing data are inherently the same problem. 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.

Semiparametric model

In this thesis, we use the term "semiparametric" frequently. Probabilistic models used in this thesis are partially specified and others are not (parametrically) specified. We use parametric (or finite-dimensional) specification to the model where the parameters are of interest, and use nonparametric (or infinite-dimensional) parameters to the model of less interest or nuisance.

Semiparametric models are often appeared in many kinds of disciplines. Famous statistical methodologies with semiparametric models are generalized method of moments (GMM), proportional hazards model, and semiparametric estimators are shown to be well-behaved (Tsiatis, 2006).

In Chapter 2, we use propensity score matching method, and this is a kind of semiparametric model. We specify parametric logistic regression model for estimating the propensity score. On the other hand, we do not assume parametric regression model between the outcome and the covariates.

From Chapter 3 to Chapter 5, we use Bayesian semiparametric models. While we assume a parametric structure for the model of the main interest (e.g. the substantive model in Chapter 3, or the structural equation in Chapter 4 and 5), we do consider Bayesian nonparametric form rather than a parametric form for the less interested model which we should avoid parametric assumptions. The examples of nonparametric models of this thesis are covariate distribution in Chapter 3, or the reduced form equation in Chapter 4 and 5. Since these distribution can be the modeled by a large number of covariates with a large number of the outcome, and moreover, prespecification of the model are generally difficult, we use nonparametric Bayes model to avoid misspecification bias (Chib, 2007).

Our nonparametric Bayes representations are based on DPM (Dirichlet Process mixture) modeling. DPM modeling is frequently utilized in applied statistical modeling when researchers intend to avoid making assumptions about parameter distribution within the Bayesian framework. For example, Hirano (2002) developed autoregressive models with individual effects where the disturbances are not restricted to a parametric class. Rodriuez et al. (2009) used DPM to develop a Bayesian semiparametric approach for functional data analysis. Miyazaki and Hoshino (2009) proposed a Bayesian semiparametric item response model with DP prior. Kunihama and Dunson (2016) constructed a method for variable selection within Bayesian nonparametric DPM. Kunihama et al. (2016) developed a nonparametric Bayes model with DPM to incorporate sample survey

weights. The theoretical properties of DPM were investigated by Shen et al. (2013). Fortunately, the DPM model can be estimated with a relatively simpler MCMC algorithm by applying blocked Gibbs sampling (Ishwaran and James, 2011).

Our Bayesian modeling of Chapter 3, 4, and 5 is the product of the parametric part and the nonparametric part with the Bayesian theorem, hence we call these models the semiparametric model.

Chapter 2. Semiparametric causal inference in positive accounting and auditing research

In this chapter, 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.

Many existing studies evaluate the relationship between audit quality and auditor size. Starting with DeAngelo (1981), studies on the U.S. market reveal that, generally, large audit firms with international brand names (hereinafter, Big N) provide better audit quality than do other firms (e.g., Becker et al., 1998; Behn et al., 2008; Francis et al., 1999). The office size of audit practices is also positively related to audit quality (e.g., Choi et al., 2010; Francis and Yu, 2009). The literature also notes that audit practice office size is positively related with audit quality (e.g., Francis and Yu 2009; Choi et al. 2010).

Auditors' incentive to provide high quality audit service can be influenced by the following environmental characteristics: litigation risk and reputation concerns for audit firms. The Japanese market is categorized in the low litigation risk group (e.g., Numata and Takeda, 2010; Skinner and Srinivasan, 2012), leaving auditors' reputation concerns as the most important audit quality incentive in Japan. East Asia cultures, including Japan, are well known for their strong emphasis on the importance of reputation or "face" (Wong and Ahuvia, 1998). It could be expected that audit firms in Japan will work to maintain a high level of reputation, and the Big N firms might provide better services because they have a higher reputation or "face" to lose.

Using discretionary accruals, our findings provide empirical evidence that no relationship between audit quality and auditor size exists in the Japanese audit market, after client characteristics

effects have been properly controlled using semiparametric causal inference. The low litigation and high reputation characteristics of Japanese audit environment shows no effect on the audit quality difference between Big N and Non-Big N after controlling for confounding variables related to Japanese companies. Since these results are not obtained from prior surveys, semiparametric causal inference seems to be useful when applied to social science in which many confounders exist.

Chapter 3. Semiparametric Bayes multiple imputation for regression models with missing mixed continuous-discrete covariates

For datasets with mixed continuous and discrete variables in various study areas, multiple imputation by chained equation (MICE), in which missing variables are iteratively imputed based on full conditional specification (FCS), has been cited numerous times by researchers from several fields including medical statistics (van Buuren, 2007; White et al., 2011; Paton et al., 2014). This is because the researchers, especially the imputers, are not required to construct an explicit joint multivariate model with mixed-scale variables (continuous, categorical, ordinal, and so on). More specifically, the MICE-FCS approach specifies a multivariate imputation model using a sequence of seemingly "appropriate" univariate regression models corresponding to the types of missing variables; namely, one only needs to assign a univariate linear regression with a normally distributed error term for an incomplete continuous variable, a logistic regression for an incomplete binary variable, an ordered logistic regression for an incomplete ordinal variable, and so on. Moreover, researchers can easily implement MICE-FCS using several existing statistical software packages, such as the mice package in R and S-plus, proc mi with the FCS option in SAS, and mi impute in STATA.

In spite of the widespread use of MICE-FCS, recent studies showed that it leads to severely biased estimates in various setups. Liu et al. (2014) proved that using MICE-FCS does not guarantee that the asymptotic distribution is equivalent with the existing Bayesian MI estimator when the families of the conditional models are "incompatible" (see Section 4 in Liu et al. (2014)). In fact, simulation studies by Bartlett et al. (2015) showed that MICE yields biased estimates when treating incompatible conditional models. Unfortunately, violation of the compatibility assumption is not uncommon (the example of the violation of the compatibility assumption is provided in section 2.1). Therefore, although MICE-FCS is simple and convenient to use, it can provide statistically valid estimates in very limited cases.

In this chapter, we propose a new flexible semiparametric Bayesian framework for MI, which is capable of treating mixed-scale incomplete variables. The model formulation is different from that seen in the existing literature in two ways.

First, we express the full model as the product of the covariate distribution (conditional distribution of incompletely observed covariates given completely observed covariates) and the substantive model (the regression model researchers are interested in). We assume the parametric model to the substantive model since the researchers conducting applied research are generally concerned with the parameters of the functions in the substantive model, which should be built upon the existing theories or previous literature in the field of study. Examples of the parametric substantive model are the Cox regression and the logistic regression in epidemiological and clinical research. On the other hand, with regard to the covariate distribution, we specify a joint distribution of the missing variables using the probit stick-breaking process mixture (PSBPM) model proposed by Chung and Dunson (2009), whose model specification is based on the Dirichlet process mixture (DPM) model. Ibrahim et al. (2005) also pointed out that one of the caveats of treating missing covariates lies in specifying the parametric model of the covariate distribution. However, it is nearly impossible to correctly prespecify the covariate distribution based on existing theories or some inferences, because the relationships of the missing variable and the complete variables are often "multivariate-to-multivariate", they can be non-linear relationships, or they may be non-normally distributed. Therefore, we employ the nonparametric Bayesian specification; specifically, we use PSBPM modeling instead of DPM since the stick-breaking weights can vary depending on the predictors. Since our approach do not rely on FCS approach, we do not have to consider the compatibility assumption holding. Murray and Reiter (2016) proposed fully nonparametric multiple imputation method using DPM model with local dependence. However, they do not consider the existence of the substantive model, hence their main scope of the inference is the means or the variances of the imputed variables, and it cannot estimate the interested parameters of the substantive model.

Second, we express mixed-scale variables through the transformation of the latent continuous variables for probit modeling. This underlying continuous variables approach is used in the context of the DPM model, as in Kottas et al. (2005) for ordinal variables; in Canale and Dunson (2011) for count variables; and in Kim and Ratchford (2013) for ordinal variables. This approach enables us to deal in a straightforward manner with many types of variables in the joint covariate distribution without specifying the complicated conditional joint distribution of mixed-scale variables.

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.

Chapter 4. Semiparametric Bayes instrumental variable estimation with many weak instruments

This article presents a new semiparametric Bayes model for instrumental variables problems. We treat the reduced-form equation (or the "first-stage" regression model) and the joint distribution of the error terms as nonparametric and potentially changing in form, corresponding to the values of the instrumental variables. In addition, our emphasis is on the semiparametric model formulation. Structural equation models (or the "second-stage" regression models), which are of interest in inference, are formulated parametrically, whereas the reduced-form equation and the disturbance terms are formulated nonparametrically.

Instrumental variables (IV) methods have become increasingly important in many empirical fields. Initially, IV estimation and its application were restricted to empirical economics. However, IV methods have been recently applied in other fields, such as epidemiology for causal inference (Ramsahai and Lauritzen, 2011; Baiocchi et al., 2014; Wang et al., 2017). In particular, IV approaches are employed when it is not feasible to carry out randomized controlled trials (RCT) or standard causal inference methodology, which assumes that no unobserved confounding exists. If it is possible to find sufficient IV that are predictive of endogenous variables, have no direct effect on the outcome, and are independent of the unobserved confounders, then the effect of the unobserved confounders can be controlled.

However, inference based on IV is prone to be imprecise when the instruments explain only a small portion of the variation in the endogenous variable (weak instrument case). This problem is inherent in samples of small size, as the data is insufficient to identify the parameters of interest (Conley et al., 2008). In general, frequentist methods depend on asymptotics and this property can be a hindrance when the sample size is small. Therefore they are not occasionally suitable for IV problems. By contrast, since Bayesian methods do not rely on asymptotics, applying these methods to IV problems is a reasonable choice. Even though there are no direct incentives for adopting Bayesian methods, Conley et al. (2008) showed that they incur smaller mean squared error (MSE) and provide better interval estimation compared with non-Bayesian methods when the structural equation and the reduced-from equation are properly specified. Moreover, Bayesian methods allow for more flexible modeling of the structural equation, since Bayesian inference depends only on the joint model of the structural and the reduced-form equation, whereas classical methods require the

development of different estimation procedure according to whether we have discrete, clustered, or panel data.

One of the disadvantages of Bayesian methods is that they impose strong distributional assumptions on the parameters. This is the case in IV problems, since Bayesian IV generally assumes that the joint distribution of disturbances is bivariate normal. Conley et al. (2008) proposed another Bayesian IV method that uses a Dirichlet process mixture (DPM) model for the error terms. It moderates the assumption on the disturbances by using DPM nonparametric specification.

Our proposed procedure is a semiparametric Bayesian IV method and is more flexible than Conley's method. We also assume that disturbances have nonparametric structure. Moreover, we use nonparametric formulation in the reduced-from equation. In general, the true functional form of the reduced-from equation is unknown, and its parameters are not of interest. In addition, if we use many instruments, parametric modeling of a large number of variables may result in misspecification bias (Chib, 2007). The approaches assuming that the reduced-form equation has some specific functional form (linear regression is often assumed), including frequentist and Conley's IV methods, yield unbiased estimates of the structural equation. However, they are less efficient. By contrast, if, for example, the reduced-from equation is not a simple linear combination with additive disturbances, our semiparametric Bayes model fits the data better and yields efficiency gains compared with classical parametric method and Conley's method.

Since the parameters of the structural equation are important in applied research, we assume that the structural equation regression model has parametric structure. Moreover, our model is different from other frequentist nonparametric IV approaches in that these approaches use nonparametric specification in the structural equation and parametric specification in the reduced-form equation.

We employ a probit stick-breaking process mixture (PSBPM) model proposed by Chung and Dunson (2009) to realize more flexible semiparametric representations for IV. Nonparametrics based on Dirichlet process makes it possible to represent a distribution by infinite mixture of well-known "base" distributions. Whereas the mean regression structure of the DPM is reduced to a linear regression model, PSBPM is more flexible than DPM since it enables us to make a probability weight of the components change by predictors in the regression model. Hence, we can treat reduced-form equation and the joint distribution of error terms as potentially changing in shape as the value of instruments vary. Even in the case that the reduced-form equation and the error terms are truly linear and bivariate normal, respectively, our procedure has small efficiency loss, since it is not necessary to prespecify the number of components. The optimal number of components, which is needed for a finite mixture of regression models, is defined by the data.

We conduct a Monte Carlo simulation study in order to evaluate the performance of the proposed method. We investigate the finite sample performance of the estimators when the reduced-

from model is not a simple linear combination. The proposed method may incur as little as 1/30 of the MSE incurred by existing procedures. Moreover, the coverage of nominal 95% confidence (or credible) intervals of the proposed method is very close to 0.95, whereas the other methods provide significantly narrower or wider interval estimates.

The proposed method is applied to a real Mendelian randomization dataset. In general, the number of instrumental variables in Mendelian randomization (i.e. information of the genotype) is large, and correct specification of the reduced-form regression model is difficult. In addition, the instruments may not explain the endogenous variables satisfactorily, and non-Bayesian methods that rely on the asymptotic approximation yield biased results. Therefore, the proposed Bayesian nonparametric formulation for the reduced-form equation is appropriate and results in obtaining efficient endogenous (causal) parameters. In fact, we provide statistically significant results that are not obtained by the standard Bayesian IV approach.

Chapter 5. Semiparametric Bayes missing instrumental variable estimation with population information

When it is not feasible to conduct randomized controlled trials (RCT) or quasi-randomized experiments, the IV approach can be a very useful tool to infer the causal effect if it is possible to find sufficient IVs. Since they can properly eliminate the confoundings caused by unobserved factors, IV models are developed and applied in many empirical economics researches, where unobserved confoundings are ubiquitous. Therefore, introducing sufficient IVs can in itself be great invention, and many researchers are trying to find them.

Despite the desperate efforts, IVs tend to be missing. For example, information of twin are often used as an instrument, but it is only observed for sub-samples. Aaslund and Gronquist (2010) used twin birth as an instrument to survey the effect of family size on the quality of children. In this case, a twin birth can be observed for families with twins. Of course, the complete case analysis seems to result in biased results, and as an alternative approach, restricting the sample to families with more than two children lose the efficiency in sample size. Variables on children, such as child BMI, are also frequently employed as instruments. However, since child information comes from different sources than the endogenous parents' information (e.g. BMI), there is a tendency for the former to be missing. Real data analysis shown in Section 5.4 is an example wherein instruments are missing for many observations since they are sourced from other surveys.

In addition to economics, missing IV is a common problem in other fields. Mendelian randomization uses genotype information as an instrumental variable to infer the causal effect of a

biomarker to a disease. Since the appropriate genetic variant is independent of the confounders of the intermediate phenotype-outcome association and can affect the outcome only through the causal intermediate phenotype as long as it is related to the intermediate phenotype, it has recently been applied in economics as well as in biostatistics. In general, as genetic variants explain only a small portion of the endogenous population, Mendelian randomization requires large sample sizes (Smith, 2006) to satisfy enough causal associations. However, Mendelian randomization datasets are often missing (Palmer et al., 2011) and a large enough sample size cannot be guaranteed.

These example of missing IV shows us that complete case analysis, namely, only those samples where all the instruments are observed, are used in the analysis, thus resulting in biased results and wrong decision-making. In this chapter, we develop a semiparametric method to impute the missing portion of IV and simultaneously infer the causal effect. A point that differs from the model proposed in Chapter 3 is that we consider the case with not missing at random (NMAR). Therefore, missingness of instruments remains associated with the missing instruments even after controlling for other observed variables. In the NMAR case, the interested regression model cannot be identified without an additional assumption (Little and Rubin, 2002). An example of such an assumption is strong parametric assumption on the regression and missing mechanism (Kott and Chang, 2010). However, Miao et al. (2015) showed that probit specification on the missing mechanism can identify normal and normal-mixture models while logit specification can less identify them.

We take another assumption that the IV distribution of the original population is available as an auxiliary information. In many cases, the population-level information is available from other data sources such as government statistics or research institutions, and some researches utilize this auxiliary information to estimate individual-level causality. Imbens and Lancaster (1994) and Hellerstein and Imbens (1999) incorporated population-level information as momentary conditions to infer individual-level models using the generalized method of moments (GMM). Another instance where population-level information is used is the empirical likelihood estimation (Qin, 2000; Chaudhuri et al., 2008). Such approaches are also applied to the missing data issues. Nevo (2003) proposed the propensity score weighting method using the moment conditions obtained from auxiliary population-level information. Igari and Hoshino (2018) introduced the Bayesian method with population-level information that dealt with repeated durations under unobserved missing indicators.

Although prior works incorporating population-level information to deal with missing variables use moment conditions, our proposed method uses probability distribution of population as auxiliary information since the momentary conditions have less information than original distribution. Under the condition that the original population distribution of the missing IV is known, followed by the theorem in Hirano et al. (2001), we show that the missing mechanism is

nonparametrically identified with generalized additive model, and the substantive IV regression models are also identified. In general, since fully nonparametric missing mechanism are not identified, parametric missing mechanism are frequently assumed (Kott and Chang, 2010). However, misspecification of missing mechanisms results in severely biased estimates (Kim and Yu, 2011). Kim and Yu (2011) developed a semiparametric missing mechanism approach which incorporates nonparametric specifications on observed variables but not on unobserved variables. However, their method, as well as other prior works, cannot identify the nonparametric part of unobserved variables. On the other hand, we assume the availability of the information of the original population distribution of missing IV so that our proposed method can specify the fully nonparametric missing mechanisms on observed and unobserved variables. Furthermore, our missing mechanism can incorporate cross terms of observed and unobserved variables, which cannot be identified by the existing methods.

We express the full model as the product of the conditional IV distribution (conditional distribution of missing IVs given completely observed exogenous), the substantive IV model (the structural equation of the main interest and the reduced-form equation), and the missing mechanism. We assume the parametric model to the structural model since the researchers conducting IV regression are generally concerned with the coefficient parameters of the endogenous variable. On the other hand, with regard to the reduced-form equation and the error terms, we specify the PSBPM nonparametric model since we can achieve more efficiency when these distributions are misspecified as described in Chapter 4. Conditional IV distribution is also represented by PSBPM nonparametric formulation since it is nearly impossible to correctly prespecify the covariate distribution based on existing theories or some inferences, as the relationships of the missing variable and the complete variables are often multivariate-to-multivariate.

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. Conclusion

In this thesis, we focused on semiparametric causal inference (propensity score analysis and instrumental variable method) and semiparametric missing data analysis.

Future works will be two directions. First one is the development of theoretical aspects of proposed statistics. Semiparametric Bayesian instrumental variable method proposed in Chapter 4 yields considerably more efficient estimates than the existing estimator. The proposed method changed the reduced-form equation from the classical specification and some approaches have

recently proposed which modify the reduced-form. For example, Andrews and Armstrong (2017) have proposed mean-unbiased IV estimator by modifying the reduced-form equation. We will clear up the statistical property of the estimators including the asymptotics. As stated in Chapter 5, identification conditions within NMAR framework for multiple instruments should be revealed.

Another direction is the applications of the proposed method to wider fields. We focused on the application to the medical statistics and empirical economics, we should consider the adoption to other kinds of dataset. Such dataset may contain mixed-scale variables, large dimensional covariates, and may be sparse. We will extend our semiparametric model to more flexible one by accommodating the knowledge such as latent variables, variable selection, and machine learning.

References

Aaslund, O. and H. Gronquist (2010). Family size and child outcomes: is there really no trade-off. Labour Economics, 17, 130-139.

Andrews, I. and Armstrong, T. B. (2017). Unbiased instrumental variables estimation under known first-stage sign. Quantitative Economics, 8, 479-503.

Baiocchi, M., Cheng, J. and Small, D. S. (2014). Instrumental variable methods for causal inference. Statistics in Medicine, 33, 2297-2340.

Bartlett, J.W., Seaman, S. R.,White, I. R. and Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. Statistics in Medicine, 24, 462-487.

Becker, C. L., DeFond, M. L., Jiambalvo, J. and Subramanyam, K. R. (1998), The effect of audit quality on earnings management. Contemporary Accounting Research, 15, 1-24.

Behn, B., Choi, J. H. and Kang, T. (2008), Audit quality and properties of analyst earnings forecasts. The Accounting Review, 83, 327-359.

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. Journal of the American Statistical Association, 106, 1528-1539.

Chaudhuri, S., Handcock, M. S. and Rendall, M. S. (2008). Generalized linear models incorporating population level information: an empirical-likelihood-based approach. Journal of the Royal Statistical Society: Series B, 70, 311-328.

Chib, S. (2007). Analysis of treatment response data without the joint distribution of potential outcomes. Journal of Econometrics, 140, 401-412.

Choi, H. K. and Ford, E. S. (2007). Prevalence of the metabolic syndrome in individuals with hyperuricemia. The American Journal of Medicine, 120, 442-447.

Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with

variable selection. Journal of the American Statistical Association, 104, 1646-1660.

Conley, T. G., Hansen, C. B., McCulloch, R. E. and Rossi, P. E. (2008). A semi-parametric Bayesian approach to the instrumental variable problem. Journal of Econometrics, 144, 276-305.

DeAngelo, L. E. (1981). Auditor size and audit quality. Journal of Accounting and Economics, 3, 183-199.

Francis, J. R. (2011). A framework for understanding and researching audit quality. Auditing: A Journal of Practice & Theory, 30, 125-152.

Francis, J. R. and Yu, M. (2009). Big four office size and audit quality. The Accounting Review, 84, 1521-1552.

Hellerstein, J. K. and Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. Review of Economics and Statistics, 81, 1-14.

Hirano, K., Imbens, G. W., Ridder, G. and Rubin, D. B. (2001). Combining panel data sets with attrition and refreshment samples. Econometrica, 69, 1645-1659.

Hoshino, T. (2013). Semiparametric Bayesian estimation for marginal parametric potential outcome modeling: Application to causal inference. Journal of the American Statistical Association, 108, 1189-1204.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R. and Herring, A. H. (2005). Missing data methods for generalized linear models: A comparative review. Journal of the American Statistical Association, 100, 332-346.

Igari, R. and Hoshino, T. (2018). A Bayesian data combination approach for repeated durations under unobserved missing indicators: Application to interpurchase-timing in marketing. Computational Statistics & Data Analysis, 126, 150-166.

Imbens, G. W. and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. The Review of Economic Studies, 61, 655-680.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stickbreaking priors. Journal of the American Statistical Association, 96, 161-173.

Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. Journal of the American Statistical Association, 106, 157-165.

Kim, J. S. and Ratchford, B. T. (2013). A Bayesian multivariate probit for ordinal data with semiparametric random-effects. Computational Statistics and Data Analysis, 64, 192-208.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. Journal of the American Statistical Association, 105, 1265-1275.

Kottas, A., Muller, P. and Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. Journal of Computational and Graphical Statistics, 14, 610- 625.

Kunihama, T. and Dunson, D. B. (2016). Nonparametric Bayes inference on conditional independence. Biometrika, 103, 35-47.

Kunihama, T., Herring, A. H., Halpern, C. T. and Dunson, D. B. (2016). Nonparametric Bayes modeling with sample survey weights. Statistics and Probability Letters, 113, 41-48.

Lawrence, A., Minutti-Meza, M. and Zhang, P. (2011). Can Big 4 versus non-Big 4 differences in audit quality proxies be attributed to client characteristics?, The Accounting Review, 86, 259-286.

Little, R. J. and Rubin, D. B. (2002). Bayes and multiple imputation. Statistical analysis with missing data, 200-220.

Liu, J., Gelman, A., Hill, J., Su, Y. S. and Kropko, J. (2014). On the stationary distribution of iterative imputations. Biometrika, 101, 155-173.

Louis, H. (2005). Acquirers' abnormal returns and the non-Big 4 auditor clientele efeect. Journal of Accounting and Economics, 40, 75-99.

Miao, W., Ding, P. and Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. Journal of the American Statistical Association, 111, 1673-1683.

Miyazaki, K. and Hoshino, T. (2009). A Bayesian semiparametric item response model with Dirichlet process priors. Psychometrika, 74, 375-393.

Murray, J. S. and Reiter, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. Journal of the American Statistical Association, 111, 1466-1476.

National Research Council (2010). The prevention and treatment of missing data in clinical trials, National Academic Press, Washington, DC.

Numata, S. and Takeda, F. (2010). Stock market reactions to audit failure in Japan: the case of Kanebo and ChuoAoyama, International Journal of Accounting, 45, 175-199.

Palmer, T. M., Lawlor, D. A., Harbord, R. M., Sheehan, N. A., Tobias, J. H., Timpson, N. J., Smith, G. D. and Sterne, J. A. (2012). Using multiple genetic variants as instrumental variables for modifiable risk factors. Statistical methods in medical research, 21, 223-242.

Paton, N. I., Kityo, C. and Hoppe, A. and Reid, A. and Kambugu, A. and Lugemwa, A. (2014). Assessment of second-line antiretroviral regimens for HIV therapy in Africa. The New England Journal of Medicine, 371, 234-247.

Qin, J. (2000). Combining parametric and empirical likelihoods. Biometrika, 87, 484-490.

Ramsahai, R. R. and Lauritzen, S. L. (2011). Likelihood analysis of the binary instrumental variable model. Biometrika, 98, 987-994.

Rodriuez, A., Dunson, D. B. and Gelfand, A. E. (2009). Bayesian nonparametric functional data analysis through density estimation. Biometrika, 96, 149-162.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41-55.

Shen, W., Tokdar, S. T. and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation

with Dirichlet mixtures. Biometrika, 100, 623-640.

Skinner, D. J. and Srinivasan, S. (2012). Audit quality and auditor reputation: evidence from Japan. The Accounting Review, 87, 1737-1765.

Thoemmes, F. J. and Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. Multivariate behavioral research, 46, 90-118.

Tsiatis, A. (2006). Semiparametric theory and missing data. Springer Science & Business Media.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, 16, 219-242.

Wang, L., Robins, J. M. and Richardson, T. S. (2017). On falsification of the binary instrumental variable model. Biometrika, 104, 229-236.

White, I. R., Royston, P. and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. Statistics in Medicine, 30, 377-399.

Wong, N. and Ahuvia, A. C. (1998). Personal taste and family face Luxury consumption in Confucian and Western societies. Psychology & Marketing, 15, 423-441.